Hate Speech Annotation

Anna Astori, Annie Thorburn, Jake Freyer, and José Molina

9 February 2016

Goals

Our task is to annotate hate speech in social media posts, with the goal of training a machine learning algorithm to detect it independently. This will have potential applications in automated forum moderation and filtering apps for social media.

A key component to this task is identifying what constitutes as hate speech and what does not, which is by no means a trivial task. The line between the two is often not clear-cut, and the annotation guide will have to be very specific in order to insure good inter-annotator agreement. Some issues we expect to face include:

- Distinguishing between use of a slur by someone outside the group it refers to (probably hate speech) and by someone within the group (probably a reclamation rather than hate speech)
- Identifying statements that seem innocuous out of context but discriminatory in a particular context
- Conversely, identifying statements that may sound discriminatory out of context but appear in a context where it is clear that prejudiced implications are not intended
- Drawing a boundary through fuzzy areas that toe the line between hatefulness and legitimate discourse on a sensitive topic—including areas like comedy and political speech where reasonable people disagree on what constitutes an offensive statement

Corpus

The data for this annotation project will be collected from a Twitter feed. Twitter is an ideal source for collecting and building a corpus for this topic, because it represents an extensive, publicly available, and easily accessible source of relatively unfiltered speech.

We will improvise a body of tweets likely to contain hate speech by filtering for general terms, as well as targeting accounts known to produce hateful posts. A control group can be taken from a general Twitter feed.

We will not know how large our corpus will be until we see how much hate speech presents itself. We anticipate that it will be fairly large, as annotating each tweet should be fairly quick.