Assignment 1: Task 1

Matthias Agema (2707560)

16 September 2022

1 Paper A: Kumar et al. (2018)

- The corpus is annotated by the use of a tagset including 3 top-levels, namely overtly aggressive, covertly aggressive and non-aggressive. One level down, each of the two aggressive levels contain 2 attributes, discursive role and discursive effects. Discursive roles are attack, defend and abet. Discursive effects are of 10 kinds: physical threat, sexual aggression, gendered aggression, racial aggression, communal aggression, casteist aggression, political aggression, geographical aggression, general non-threatening aggression and curse/abuse.

- Motivation of the paper comes from the almost completely impractical and ineffective way of handling online aggressive behaviour manually. Due to the amount and pace which new data is being created through users.

- A possible research question for this paper could be something like: How can an annotated corpus be achieved using an aggression tagset of Hindi-English code-mixed data?

- After collecting data, set up an annotation scheme and annotating the corpus, they found that people on Facebook are more vocal and overtly aggressive than on Twitter. The data also shows that the majority of code-mixed comments and tweets are aggressive. Posts in English are largely non-aggressive and posts in Hindi turned out to be equally distributed. They also found out that some posts could be interpreted as depicting more than one discursive effect. Therefore, in the second round annotators were given an option to annotate the posts with more than one discursive effect, what led to a better inter-annotator agreement.

- In the conclusion they state that their dataset can be an invaluable resource for understanding and automatically identifying aggressive messages and tweets.

- An interesting aspect of this paper is in my opinion the multi-level annotation scheme they used. What makes it also difficult and vague for annotators.

2 Paper B: Zampieri et al. (2019)

- In this paper is a fine-grained three layer annotation scheme established. Layer A discriminates cases between being offensive or non offensive. Something is offensive when it contains non-acceptable/profane language or a targeted offense. In the next layer offensive posts will be distinguished in untargeted and targeted insults. Where targeted posts are targeted to a individual, group or others. The latter three targets imply the last layer C of this annotation scheme. Something is targeted at a group when they are considered as a unity due to ethnicity, gender or sexual orientation, political affiliation, religious belief, or other common characteristics. Individual speak for themselves and what remains are others, such as organizations, events etcetera.

- Motivation for this paper is that previous work only focused on very specific types of offensive content, such as cyberbulling, hate speech or cyber-aggression.

- A possible research question for this paper could be: How can the Offensive Language Identification Dataset be annotated by identifying the type and target of the offense and can it be used for classification?

- They found a high class imbalance between the different classes, what was one of their key challenges. Therefore it was difficult to predict for example the OTH class. Furthermore, there was no need for them to use more than three annotators to reach their minimal agreement scores.

- Their conclusion is that they provided the first dataset to contain annotation of type and target of offenses, that is challenging to predict with, but doable.

- What I found interesting is that they showed how well the different models performed per layer in their three layer annotation scheme.

3 Paper C: de Gilbert et al. (2018)

- In this paper is the annotation scheme distinguishes four classes, Hate, NoHate, Relation and Skip. Something is hate speech when it satisfies the following three requirements:

  - It is a deliberate attack
  - Directed towards a specific group of people
  - Motivated by aspects of the group's identity

  If a post does not satisfy the above mentioned requirements, it can be annotated as NoHate. The third class, Relation, can be used for posts that do not contain hate speech on their own, but the combination of

several posts does. At last, posts that are not written in English or can not be labeled as Hate or NoHate, can be skipped by the annotator.

- The motivation for this paper is because of the rapid growth of social media over the past decade, that makes it impossible to monitor what is being said. Therefore, automatic hate speech classifiers are needed to detect what can and what cannot be posted on social media. And all this starts with a useful labeled dataset.
- How can a detailed annotation scheme be set up for automatic classifying hate speech at a White Supremacy Forum?
- They found that the dataset is unbalanced as there exist many more sentences not conveying hate speech than the hateful ones. Another thing they found out is that classifying non-hateful sentences yield to higher accuracy scores than the hateful ones.
- They concluded that their study is a good starting point for discussion and further research.
- The most interesting part was for me the extra option of annotating sentences as Relation. I'm very curious if future studies are able to find if this class affects the classification task.