Subjectivity Mining: Assignment 1

Group AI-11

Matthias Agema[1] (2707650), Noa van Mervennée[2] (2710633), and Matteo de Rizzo[3] (2749303)

[1] Business Analytics, Vrije Universiteit, 1081 HV Amsterdam
m.j.agema@student.vu.nl
[2] Artificial Intelligence, Vrije Universiteit, 1081 HV Amsterdam
n.van.mervennee@student.vu.nl
[3] Artificial Intelligence, Vrije Universiteit, 1081 HV Amsterdam
m.de.rizzo@student.vu.nl

1 Introduction

There are a lot of factors that bolster the use of hate speech on online platforms. Anonymity and presence of communities with like minded users tends to polarization of thinking, with mutual support over behaviors that might be seen or are indeed offensive to individuals or to minority groups. Research has shown exposure to these hateful or threatening messages increase out-group prejudice [3], which in turn leads to further hateful behaviours, both in an online and offline setting. With the overwhelming amount of messages being posted on various platforms every second it is therefore crucial to figure out methods to automate hate speech detection, in order to mitigate the negative impact. This task has been looked upon by a multitude of parties, each developing different annotations methods regarding hateful comments. In this paper, three different annotation schema's were applied to a sample of 45 messages by two independent annotators, and consequently analyzed and compared.

1.1 Paper A: Kumar et al.

The annotations applied in the first paper yield: overt aggression (OAG), covert aggression (CAG) and non-threatening aggression (NAG). Overt aggression refers to any speech/text that is expressed openly and is considered to be aggressive. In contrast, covert aggression is an indirect attack against the victim and is often packaged as (insincere) polite expressions (through the use of conventionalised polite structures) [2].

1.2 Paper B: Zampieri et al.

The paper describes to use a hierarchical annotation schema split into three levels: offensive / not-offensive (A), the type (targeted / untargeted insult) (B), the targets (individual / group / other) (C). On the first level (A) there is a simple

distinction that determines whether the message contains offensive language / profanity, or if it does not. Following, level B determines if the profanity is directed to a specific individual or group, or if it is not directed to anyone, but rather a common use of profanity. Lastly, if level B indicated the message to be targeted, there is a further distinction based on the recipient of the insults (as stated earlier individual / group / other) [4]

1.3 Paper C: de Gibert et al.

Out of the three papers here analyzed, de Gibert and colleagues proposed the simplest method of annotation, which is simply a binary categorization, with the two options being $Hate$ or $NoHate$. While the other methods went deeper into the characteristics of the hateful messages (establishing a target or whether it's an overt or covert hostility), this one only determines if the overall meaning is hateful or not. [1]

2 Inter-Annotator Agreement

Two annotators tagged 44 tweets according to the guidelines of the three papers. The annotators independently tagged the tweets without prior discussion. In order to test the validity and efficiency of applying the annotated dataset for further experiments, the inter-annotator agreement score is measured for each set of labels. To calculate the inter-annotator agreement score, the Cohen's Kappa coefficient is pertained. Formula 1 expresses the equation to calculate Cohen's Kappa K-coefficient, whereas p_o refers to the relative observed agreement among annotators and p_e is the probability of a random agreement.

$$K = \frac{p_o - p_e}{p_e} \tag{1}$$

A score of 0 implies a random agreement among the annotators, whereas a score of 1 implies a complete agreement.

2.1 Paper A: Kumar et al.

The Cohen's Kappa coefficient for the tagged set CAG, NAG and OAG yields 0.28. This result is remarkably low, requiring certain changes to the annotation guidelines to continue further experimentation with this dataset. After discussing this result with the two annotators, it is concluded that the score can be explained by the differences in interpretations of tweets. For example, one annotator would assign the label non-aggressive (NAG) to a tweet, while the other interpreted it as an indirect attack (CAG).

2.2 Paper B: Zampieri et al.

In addition to the calculation of the inter-agreement score of the individual annotation levels (A, B, and C as introduced in section 1.2), the score for the combined levels is measured, seen in table 1.

Table 1. Cohen's Kappa coefficients

Tagset	K
Combined three-level-tagset	0.46
Offensive / Non-Offensive (A)	0.52
Target / No Target (B)	0.48
Group / Individual / Other (C)	0.49

All of the results are approximately below par. Resulting again to the conclusion that certain changes to the annotation guidelines should be made to continue further experimentation with this dataset.

2.3 Paper C: de Gibert et al.

The last calculated inter-annotator agreement score is relatively high in contrast to the previous measurements. The Cohen's Kappa coefficient score yields 0.72. Therefore, it is concluded that the definition of Hate / No Hate is approximately in line for the two annotators, although an improvement is desirable.

3 Error Analysis

3.1 Paper A: Kumar et al.

In pursuance of understanding the extremely low Cohen's Kappa (0.28), a confusion matrix was created (see Fig. 1) The graph shows a tremendous discrepancy in annotations between Noa's and Matthias'. The biggest difference resides in the interpretation of covert aggression. Indeed, while Noa categorized 17 messages as CAG, Matthias only attributed to this category 6 tweets. Some examples of messages that created this divergence are:

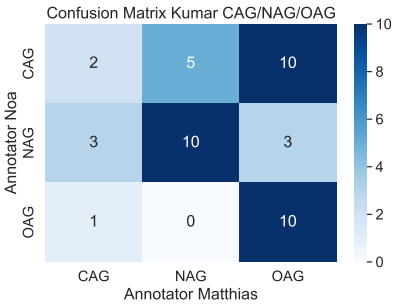

Fig. 1. Confusion table Kumar et al.

- "Indian government needs to stop blaming all it's self inflicted problems on it's neighbors. This sounds like the work of people angry at being occupied."
- "In JNU students is learning from @ArvindKejriwal how to speak from a negative.. Shame on you kejri sir no reaction on it #ShutDownJNU"
- "our President has no history like modi who kills thousand Muslims in gujrat"

All of these examples have been evaluated as OAG by one annotator, while CAG by the other. When considering the last item in the list, we can see how difficult the distinction between OAG and CAG ccan be. While this message could be indicate direct (overt) aggression towards modi, it is also true that it is written in a rhetorical form, which could make it appear as overt. As demonstrated by Noa and Matthias, the perceptive distinction between seeing one form of aggression as overt of covert often comes down to personal judgment, making this kind of annotation challenging to utilize.
Furthermore, a third of the time that a message was considered non-threatening, the other party considered it covert aggression. One of these cases is the following:

- "Thomas Kuruvilla Baven BJP is creating an image with illiterat hindus that their religion is in danger is only a vote gathring strategy"

Here we can see how different judgment was performed on this sentence. While Noa has seen this message as a covert aggression, Matthias perceived it as a non-threatening aggression, demonstrating once again the difficulty of recognizing the delicate differences between covert form of languages and others.

3.2 Paper B: Zampieri et al.

Figure 2 depicts the confusion matrix for the annotations of Matthias and Noa after following the guidelines of Zampieri et al. [4]. As previously discussed the followed annotation schema exists of three levels. At first, the class NonOffensive was classified for 9 cases through both annotators. Noa did three extra, where Matthias disagreed with. And vice versa Matthias annotated 6 other cases, where Noa assigned them as offensive. Some examples are:

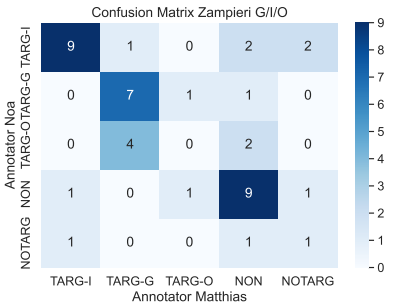

Fig. 2. Confusion table Zampieri et al.

– "This is disgusting"
– "Its not good signs....Webcast crashed...CNBC tvs 18 webcast playing like cat and mouse...awefull songs"
– "Reservation seats from different organizations. ..Families of such students thank God for their absence in their homes."

In the first two cases the words "disgusting" and "awefull" occur. Both words could be interpreted as non-acceptable language, but are in a gray area in terms of admissibility. In the third sentence there is a reference to 'God' in the message. According to Matthias this is profane language, while Noa did not interpret it this way. Error types on this level originates all through different interpretations.

One level down, in which distinction had to be made between untargeted and targeted insults, the annotators agreed at only one case. Where Matthias annotated a total of 4 tweets as untargeted, Noa did three. The untargeted class is only confused with the non offensive and offensive targeted at individual class. We already discussed some examples between the distinction of offensive and non offensive cases. Therefore, the next is example is labeled by both annotators as offensive, but untargeted by Matthias and targeted at individual by Noa.

– "@USER No fucking way he said this!"

This tweet is for sure offensive, because of the word "fucking". Noa interpreted it as a direct insult to the "he" in this post and therefore classified it as targeted at individual. In contrary, Matthias did not interpret it that way, which resulted in the untargeted label for him, because of the non-acceptable language.

At last, the third level, in which targeted insults or threats need to be distinguished between individual, group or others. There is only 1 confusion between the classes individual and group. This confusion concerns the following case:

– "Thomas Kuruvilla Baven BJP is creating an image with illiterat hindus that their religion is in danger is only a vote gathring strategy."

Matthias annotated this tweet as offensive targeted at the group "illiterat hindus", while Noa thought it was an offense, in the form of a reproach, to Thomas Kuruvilla Baven. Which means that in this case there are several possibilities for annotators to base their choice on. The most common confusion at this level has been made between the classes group and other. There are null cases they agreed to belong to the targeted at others class. That being said, in four cases Matthias has chosen the targeted at group class, while Noa chose the targeted at others class. Two of the four were offenses targeted at respectively the media, journalism and the Indian government, which are not groups that meet the requirements described in Zampieri et al. [4], which implies that Matthias was not right with his annotations. The other confusion is the following.

– "@USER Yes do hurt your selves gun control freaks."

In this case "gun control freaks" could be interpreted as both group and other.

3.3 Paper C: de Gibert et al.

In figure 3, the confusion matrix for the annotations of Matthias and Noa according to the guidelines of de Gilbert et al. [1] can be found. The relative high KAPPA score showed before that there was a high level of inter-annotator agreement. Overall there is a high class imbalance with only 7 tweets that are unanimously classified as *Hate*. Furthermore, only 4 of the 44 tweets were classified different. Those four tweets were all classified as *Hate* by Noa and *NoHate* by Matthias. Both annotators are relatively hesitant in assigning tweets as *Hate*. A cause of this is probably the clear point by point guideline, which states what a hateful tweet should consist of at least.

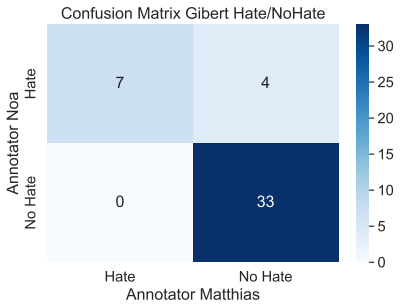

Fig. 3. Confusion table de Gibert et al.

According to the four tweets classified in different ways there, we suggest three possible causes of errors. Matthias was too conservative in annotating tweets as *Hate* or missed some interpretation as a result of which, according to him, the guidelines were not satisfied. Furthermore, Noa could also misunderstood the tweet as a result of which, according to her, the ingredients were complete for classifying the tweets as hateful. It can also happen that a tweet overall can be interpreted different. Three of the differently classified tweets are shown below:

– "People of Kerala will give befitting reply to these BJP idiots. They Are wiser than other northern States"
– "Terrorist"

– "I'm Assuming I'm Not Going to Get an Answer': Brainless #Liberal Stooge Ocasio-Cortez .@USER on $40T Policy Plan URL #WakeUpAmerica THIS is NOT America's future VOTE Republican or surrender America to these #DEM morons @USER #MAGA"

Noa annotated the first tweet as *Hate*, because of the "BJP idiots" part, what was according to her a deliberated attack motivated by the aspects of the group's identity. In contrast, it was not clear for Matthias what aspects of the group's identity were and therefore did not annotated it as *Hate*. The second tweet was according to Matthias not directed to a specific group of people, while Noa thought terrorists were the group of people. After discussing this one, the conclusion was drawn that this tweet should be labeled as *NoHate*. At last, "Brainless liberal stooge" in the third tweet was according to Matthias aimed on the individual Ocasio-Cortez, which is not directed to a group of people. Noa interpreted liberals as the attacked group which led her to classifying this tweet as *Hate*.

Most of the cases were unanimous classified as *Hate* or *NoHate*. An example of a hateful tweet is:

– "#Dutch people who live outside of #NewYorkCity are all white trash."

This tweet contains a deliberated attack to Dutch people who lives outside of New York city. The group of people is clearly motivated by aspects of the group. So all requirements for classifying this as *Hate* are fulfilled. Finally, an example of a clear non hateful case is the following:

– "He is coward"

This tweet is a deliberated attack, but is not directed to a specific group of people and therefore it cannot be classified as *Hate*. Both annotators did this correct.

4 Discussion

After receiving the various inter-annotator agreement scores, it was perceived that the two annotators acted with a dissimilar set of guidelines. The first score obtained was the Cohen's Kappa coefficient score that defines whether there is an agreement of the tagset Hate / No Hate. In comparison with the other scores, this score of 0.72 is relatively high. After the annotation procedure, the set of guidelines applied by each annotator was discussed. Both annotators applied the guidelines designated by Zampieri [4]. According to Zampieri's guidelines, hate speech is a:

(a) deliberate attack
(b) directed towards a specific group of people
(c) motivated by aspects of the group's identity.

Although the same set of guidelines were applied, many disagreements of tagged labels related to Hate / No Hate provoked various discussions. Several disagreements were predominantly the result of the misinterpretation of a tweet. Misinterpretation of a tweet could be due to the lack of understanding or the lack of context. Substantially, when the annotators applied different tags, the annotators were already in doubt about what label to apply. When in doubt, for further research it is recommended to either appeal for a third annotator (or more), or, an extra guideline should be followed that decides to label a doubtful tweet as, for example, No Hate.

After discussing the differences of both tagsets, the guidelines presented in the articles have been reviewed together. Both annotators indicated that they found it difficult to tag Covert Aggression, as it is difficult to recognize a rhetorical question or sarcasm without the sound of speech. For a second round of annotation, a new rule have been set to the annotation of these doubtful tweets: ask a third person for advise, if in consequence both persons are doubtful, mark it as No Hate, Non Offensive or Non-threatening-aggression. Therefore, the system is less likely to learn from misinterpreted/ambivalent tweets. The results of the second round for each individual tagset is presented in table 2.

Table 2. Cohen's Kappa coefficients

Tagset	K
Hate / No Hate	0.78
Offensive / Non-Offensive	1
CAG / OAG / NAG	0.93
Target / No Target	0.72
Group / Individual / Other	0.66

As seen in table 2 the coefficient scores have remarkably increased. The score of 1 for offensive/non-offensive implies a complete agreement. Questions can arise such as: is this a reliable annotation procedure or did the annotators steer too much in one direction which could result in a biased dataset? For future research it is therefore recommended to account for more annotators in order to prevent this bias.

Appendix

Contributions

- Matthias Agema - 3.2, 3.3
- Noa van Mervennée - Confusion matrices, 1.1, 2, 2.2, 2.3, 4
- Matteo De Rizzo - 1, 1.2, 1.3, 3.1

References

1. de Gibert, O., Perez, N., García-Pablos, A., Cuadros, M.: Hate speech dataset from a white supremacy forum. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). pp. 11–20. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/W18-5102, https://aclanthology.org/W18-5102
2. Kumar, R., Reganti, A.N., Bhatia, A., Maheshwari, T.: Aggression-annotated corpus of Hindi-English code-mixed data. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://aclanthology.org/L18-1226
3. Soral, W., Bilewicz, M., Winiewski, M.: Exposure to hate speech increases prejudice through desensitization. Aggressive Behavior **44** (09 2017). https://doi.org/10.1002/ab.21737
4. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Predicting the type and target of offensive posts in social media. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 1415–1420. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1144, https://aclanthology.org/N19-1144