Subjectivity Mining: Assignment 1 - Part 1

Group AI-11

Matteo de Rizzo[1] (2749303)

Artificial Intelligence, Vrije Universiteit, 1081 HV Amsterdam
m.de.rizzo@student.vu.nl

1 Kumar et al. - Aggression-annotated Corpus of Hindi-English Code-mixed Data [2].

1.1 Annotation guidelines

According to Kumar and colleagues, in order to annotate Verbal Aggression, the first step is distinguishing between *Overt* and *Covert* aggression. Overt (OAG) refers to any kind of message that explicitly uses lexical items that are considered aggressive, while covert (COG) masks aggressive behavior behind disingenuous polite words.
Once this distinction was done, the next categorizational step identifies the target of the aggressive expression. The four types are:

- Physical threat: contains threats of physical harm
- Sexual threat: portrays sexual acts in an aggressive way, or threats related to it
- Identity threat: aggressions towards gender, geographical location, political views, caste, religion or ethnicity
- Non-threatening Aggression: contains personal insults, or aggression towards one's choices (That do not adhere to any of the previously stated categories)

1.2 Motivation

The motivation for their paper was to safeguard users of the web, as incidents of aggression have exponentially grown, generating a lot of psychological dread, with extreme cases leading to suicides. Therefore, Kumar and colleagues tried to create a new annotation method for aggressive speech, that could then be used to automatically detect and strike these comments in the future.

1.3 Findings

Their first experiment using automatic identification of aggression using their annotation guidelines were not as satysfying as they hoped. Indeed, with the main distincion of OAG, CAG and NAG they only reached an F1 score of 0.70.

1.4 Conclusion

They concluded that classification of aggression is a very complex task even on the most basic level, and therefore it need further exploration

1.5 Interesting aspects

I found the attempt of categorizing aggression in covert and overt ways an ambitious method for automatic analysis of hate speech, as it is quite difficult for me to distinguish them in a lot of cases. It tends to be extremely subjective, and therefore difficult to automate.

2 Zampieri et al. - Predicting the Type and Target of Offensive Posts in Social Media [3]

2.1 Annotation guidelines

The annotation proposed by Zampieri and colleagues is based on a hierarchical system. The three steps, starting from broader to narrow are: Offensive Language Detection (A), Categorization of Offensive Language (B) and Offensive Language Target Identification (C). Level A distinguisher between Offensive (OFF) and Not Offensive (NOT). Level B determines if the offense is Targeted (TIN) or Untargeted (UNT). Lastly, if level B was OFF then there is a further distinction between Individual (IND), Group (GRP) or Other (OTH).

2.2 Motivation

The motivation for their paper was that previous work on identification of offensive messages focused of detecting specific types of offences, lacking a generality that could be applied to multiple types of offensive messages.

2.3 Findings

They generated a new dataset able to annotate many different types of offensive language, which seems to be the first dataset containing annotations of type and target of offensive messages. The Fleiss' Kappa for their experiment resulted in 0.83

2.4 Conclusion

They concluded that their dataset has shown to be somewhat effective when used in SVMs and neural networks, providing openings for future research and application.

3 Gibert et al. - Hate Speech Dataset from a White Supremacy Forum [1]

3.1 Annotation guidelines

In contrast to the previous papers,the guidelines proposed by Gibert and colleages are stripped down to the most basic level. The only distinction made in this paper are between $Hate$ and $NoHate$. Each of the two categories go in detail of the characteristics that determine whether a message is part of one group or the other.

3.2 Motivation

The motivation for this paper was to carefully describe the distinction between Hate an No Hate messages, in order to provide a solid form of annotation, as it is a very subjective topic.

3.3 Findings

The Inter Annotator Agreement shows an average just over 90% of agreed annotations, probably determined by the binary distinction that was performed.

3.4 Conclusion

They concluded that $NoHate$ messages were easier to agree on, and that the model that provided the best results is the LSTM classifier

3.5 Interesting aspects

The data used in this paper is all derived from $Stormfront$, a white supremacist forum, which is thefore a biased community, with a higher percentage of hateful comments than seen on other platforms.

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. 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