TxMM research idea summary
Mining of Misogyny Categories in Tweets

Evgeniia Martynova (s1038931)

November 17, 2019

Motivation

Examination of misogynistic language in online social media is a relatively new research area in computer science. Among all the studies related to hate speech and offensive language ([3] gives a good overview) there are only a few papers concentrated on misogyny specifically. Given that online misogyny is a common phenomenon and women constitute 50% of population, work in this area has high social value.

Research question

The recent studies were concentrated on automatic detection and classification of misogynistic language. I am going to look at this problem from another angle and check whether it is possible to extract different categories of misogyny with clustering. Also it will be interesting to see if the known misogyny categories contain subcategories.

Data set

I will use the data set created for Automatic Misogyny Identification (AMI) task [2] from EVALITA 2018 evaluation campaign of NLP. The data set contains just over 4000 tweets, each annotated with the following fields:

- **id** denotes a unique identifier of the tweet.
- **text** represents the tweet text.
- **misogynous** defines if the tweet is misogynous or not misogynous, 0 or 1.
- **misogyny_category** denotes the type of misogynistic behaviour. The defined categories are: stereotype, dominance, derailing, sexual_harassment, discredit, 0 (for non misogynous tweets).
- **target** denotes the subject of the misogynistic tweet; it takes value as: active (individual), passive (generic), 0 (for non misogynous tweets).

The data set is available for free, however it is distributed via private google group. Thus, to get the data, one need to request to join the group and get an approval from the administrators. Another data set that might be useful is the one created for SemEval 2019 Task 5 [1]. It contains the collection of tweets with hate speech against immigrants and women. The access to the data set must be requested via CodaLab.

Evaluation

To evaluate the clustering I will compare the results with the distribution of tweets into the categories defined by AMI organizers. For detection of possible subcategories I will need to evaluate the results manually and decide whether the tweets which are put into one cluster form a meaningful category.

References

[1] Valerio Basile et al. "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 54–63. DOI: 10.18653/v1/S19-2007. URL: https://www.aclweb.org/anthology/S19-2007.

[2] Elisabetta Fersini, Debora Nozza, and Paolo Rosso. "Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI)." In: *EVALITA@ CLiC-it*. 2018.

[3] Marcos Zampieri et al. "SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 75–86. DOI: 10.18653/v1/S19-2010. URL: https://www.aclweb.org/anthology/S19-2010.