Wrocław University of Science and Technology
Alicja Figas | Marcin Gruza | Daria Puchalska

Do we speak OFFENSIVE in social media?

Based on the survey run by Pew Research Center in 2017, 41% of Americans have been subjected to online harassment and 27% have decided not to post something online after witnessing the harassment of others.
We decided to check if cyberbullying, hate speech and similar behaviour might be a problem on Polish social media.

Dataset

We collected almost 100,000 posts with more than 500,000 comments from *wykop.pl, starting from November 2019 to November 2020.*

We gathered not only texts but also information about the post and users (author and receiver) metadata (gender, no. of posts, comments, followers, followings).

Frameworks & Models

We trained multiple uni- and multimodal models (based on text features, user attributes and network embeddings) to predict if the text was offensive.
The dataset was highly unbalanced (class ratio 1:9) but we managed to reach 72% in macro F1-Score.

Using the igraph tool we created a network of Wykop users based on their conversations to further analyze its characteristics.

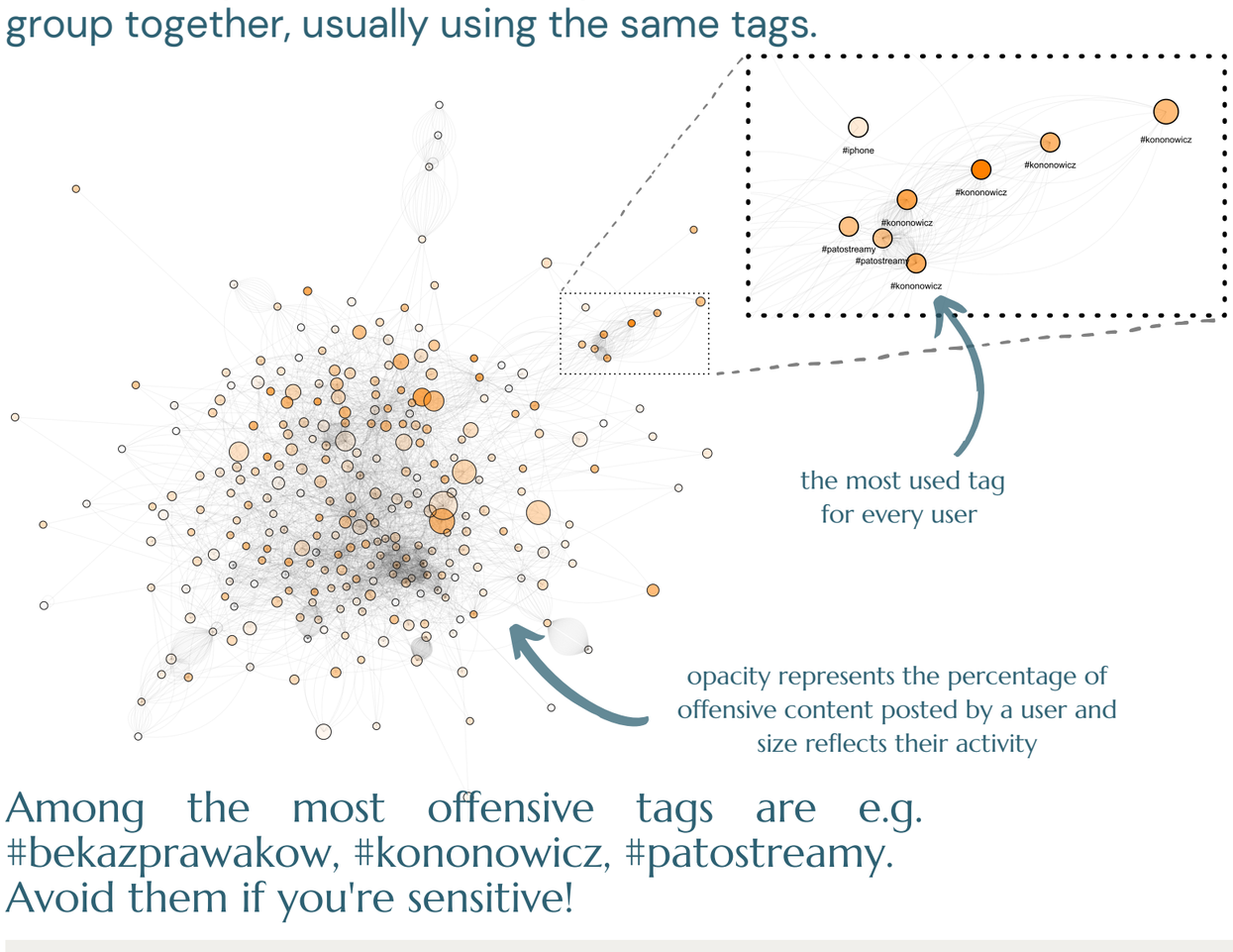
TensorFlow PyTorch igraph

Subjective problem

We manually annotated 6,000 texts deciding if the content was **offensive** or not. We had to take several discussion sessions to reach the satisfactory agreement level - this shows that the problem is very subjective and prediction models may need personalisation.

HateNetwork

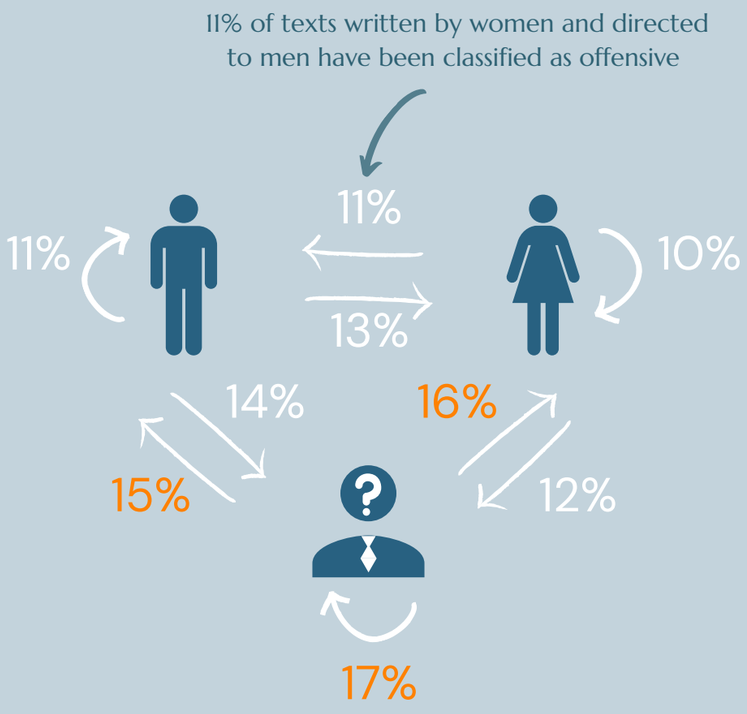
Users who publish offensive posts tend to group together, usually using the same tags.



Among the most offensive tags are e.g. #bekazprawakow, #kononowicz, #patostreamy. Avoid them if you're sensitive!

Who tends to offend others?

Applying our model to the entire dataset has shown that people who hide data about their gender are more likely to post offensive content in discussions with others. Moreover, texts written by men and directed to women are more often offensive than vice versa.



Seeking attention...

Our model confirmed that posts that contain offensive content are controversial - comments to them are also rude in **25%** of cases!

Offensive posts also get more attention. They get 76% more upvotes on average!

25% offensive comments to offensive posts

9% offensive comments to non-offensive posts

Final thoughts

- "What goes around comes around" - it is much more likely to get unpleasant responses when your post was also offensive.
- Detecting offensive content is an extremely difficult task and textual features may be insufficient to reach high classification quality - context is important.
- In general, **13%** of the texts have been found offensive by the model - that's a lot!

Try it out! • We created a tool where you can explore the results and check the offensiveness of your own texts: offensiveai.pl