DDMCS Project: Twitter and Mastodon Social Toxicity Analysis

Valentino Sacco, Camilla Savarese

January 2023

1 Abstract

Mastodon is free and open-source software for running self-hosted social networking services. It has microblogging features similar to the Twitter service, which are offered by a large number of independently run nodes, known as instances. Each user is a member of a specific Mastodon instance (also called a server), which can interoperate as a federated social network, allowing users on different instances to interact with each other.
Why talk about this? Because it's the most talked-about alternative now that Twitter is experiencing a particular period under Elon Musk's ownership.

Brian Lloyd, an editor at the Irish Web site entertainment.ie, belongs to mastodon.ie, a server focussed on Ireland with sixteen thousand active members said about Mastodon, compared with Twitter, "you don't feel the same level of hostility there and there's no "swathe of American bullshit to cut through"[1].

But will this really be the case? That is what we will try to find out.

Will it be true that Mastodon is designed to cultivate an environment very different from that of Twitter? In this project we show from the study of users having accounts on both social networks how Twitter is, generally speaking, more toxic than Mastodon but the latter, although advertised as a positive environment, contains more extreme behaviour. We also explore the possibility of echo-chamber[2] effects by studying toxic scores relationships between toxic and non-toxic accounts and their followers.
For more details on code and implementation we leave you to the GitHub Repository

2 Data

For this particular case study we make use of data scraped from approximately 370 independent users that have an account on both Twitter and MastodonSocial, which we identified by querying Twitter's V2 API asking for tweets that contain both a link aswell as the hashtags "#MastodonSocial" or "#MastodonMigration".
After extracting said users we proceed by scraping posts on both socials, which we then feed to Google's Perspective API[3] in order to measure each user's level of toxicity.

Really Perspective models provide scores for several different attributes:

- Toxicity
- Severe Toxicity
- Insult
- Profanity
- Identity attack
- Threat

We are going to consider all of them separately and also averaging the scores by computing their the mean.
Finally for the echo-chamber[2] analysis we extract top and bottom users in term of average scores and query both Twitter and Mastodon's APIs to extract their followers. We extract average scores for a sample of these accounts by scraping posts and feeding them to Perspective API.

3 Toxicity Analysis

3.1 Average Score

We decide to start with this score given by the average of the others since in this way we can get a more general overview.
In Figure 1 there are several graphical representations of the score, which aim to underline the differences in the distribution in the two different social networks. One the first row we have the boxplot with and without outliers, since they are present in quite significant way: 9% for Twitter and 12% for Mastodon. So the first thing that catches the eye is that the median score is higher on Twitter, but the distribution is also more symmetrical on this social network. On Mastodon, the median score is slightly lower but the distribution is shifted towards the third quartile, i.e. there are more users who reach an abnormal level of toxicity. This is also confirmed by the density approximation, in fact the peak for Mastodon is slightly higher but with a smaller frequency (fewer users); to put it in more statistical terms we can say that the average is higher for Twitter, but the maximum score (and also the third quantile) are higher on Mastodon.

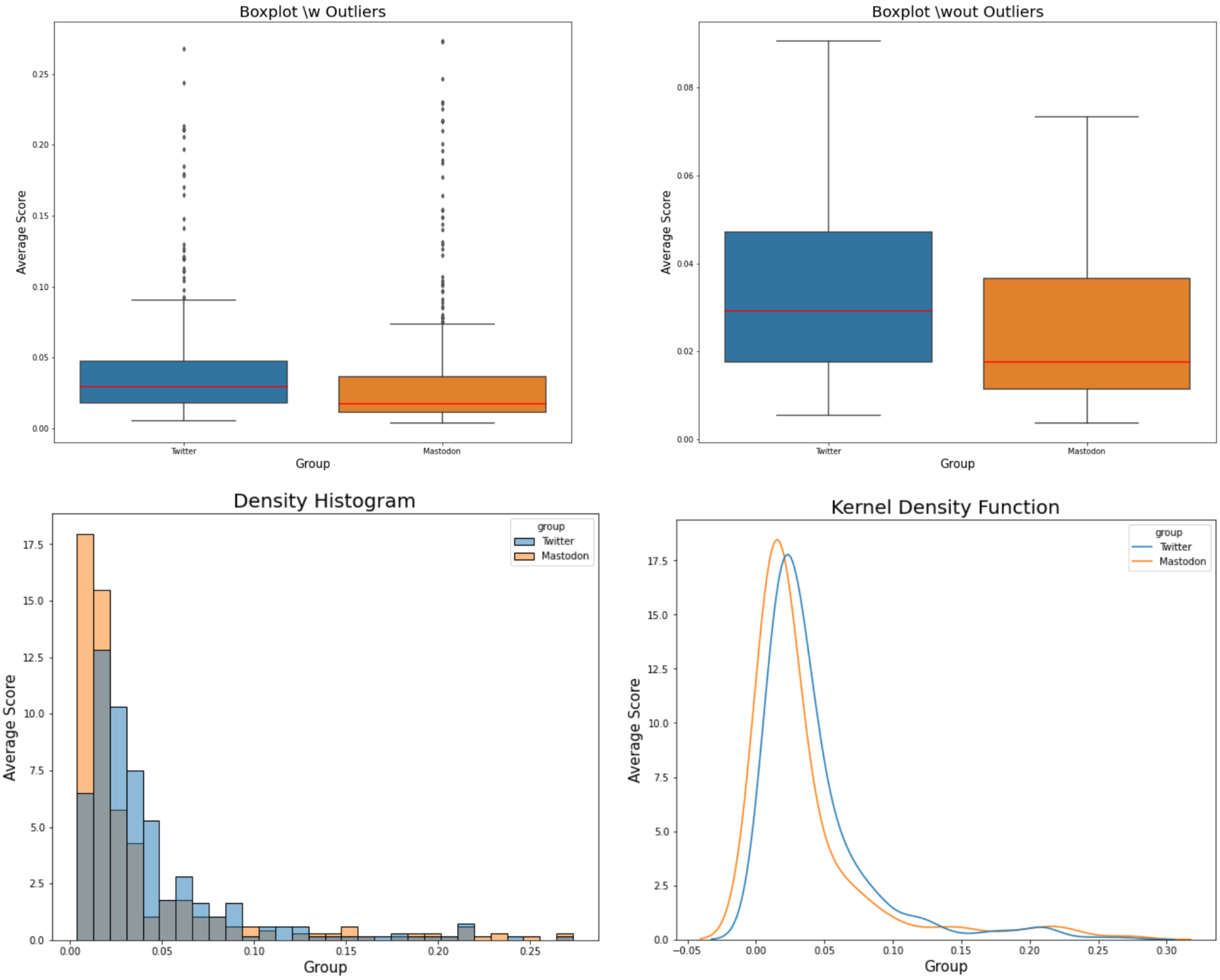


Figure 1: Average Score Distribution

We decided to extract the 5% of the users with the highest score so that we could compare them. Among the 19 users extracted, 14 are in common between the two social networks. They are plotted in Figure 2 and, as a confirmation of what we said previously, most of them are more toxic in Mastodon! This confirms the idea that with Mastodon the general trend is calmer, but a more limited circle reaches very high levels (statistically the distribution is by no means symmetrical).

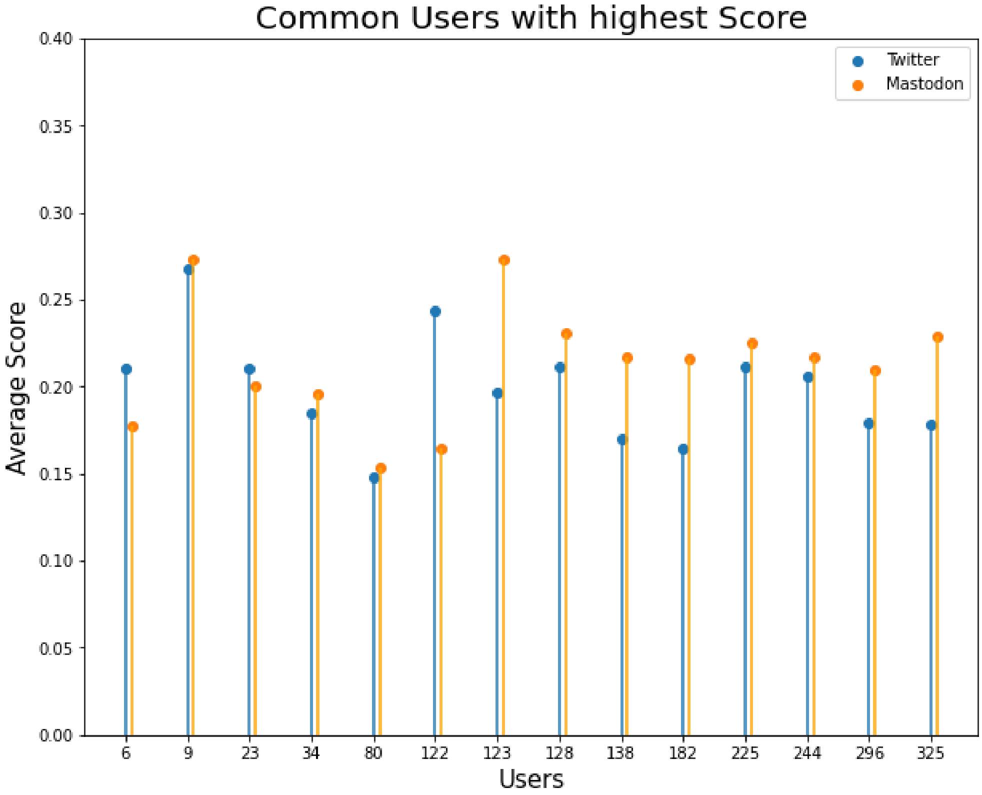


Figure 2: 95 Percentile most toxic Users

3.2 Separate Attributes

Now our goal is to analyze separately each attribute to understand which one is more discriminatory or exhibits anomalous behavior.

3.2.1 Toxicity

Description: *A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.*
In this case the outliers are the 6% for Twitter and the 9% for Mastodon, so also in this case they are more on this second social. Compared to the average score, here we are projected on decidedly higher values (practically double), so this is one of the most significant attributes. The two distributions differ more (again with respect to the "average case") and underline the fact that as an overview Twitter is more toxic, but Mastodon is more extreme (e.g. on Mastodon the maximum score is 0.5 and the kernel density has longer tails) We show this aspects in Figure 3.

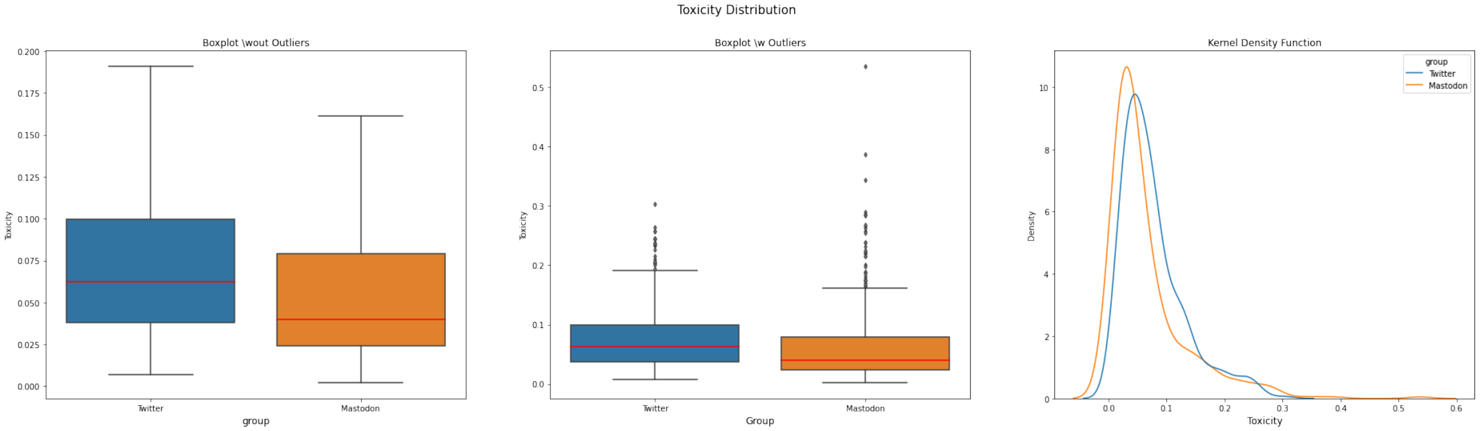


Figure 3: Toxicity

3.2.2 Severe Toxicity

Description: *A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words.*

Here the percentages of outliers are very high, 14% and 15% respectively! Even the overall range is not very wide and we remain on lower values than in the previous case; the distributions actually present a rather similar trend (both average and maximum value almost coincide and are lower than the average score, Figure 4).

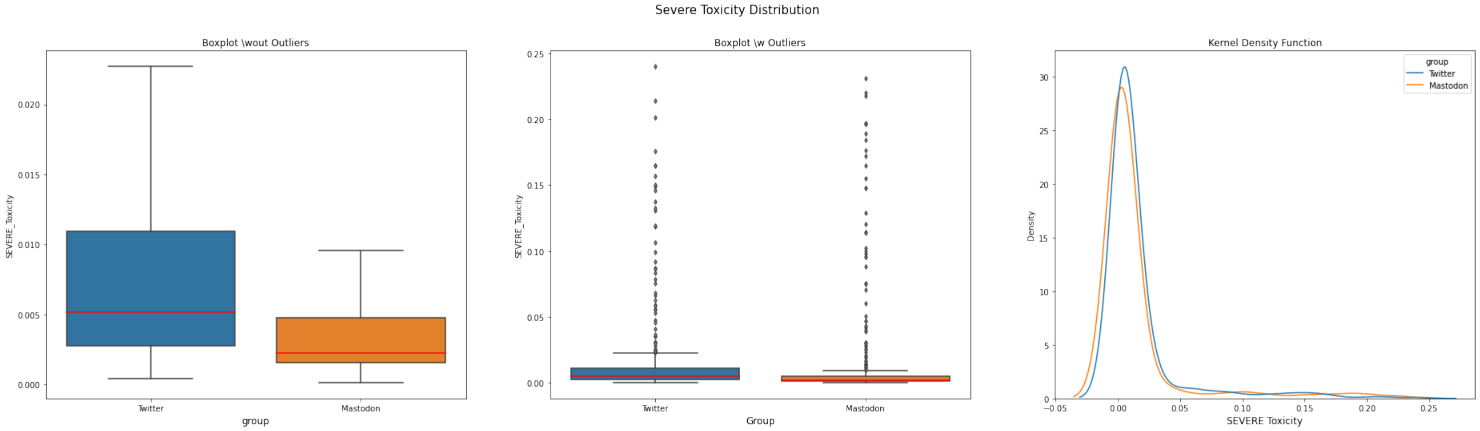


Figure 4: Severe Toxicity

3.2.3 Identity Attack

Description:*A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words.*

As shown in Figure 5, here the presence of outliers is very consistent: for Mastodon that's more than 17% and this is the reason why the two distributions are very similar if we consider all the data, but seems different in the first plot n the left. In this case the peculiarity is that in both the social the distributions are not symmetric and they are projected on the third-quantile (higher values). Moreover, on average we are below the "Average Score", but the maximum values are slightly higher.

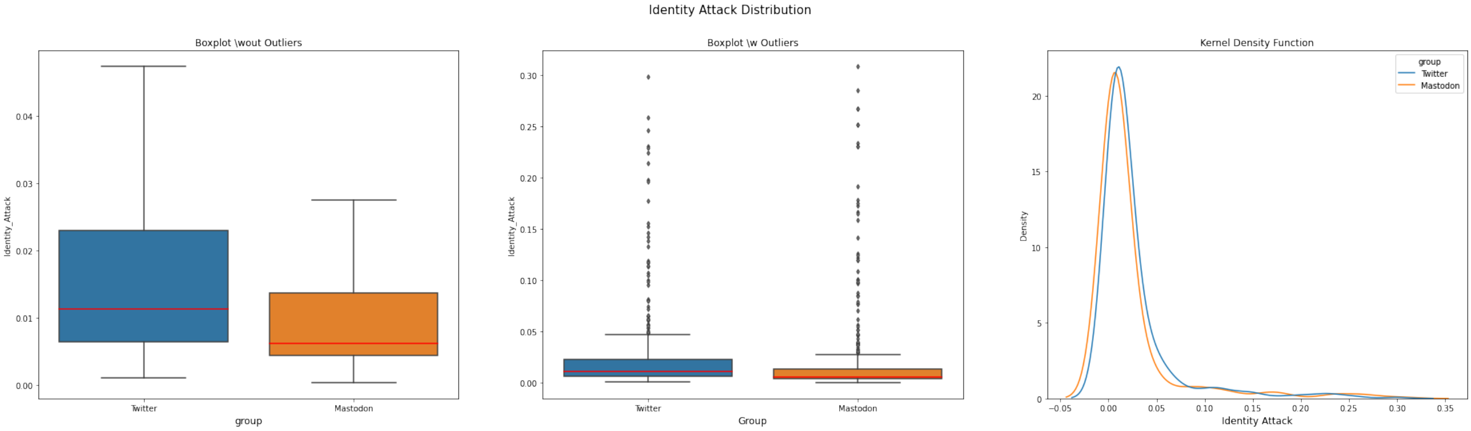


Figure 5: Identity Attack

3.2.4 Insult

Description:*Insulting, inflammatory, or negative comment towards a person or a group of people.*

This is one of the attributes for which the difference is bigger. Firs of all the outliers are just 8% on Twitter (less with respect to most of the others case) but they are 13% on Mastodon. The means are quite different,in the first case above the average score while in the second case below ; also with respect to the others attributes, distributions differ also graphically in all the three plots of Figure 6 (but they are also more symmetric).

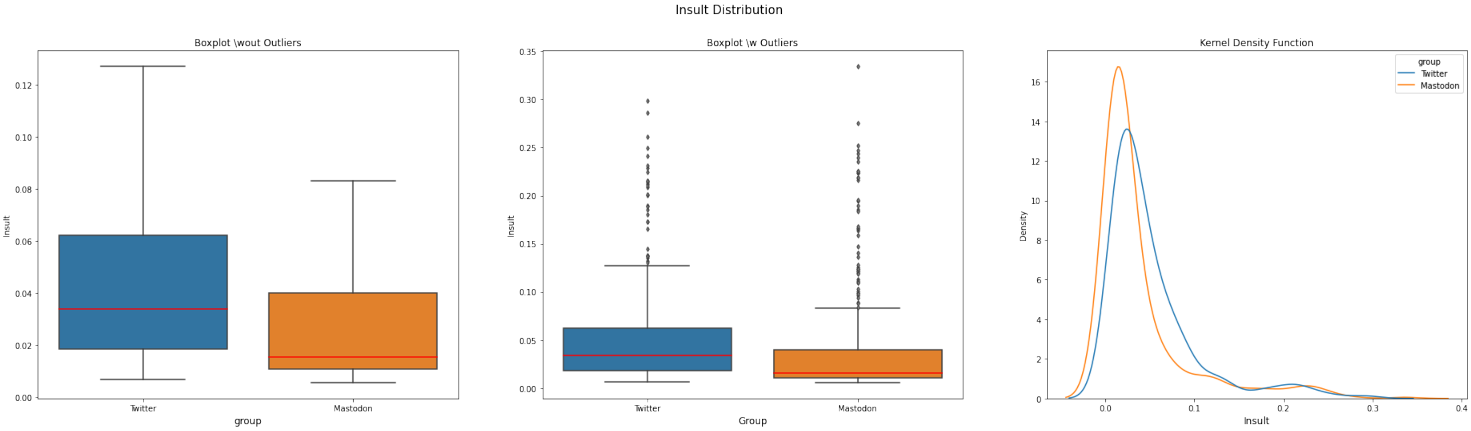


Figure 6: Insult

3.2.5 Profanity

Description: *Swear words, curse words, or other obscene or profane language.*

This attribute is interesting. There are very few outliers and the median are rather distant in the two social networks, as can be seen very well in the first plot of Figure 7. We also want to underline the fact that Twitter's extreme value is low and coincide with the one of the average score, but on Mastodon the maximum score is more than double the baseline we are considering, and it is the highest of the entire dataset (0.6 on a $0-1$ scale where the average in in general 0.04/0.05). If you take into account the fact that the third-quantile is 0.055, this is a really extreme value. In general we believe that it is an area that could show rather interesting social behaviors, above all because for some people it is a real taboo, for others it is a topic like any other that shouldn't cause "a stir".

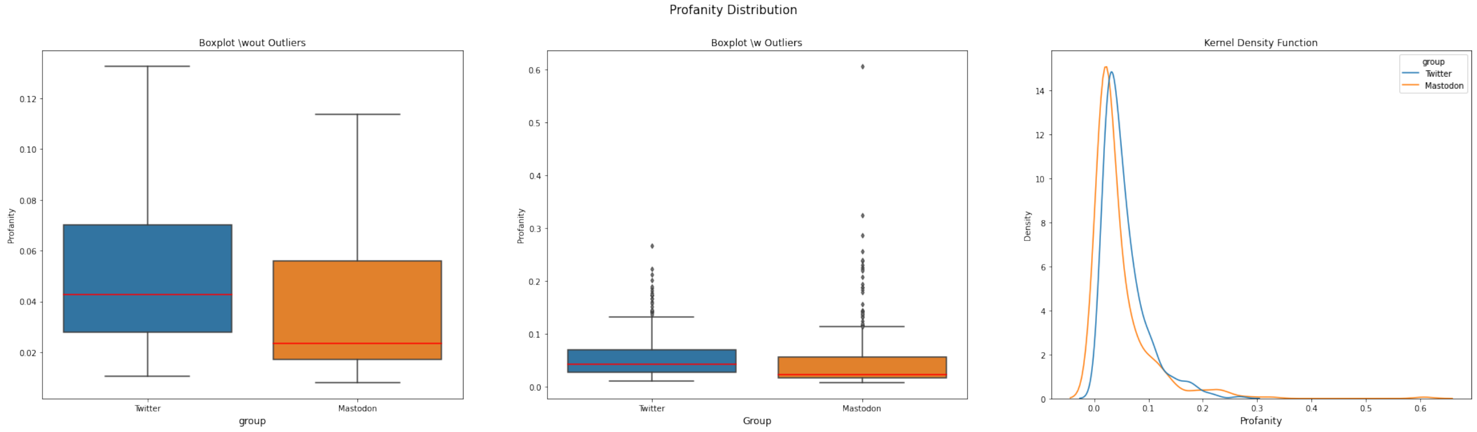


Figure 7: Profanity

3.2.6 Threat

Description: *Describes an intention to inflict pain, injury, or violence against an individual or group.*

This last attribute is definitely worth mentioning Starting from the outliers they are the 18th% on Mastodon (higher percentage of the dataset) and only the half on Twitter. Infact you can seen an huge difference in the two boxplot and the maximum value on Mastodon is 0.52, much higher than the same on Twitter and the average score. The two distributions are decidedly asymmetric and this means, with respect to the baseline, that initially the values are lower (e.g. up to the 2nd quantile), but then they "explode".

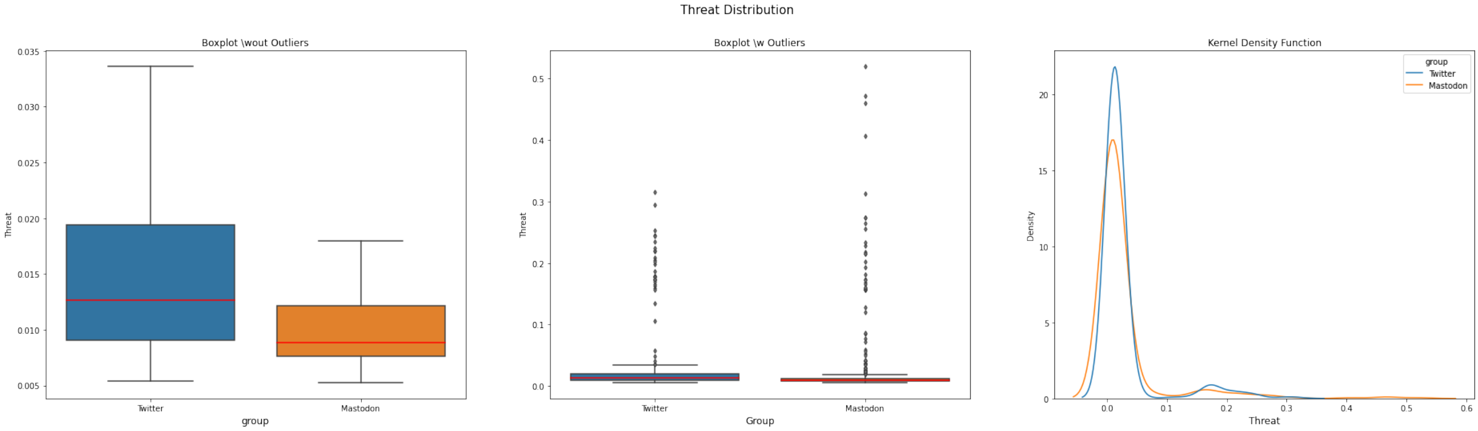


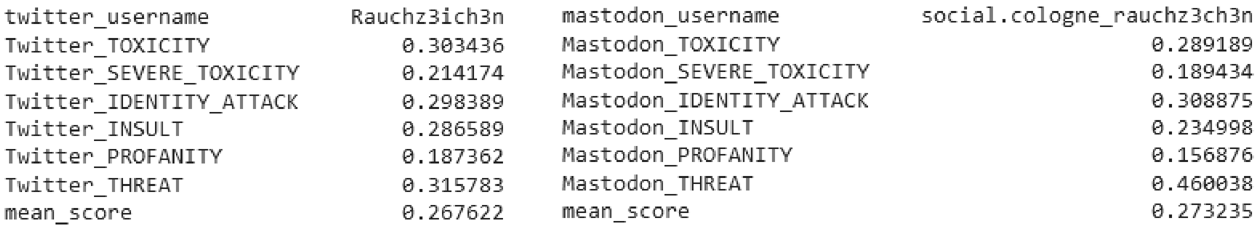
Figure 8: Threat

3.3 Common Users

At this point the idea is to extract the users with highest score for each attribute to compare them with each other and in the two different social networks. For each attribute and for each social, 19 users were selected: they represent the 5% of the highest values. First step is check how many are those in common for both social networks, which can be seen as an indicator of significant differences at the top of the distribution (including outliers):

Toxicity	7
Severe Toxicity	17
Identity Attack	15
Insult	9
Profanity	6
Threat	16

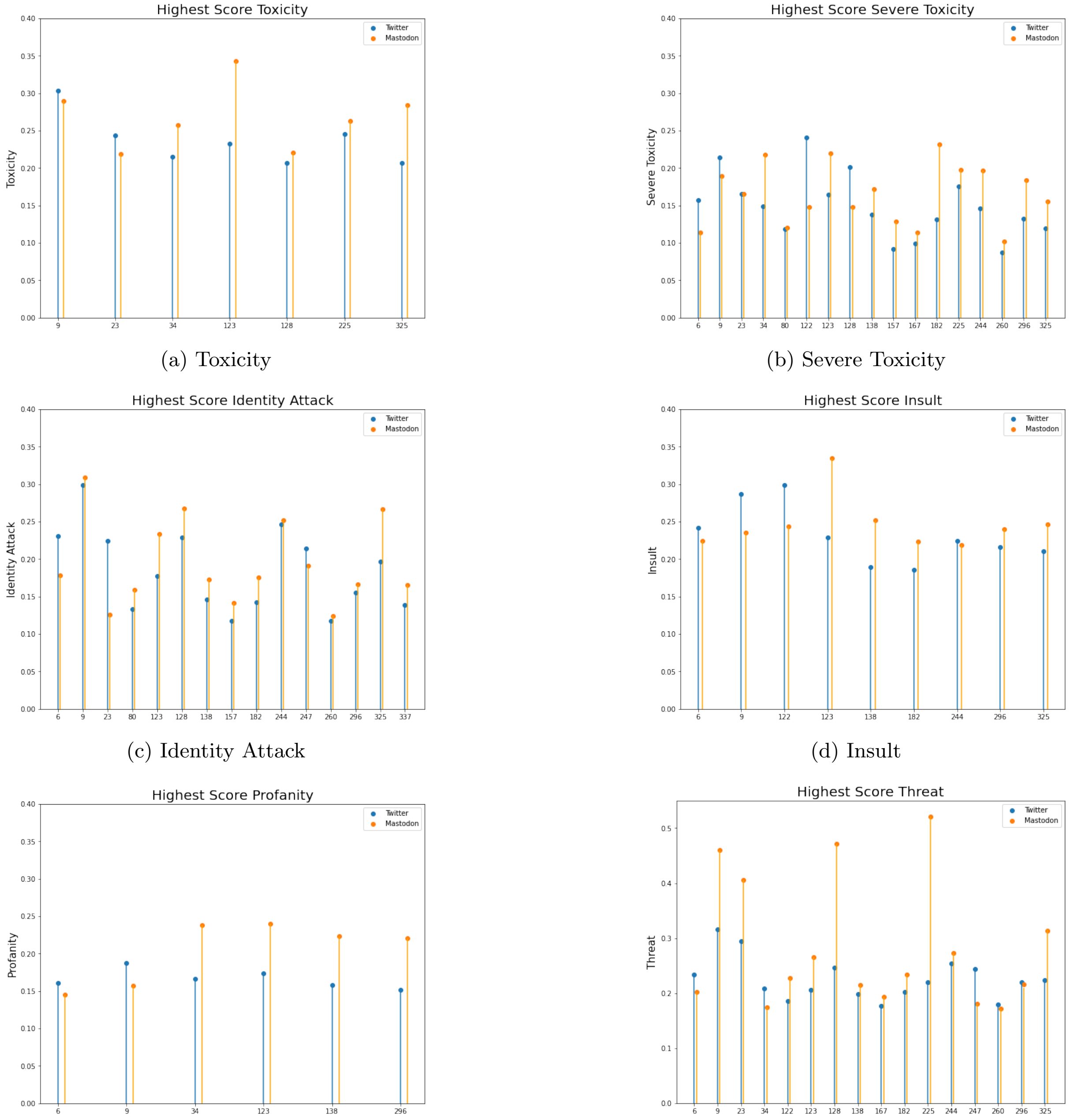
The minor intersections actually include the attributes that we believe are most significant and for which we have shown the greatest divergence: Toxicity, Profanity and Insult. Considering all the attributes only two users (number 9 and 123) are always in common, but there are other users like number 6 who are present in all attributes except one. To take a look, these are the scores of user 9, who has 3 highest attributes on Twitter and 3 on Mastodon, but on the latter the difference is greater, e.g. Threat (this explain the mean):



```
twitter_username             Rauchz3ich3n    mastodon_username                 social.cologne_rauchz3ch3n
Twitter_TOXICITY                 0.303436    Mastodon_TOXICITY                                   0.289189
Twitter_SEVERE_TOXICITY          0.214174    Mastodon_SEVERE_TOXICITY                            0.189434
Twitter_IDENTITY_ATTACK          0.298389    Mastodon_IDENTITY_ATTACK                            0.308875
Twitter_INSULT                   0.286589    Mastodon_INSULT                                     0.234998
Twitter_PROFANITY                0.187362    Mastodon_PROFANITY                                  0.156876
Twitter_THREAT                   0.315783    Mastodon_THREAT                                     0.460038
mean_score                       0.267622    mean_score                                          0.273235
```

Figure 9: User 9

Now we try to extrapolate a ”general trend” for the users with highest score in the two social.
It’s clear from Figure 10 that Mastodon’s score are higher for all the the attrbutes (more homogeneous situation for Insult) and this confirm the idea that on the new social network some people tend to let themselves go more and have more extreme behaviors.



(a) Toxicity

(b) Severe Toxicity

(c) Identity Attack

(d) Insult

(e) Profanity

(f) Threat

Figure 10: Highest Score Twitter Vs Mastodon

4 Followers Analysis

In this section because of technical reasons we could only perform our analysis with a small subset of data. This means that our aim is to show the emergence of certain behaviors which have already been highlighted in social networks and which certainly require a more in-depth study.

In particular we select the users with highest average score in common for Twitter and Mastodon and 15 among the users with the lowest scores. For each of them, the scores for all followers were then extracted, and finally we took into consideration the average for each category. We therefore want to analyze the relationship between the behavior of the individual user and his followers, and we have taken users in this way so as to have a ”polarized” situation in which we believe it is easier to see the results.

4.1 Toxicity

We show in Figure 11 for the variable ”Toxicity” what we said in the previous paragraph: actually there are two ”groups” of users.

In order to study the relationship user-follower we plot the toxicity of each users versus the average of his followers in the social, Figure 12.

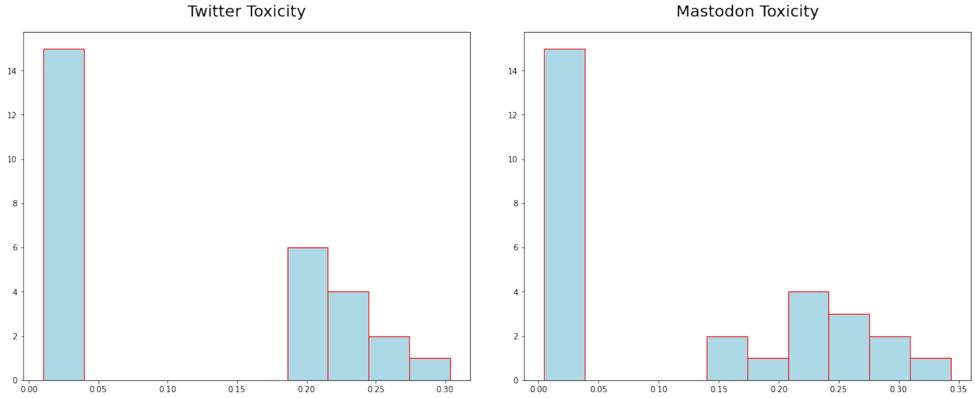


Figure 11: Toxicity Histogram

First of all we can observe that the division in two ”groups” remains in both social with a with a slightly greater dispersion on Mastodn. We don’t have the diagonal that characterize the Eco-Chamber effect; for users with lower value of toxicity a small group matched, but the others did not. Same for users with highest values, for Twitter a sort of diagonal could be hypothesized, but in general the situation is quite varied. Furthermore, these plots show us that the two social networks do not have particular differences.

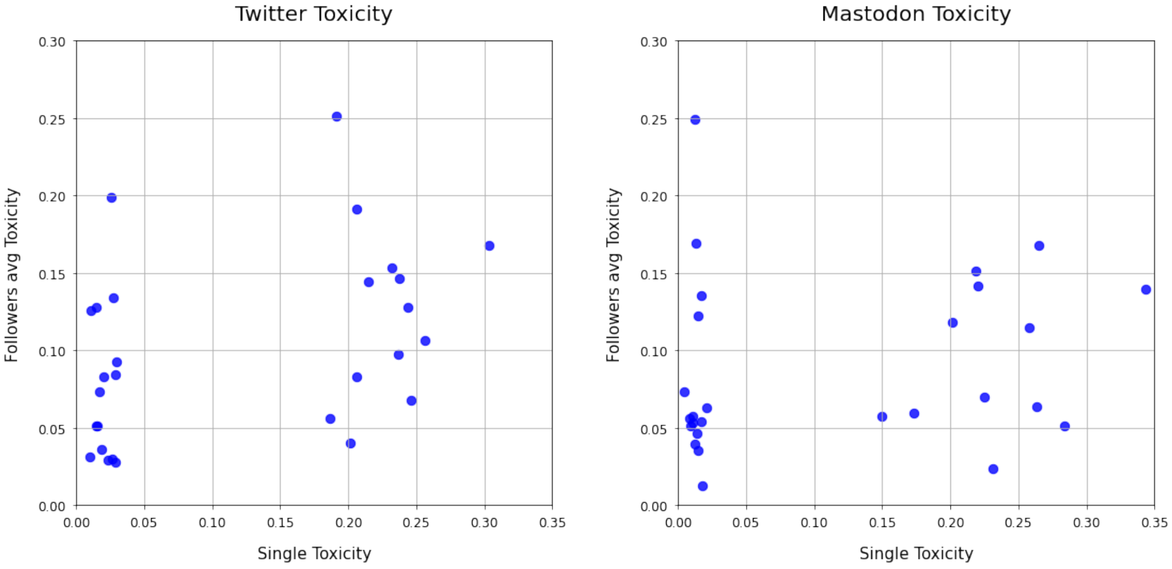


Figure 12: User VS Followers

Lastly, to see if the group structure is present also in the followers and if there are significant differences in the two social networks, we are going to fit a Gaussian Mixture Model.

First of all we compute the mean of the ”true group” that we have (and we know them because we chose the users) so that we can check if the results are consistent enough:

1. Average Group Highest scores Twitter: 0.15
2. Average Group Highest scores Mastodon: 0.151
3. Average Group Lowest scores Twitter: 0.056
4. Average Group Lowest scores Mastodon: 0.05

The GMM is fitted (separatly in the two social) with 2 as number of components.
Figure 13 shows the predicted label, that seems to better reconstruct the division for Twitter.
In both the social the mixture of the two gaussian have similar weights ([0.43930232,0.56069768] Twitter, [0.54023215,0.45976785] Mastodon). This reflects the fact that the two user groups have similar cardinality.
Also, the mean of the gaussian, coincide with the one of the data for both socialin the second group (Twitter 0.053, Mastdon0.055) while the ones of highest scores are lower (Twitter : 0.137, Mastodon 0.13).

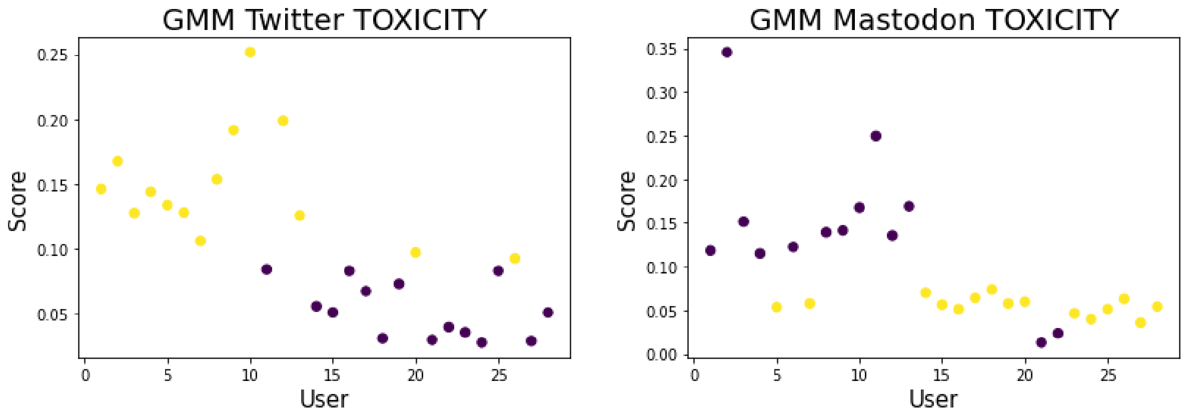


Figure 13: GMM Predicted Label

Finally we superimpose the two distributions estimate by GMM (so for the two group) with the one estimate as kernel density. From Figure 14, we can say that the GMM, overall obtains good results: in both social networks the peak of the kernel density is on slightly lower values, but the tail of the distribution is long.
To conclude, even in this ”imposed polarization”, we can't we cannot refer to a real Eco Chamber effect (which perhaps would appear using more data) but there is certainly a correlation between a user and his followers, and the fact that a sort of division into two groups always exists confirms this.

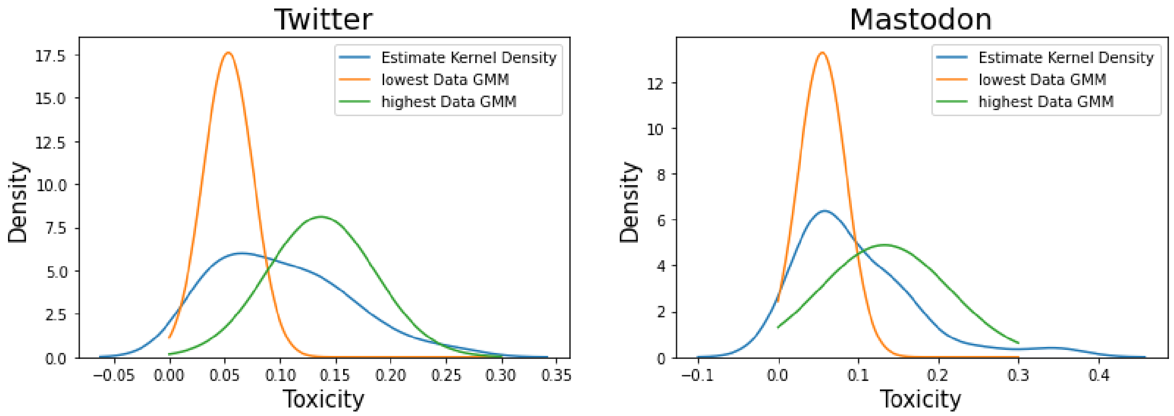


Figure 14: GMM Predicted Label

4.2 Other Attributes

We repeat the same steps for the others attributes and we give in this section only an overview that aims to underline any important features.

Firstly, Figure 15 shows that we have a polarized situation for both social for each attribute, indeed on Mastodon in various cases we could almost identify more than two groups.
Considering Figure 16, where we have the score of each user versus the average of his followers, the situation is similar to before : for low values the followers have higher scores than the single (vertical line) while for the high values, albeit with a lot of dispersion, a sort of Eco Chamber style diagonal appears (but our data is too few).

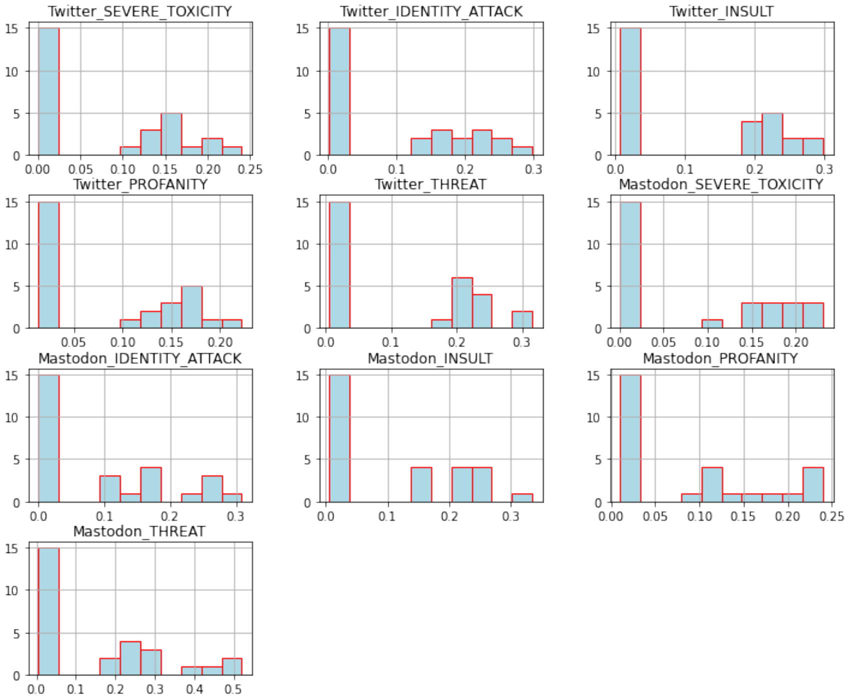


Figure 15: Histograms

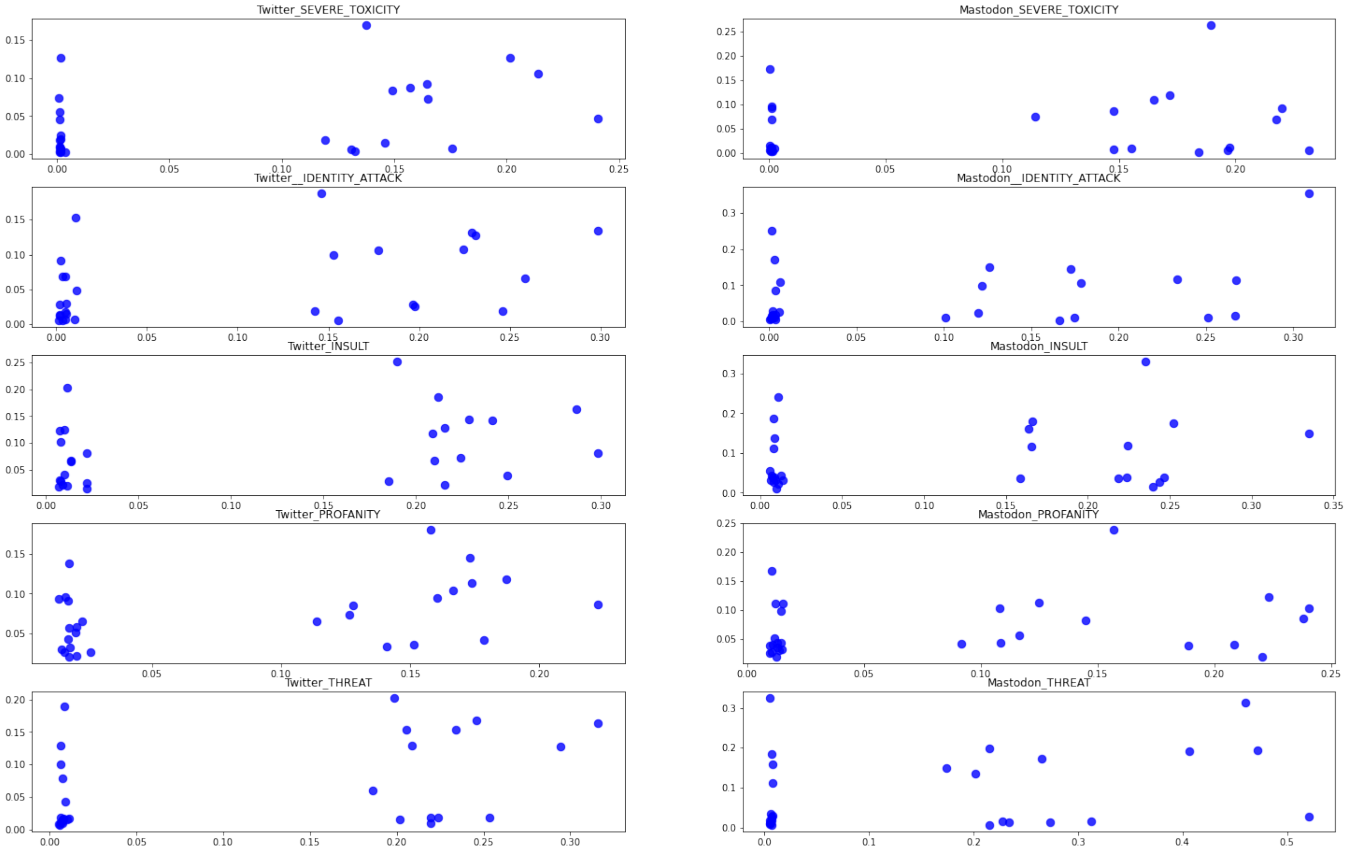


Figure 16: Single Users VS Followers

5 Conclusions

We believe this to be a good starting point for a deeper analysis. From what we have shown there are similar scores on both social networks, having higher toxicity on Twitter rather than Mastodon which instead has a larger number of extremes scores; this highlights the possibility of fueling toxicity on inclined users.

Regarding the relationship between users and their followers (echo-chamber[2]) we can notice from this study how non-toxic users have a much heterogeneous following, exposing their ideas to both calm and aggressive users. This is not the case for toxic users which we displayed have a following that reflects their attitude by having scores that are very similar to theirs. This happens for both Social Networks, which highlights more of a human behaviour feature rather than a Social Network structure.

Lastly we'd like to point out that the latter study has been influenced by the small computational power at our hands, mostly for the followers extraction and the study of their posts as we were forced to sample both accounts aswell as publications to be able to conduct this study in time. Sampling this resources exposes this study to selection bias, therefore we believe that running the study on large scale would be needed to confirm our results.

References

[1] Kyle Chayka, What Fleeing Twitter Users Will—and Won't—Find on Mastodon, https://www.newyorker.com/culture/infinite-scroll/what-fleeing-twitter-users-will-and-wont-find-on-mastodon

[2] Cinelli M. De Francisci Morales G. Galeazzi A. Quattrociocchi W. Starnini M. (2021). The echo chamber effect on social media.Proceedings of the National Academy of Sciences of the United States of America, 118(9), e2023301118. https://doi.org/10.1073/pnas.2023301118

[3] Using machine learning to reduce toxicity online, https://perspectiveapi.com/