Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)

Mastodon Content Warnings: Inappropriate Contents in a Microblogging Platform

Matteo Zignani, Christian Quadri, Alessia Galdeman, Sabrina Gaito, Gian Paolo Rossi
Computer Science Department, University of Milan, Milan, Italy
matteo.zignani@unimi.it, christian.quadri@unimi.it, alessia.galdeman@studenti.unimi.it
sabrina.gaito@unimi.it, rossi@di.unimi.it

Abstract

Our social communications and the expression of our beliefs and thoughts are becoming increasingly mediated and diffused by online social media. Beyond countless other advantages, this democratization and freedom of expression is also entailing the transfer of unpleasant offline behaviors to the online life, such as cyberbullying, sexting, hate speech and, in general, any behavior not suitable for the online community people belong to. To mitigate or even remove these threats from their platforms, most of the social media providers are implementing solutions for the automatic detection and filtering of such inappropriate contents. However, the data they use to train their tools are not publicly available.

In this context, we release a dataset gathered from Mastodon, a distribute online social network which is formed by communities that impose the rules of publication, and which allows its users to mark their posts inappropriate if they perceived them not suitable for the community they belong to. The dataset consists of all the posts with public visibility published by users hosted on servers which support the English language. These data have been collected by implementing an ad-hoc tool for downloading the public timelines of the servers, namely instances, that form the Mastodon platform, along with the meta-data associated to them. The overall corpus contains over 5 million posts, spanning the entire life of Mastodon. We associate to each post a label indicating whether or not its content is inappropriate, as perceived by the user who wrote it. Moreover, we also provide the full description of each instance. Finally, we present some basic statistics about the production of inappropriate posts and the characteristics of their associated textual content.

Introduction

In little over a decade, online social networks and social media have come over almost every aspect of the life of billion people all over the world. Undoubtedly, all these social platforms have facilitated the communications and broadly increased the audience any sort of message can reach. In fact, people have a large number of online communities available to express, discuss and exchange any kind of information, from their beliefs and views on various topics to artistic expressions. Beyond countless other advantages, this democratization and freedom of expression (Gayo-Avello 2015) is also entailing the transfer of unpleasant offline behaviors to the online life. It is becoming increasingly clear that these platforms open space for contents and conversations which often become inappropriate and hurtful towards the online communities that these systems are supporting. This negative phenomenon is further supported by a recent statement by the Facebook VP of product management saying that the Facebook community "submit tens of millions of reports a week about potentially objectionable contents"[1]; and it is becoming more urgent since children and adolescents easily access to social media but they are more susceptible to these contents and more exposed to threats such as cyberbullying (Dinakar, Reichart, and Lieberman 2011), online harassment or sexting (O'Keeffe, Clarke-Pearson, and others 2011).

To defeat these threats, social media providers, first and foremost Facebook, are developing AI systems for automatically identifying inappropriate contents (Yenala et al. 2017) on the basis of the large dataset of content moderation they have. However, this kind of data are not publicly available to the researchers' community, which in the last years has been very active in providing tools for the automatic identification of specific types of inappropriate contents such as hate speech (Davidson et al. 2017; Vigna et al. 2017; Ribeiro et al. 2017), sexually explicit texts (Jha and Mamidi 2017) or images, offensive language (Xiang et al. 2012) or racism (Founta et al. 2018).

In this respect, we aim at providing the community with a dataset that could help carrying on the research on this topic. With this in mind, our attention has been directed to a decentralized social network whose features are well suited to the purpose: Mastodon, a new and fast emerging decentralized microblogging platform. In particular, it exhibits the following two services: *i)* that impose clearly stated rules of publication and censorship which must be accepted by the user who wants to register; and *ii)* it enables the user to publish an alert if she thinks that the post she is publishing may be perceived as inappropriate by any reader. Thus, we release a dataset gathered from Mastodon consisting of two

[1] https://www.fastcompany.com/40566786/heres-how-facebook-uses-ai-to-detect-many-kinds-of-bad-content

elements: *i)* all the posts with public visibility published by users hosted on Mastodon servers, namely instances, which support the English language; and *ii)* the policy, the code of conduct and the prohibited contents of each instance. Since Mastodon comes with a built-in function which allows users to mark their posts as “inappropriate” or “sensitive”, we associate to each item a label indicating its appropriateness. Further, the definition of inappropriate – “Something that is inappropriate is not useful or suitable for a particular situation or purpose.”[2] – asks for a context in order to distinguish what is suitable. In our case the context is provided by the instance meta-data and we assume that users are aware of the policy of the instance they belong to, as they choose it at the registration time. The latter characteristics represent a novelty and a change of paradigm in the today’s social media ecosystem since they give the responsibility of publication and of a possible censorship back to the users themselves.
Effectively, the dataset have been collected by implementing an ad-hoc spider for retrieving the timelines of the instances which form the Mastodon platform and their meta-data. The overall corpus contains over 5M posts, spanning the entire life of Mastodon. The DOI associated to the dataset, hosted at https://dataverse.harvard.edu, is 10.7910/DVN/R1HKVS.

Dataset

Mastodon characteristics. Before delving into the details about the released dataset and the methodology we used to collect it, we briefly summarize what is Mastodon and the features we exploit to get and label inappropriate contents. Mastodon is a decentralized online social network with microblogging features, where each server runs open source software. The main aim of the project, which dates to 2016, is to restore control of the content distribution channels to the people by avoiding the insertion of sponsored users or posts in the feeds. Due to its decentralized architecture, Mastodon is oriented towards small/medium communities, called instances, which explicitly specify the topics their users should be interested in. To this aim, each instance declares the contents which are not allowed and the spirit of the community it supports through a full description and a list of allowed topics. All these pieces of information are available at the registration time, so that users are aware of the instance policy and whether or not a content is inappropriate for a given community. In fact, the community-orientation strongly impacts the moderation procedures. For instance, one of the most popular European instances, “mastodon.social”, bans contents that are illegal in Germany or France, including Nazi symbolism and Holocaust denial. That according to the idea of the Mastodon’s founder, i.e. small and close communities would defeat unwanted behavior more effectively than a centralized solution based on an operation team screening controversial contents.

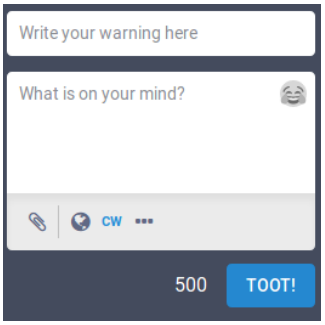


Figure 1: The box to compose a toot with a content warning.

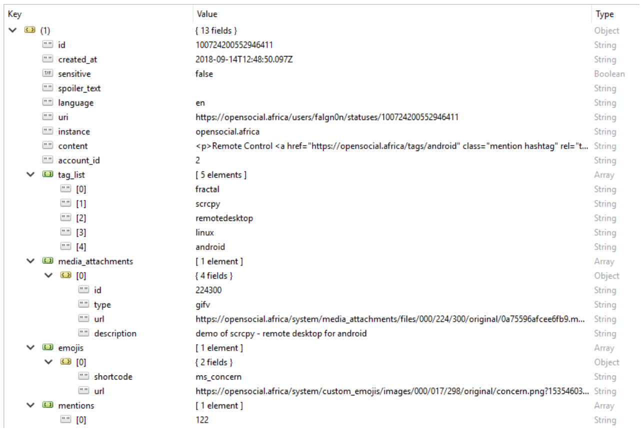
Even though Mastodon is built around the concept of instance, users belonging to a specific instance can communicate with users on other instances, as well. In fact, as every microblogging platform, the most basic way to interact in Mastodon is the creation of a textual content, also called ”toot”. A toot, or post, cannot exceed the 500 character length. Each toot has a privacy option, and users can choose whether the post is public or private (Salve, Mori, and Ricci 2018). Public messages can be aggregated and displayed on the federated timeline – a real-time wall which collects the public contents coming from all the Mastodon instances –, while private messages are shared on the timelines of the user’s followers only. A third timeline, namely the local timeline, shows messages from users hosted on a singular instance only. It is to note that Mastodon only allows a chronological ordering of the timelines, thus avoiding any ranking mechanism based on advertisement or other recommendation algorithms. Finally, for the purpose of this paper, we describe a further feature regarding the publication of toots. When a user wants to post something that should not be immediately visible or that may potentially result inappropriate for the community, she can “hide” it behind a content warning, as shown in Figure 1.

By clicking on the “CW” button, a user can enter a short summary of what the ”body” of her post contains, namely a spoiler-text, and the full content of her toot. Automatically, the system marks this toot as “sensitive” and only shows the spoiler-text in all the timelines. We exploit this latter feature to build our released dataset. This way the toots are labelled by the users, and we assume that they are aware of the policy of the instance and aware of what is appropriate or not for their community.

Collection Methodology Here we describe the collection methodology of the two main elements of our dataset: *i)* the instance meta-data and *ii)* the local timelines of all the instances which allow toots written in English.

The collection of the instance meta-data has become a straightforward process since Mastodon developers introduced an API to query different kinds of information about the instances. In fact, although they are based on a registration procedure where instance administrators have to subscribe to them for being inserted into the query results, APIs provide a lot of information about the instances. Specifically, we are interested in the full description of each instance and the list of the allowed topics. From our viewpoint, these two fields contain the information related to the context which makes a post inappropriate or not. We query these data through the API endpoint

[2] https://www.collinsdictionary.com/dictionary/english/inappropriate



Key	Value	Type
(1)	{ 13 fields }	Object
id	100724200552946411	String
created_at	2018-09-14T12:48:50.097Z	String
sensitive	false	Boolean
spoiler_text		String
language	en	String
uri	https://opensocial.africa/users/falgn0n/statuses/100724200552946411	String
instance	opensocial.africa	String
content	<p>Remote Control <a href="https://opensocial.africa/tags/android" class="mention hashtag" rel="t...	String
account_id	2	String
tag_list	[5 elements]	Array
[0]	fractal	String
[1]	scrcpy	String
[2]	remotedesktop	String
[3]	linux	String
[4]	android	String
media_attachments	[1 element]	Array
[0]	{ 4 fields }	Object
id	224300	String
type	gifv	String
url	https://opensocial.africa/system/media_attachments/files/000/224/300/original/0a75596afcee6fb9.m...	String
description	demo of scrcpy - remote desktop for android	String
emojis	[1 element]	Array
[0]	{ 2 fields }	Object
shortcode	ms_concern	String
url	https://opensocial.africa/system/custom_emojis/images/000/017/298/original/concern.png?15354603...	String
mentions	[1 element]	Array
[0]	122	String

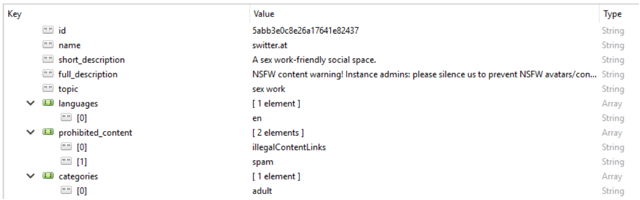
Figure 2: An example of toot in JSON format.

`instances.social/api/1.0/instances/list ?count=0`, which returns an array of JSON objects providing the data about the instances matching our criteria. In fact, we apply a filter to retrieve all the instances which include English as language for writing posts. In total, we release information about 363 instances.

As for the retrieval of the local timelines we developed an ad-hoc spider through the Scrapy framework [3]. The spider exploits the instance list obtained from the previous step and makes a pool of requests to the instance endpoint[4] which returns the latest toots of the local timeline. Since the timelines implement a pagination mechanism, the spider extracts the URL for the next request and repeat this procedure till it reaches the end of the timeline. As in the case of instance meta-data, each toot is represented by a JSON object which contains information about both the post and the user who wrote it. In the case of toots, we collected 5,877,355 items.

Data format and pre-processing Since both toots and instance meta-data are formatted according to the JSON specification we do not need to perform a strong data pre-processing to make the elements of our dataset compliant to the FAIR data principles (Wilkinson et al. 2016). In fact, the JSON format guarantees *i)* a high interoperability, since it can be processed with great ease by many different tools during data analysis; *ii)* the findability, as we associate to each object in the dataset a unique identifier; and *iii)* the accessibility, since we provide a meta-data descriptor available in JSON format. The meta-data file contains a brief description of the fields we did not discard from the original representation provided by the Mastodon API, or that we introduced to speed up further data analysis. For instance, as for toots, we created the fields “created_at_date” and “tag_list”. The former refers to a Date object built from the string date contained in the field “created_at”, and the latter is created from the field “tags”, i.e. it is an array of tags. We also added the field “instance” which contains the name of the instance – the hostname – from which the toot has been sent. Finally, each toot provides the fields related to the inappropriateness of its content, namely the entries “sensitive”, “content”, “spoiler-text” and “language”. The boolean field “’sensitive” indicates whether or not the author of the toot thinks that the content is appropriate. If the toot is inappropriate, the field is set up to “True” and the field “spoiler-text” would contain a brief and publicly available description of the content. The field “language” contains the language used to write the toot, while the “content” contains the complete text, including HTML tags. In Figure 2 we report an example of toot in JSON format.

As for the instance meta-data, we only kept the information which provides a context of the instance and the community it supports. In Figure 3 we show an example of instance meta-data. All the remaining fields of the original JSON response have been discarded.



Key	Value	Type
id	5abb3e0c8e26a17641e82437	String
name	switter.at	String
short_description	A sex work-friendly social space.	String
full_description	NSFW content warning! Instance admins: please silence us to prevent NSFW avatars/con...	String
topic	sex work	String
languages	[1 element]	Array
[0]	en	String
prohibited_content	[2 elements]	Array
[0]	illegalContentLinks	String
[1]	spam	String
categories	[1 element]	Array
[0]	adult	String

Figure 3: An example of instance meta-data in JSON format.

Ethical Considerations Through the above methodology we are able to collect a large amount of data containing different kinds of content produced by individuals from around the world. This fact rises some considerations about the privacy of the Mastodon users, that must be taken into account. In particular, the JSON response about a toot contains plenty of information about the user who has published the post. Since the Mastodon user may be unaware of their data being public and reusable for research purposes we disposed of the information about the users and we fully anonymized them by hashing the Mastodon user identifier. This latter aspect might limit the reuse and the fusion of this dataset with the topology of the Mastodon social network. To overcome these limitations and to integrate our dataset with the dataset about the Mastodon social network we previously released, we re-hashed the node/user identifier in the latter dataset, so that the same individual in both datasets corresponds to the same hash code.

A final remark concerns the diversity of the contents published in the Mastodon instances. Since each instance defines its policy and a code of conduct of its members, it is worth noting that pornography or contents for an adult audience may not be prohibited in some instances. This might hurt the feelings of some researchers who wish to avoid adult contents.

Legal considerations Even though the distributed nature of Mastodon allows each instance adopts a specific terms of use and service, many instance are used to adopt the standard terms of service and privacy policy provided by the Mastodon developers.[5]. In the terms of service and privacy

[3] https://scrapy.org/

[4] http://<instance>/api/v1/timelines/public?local=true

[5] https://mastodon.social/terms

policy the gathering and the usage of public available data is never explicitly mentioned, consequently our data collection seems to be complaint with the policy of the instance. Moreover if the server of an instance is in the EU or the EEA we also fulfill the requirements of the GDPR since we do not store and release personally identifiable information of the users. Finally, we have also respected the limitations imposed by the *robots.txt* files of the different instances.

Inappropriate contents: producers and properties of the text

In this section we provide some descriptive statistics about the dataset in order to offer an overview on how inappropriate contents are produced in the Mastodon network and to what extent the practice of labelling toots is spread among the Mastodon users. Moreover, we present a brief comparison between inappropriate and "appropriate" textual contents and a characterization of the field "spoiler-text".

Content production We start our analysis from the production of contents. In particular, in Table 1 we report the list of the instances that cover up to 70% of the total number of collected toots, ordered by number of toots. It is evident how the production of toots is unevenly distributed across the different instances, indeed a remarkable percentage of toots, around 39.5%, has been published in the instance "mastodon.social". This instance is one of the largest instance of Mastodon[6], it is not focused on any particular topic and most of its toots are in English.

Instance name	Number of toots	Percentage
mastodon.social	2,323,827	39.53%
switter.at	493,245	8.39%
mamot.fr	398,691	6.78%
mastodon.at	267,500	6,78%
niu.moe	212,093	3.60%
octodon.social	170,070	2.89%
social.tchncs.de	91,559	1.55%
mastodon.art	85,531	1.45%
todon.nl	73,580	1.25%
social.lescorpsdereve.space	73,289	1.24%

Table 1: Toots per instance. We report the number of toots published in each instance and its percentage with respect to the overall amount of posts.

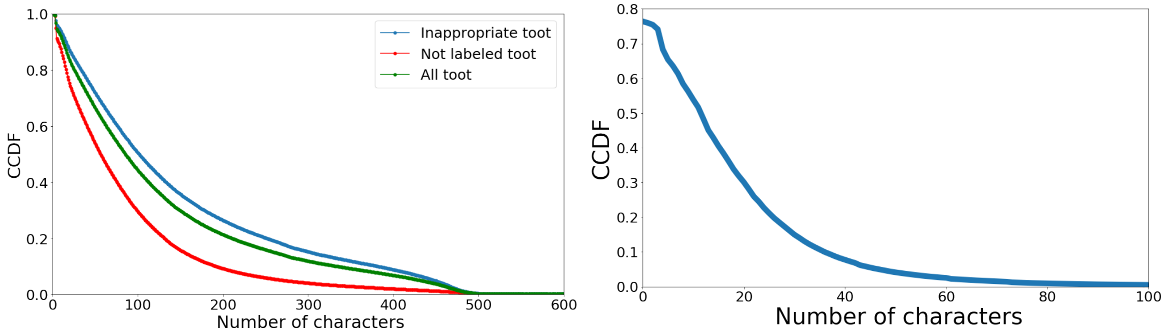
As for the production of inappropriate contents, we compute for each instance in Table 1 the fraction of toots labelled as inappropriate and the percentage of users who has produced at least one inappropriate content. Then, to assess whether or not super producers of bad contents exist, we quantify how inappropriate toots are distributed within the producers by using the Gini index. Gini index takes values between 0 – perfect equality – and 1 – complete inequality, all toots are produced by one super producer –, and provides a measure of the inequality of the productivity of instance users in terms of inappropriate toots. The above measurements have been reported in Table 2. From the results it emerges that most of the inappropriate toots are produced by a very small fraction of users, i.e. the highest percentage is 1.40%, got by the instance "switter.at". We also observe that the production of inappropriate contents is unevenly distributed across the bad content producers. In fact, for all the instances the Gini index is high, namely above 0.8, meaning that most of the sensitive toots belong to a very few users. This result confirms the presence of super producers of inappropriate contents. This specific type of user may correspond to *i)* people who maliciously publish inappropriate contents or *ii)* people aware of the content warning function and of the instance policy, who mark their contents as inappropriate when they think they may hurt the feelings of the instance members. This latter aspect will be the subject of a further investigation.

Instance name	% sensitive	% users	Gini
mastodon.social	5.42%	0.35%	0.825
switter.at	19%	1.40%	0.859
mamot.fr	0.6%	0.08%	0.858
mastodon.at	0.5%	0.03%	0.858
niu.moe	3%	0.22%	0.856
octodon.social	10.3%	0.36%	0.855
social.tchncs.de	10%	0.23%	0.857
mastodon.art	6%	0.86%	0.856
todon.nl	10.4%	0.25%	0.857
social.lescorpsde reve.space	0.002%	0.003%	0.857

Table 2: Production of inappropriate contents. Percentage of inappropriate toots and users who produce inappropriate toots, and the Gini index.

Since Mastodon allows users to enrich their toots with URL referring external pages and pictures, the identification of which content is inappropriate is not straightforward. To this aim, we limit our analysis to textual contents only. Specifically, we filter out all toots which contain URL and/or picture: in fact, if a toot contains a picture, the sensitive information is the picture itself, whereas if an URL is included in a toot, it means that the sensitive information could be in the referred content and not in the text of the toot. Moreover, contents referred by URL can be heterogeneous, e.g. web pages, videos, pictures, etc., and require ad-hoc solutions to be collected and labelled as inappropriate or not; a task which is out of the scope of this work. Finally, we discard all toots whose language is not English. By applying these filters we obtain 2,313,981 toots of which 206,486 (8.9%) are labelled as "sensitive". In Table 3 we report the list of the instances that cover up to 70% of the total number of the filtered toots, ordered by number of textual toots. We note that six instances, instead of ten, are enough to cover the 70% of the toots. In fact, more that 50% of filtered toots belong to the instance "mastodon.social". Finally, we surprisingly note that in the second instance, namely "switter.at" – an instance accepting pornography – only text posts are more frequent than expected as images or video are usually

[6]"mastodon.social" is the forth largest instance in Mastodon, the top-3 instances are Japanese instances.



(a)

(b)

Figure 4: Length of contents. In (a) the distribution of the number of characters per toot measured on all the toots (green), inappropriate toots (blue) and unlabelled toots (red). In (b) the distribution of the number of characters for the field “’spoiler-text’.

predominant in adult contents.

Instance name	Number of toots	Percentage
mastodon.social	1,162,746	50.24%
switter.at	141,867	6.13%
niu.moe	135,080	5.83%
octodon.social	88,481	3.82%
knzk.me	53,830	2,32%
mst3k.interlinked.me	47,869	2.06%

Table 3: Toots per instances after filtering. We report the number of toots published in each instance and its percentage with respect to the overall amount of posts.

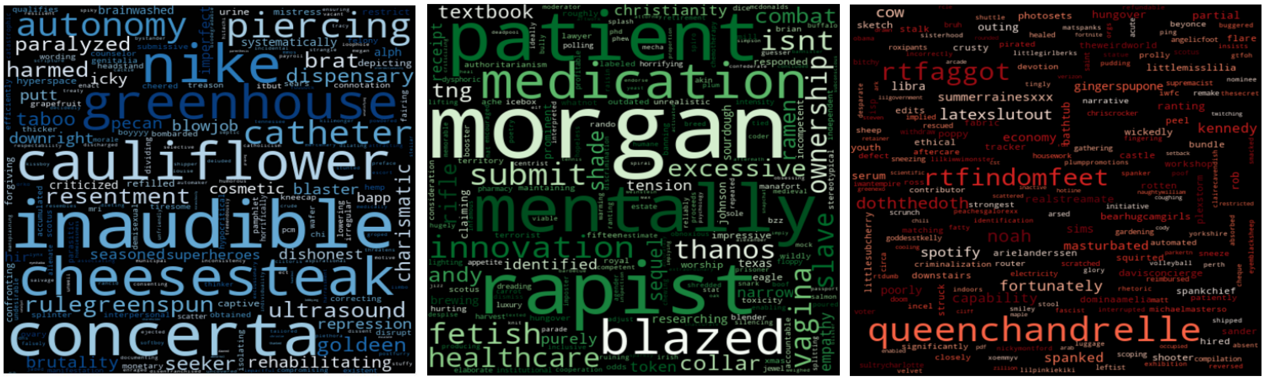
Text content properties We report some basic properties about the length of the text contents and brief overview of the words exclusively used in inappropriate contents. As for the first aspect, in Figure 4a we display the distribution of the number of characters per toot (green line). The toot length concentrates on values far from the maximum limit, indeed the mean and median values equal to 129 and 82, respectively. These statistics indicate that, generally speaking, users do not take full advantage of all the expressivity of the media, but use a much more concise style of communication, as also happens in other platforms such as Twitter (Gligoric, Anderson, and West 2018). But, it would be interesting to investigate whether it exists a dissimilarity in the communication style in terms of conciseness when user posts a inappropriate toot with respect to the publication of an “appropriate” content. To this purpose, we divide the two types of toots, then we compute the number of characters in the two populations and their respective distributions, which are shown in Figure 4a. The distribution of the number of characters for the inappropriate toot class (blue line) is quite dissimilar from the appropriate one (red line): the median value is equal to 148 characters per toot, doubling up the 73 characters per toot in the “appropriate” class. This result shows how the style of communication changes according to the content you want to publish, becoming much more direct and short in case you want to communicate “appropriate” contents. Note that when a user decides to censor her own toot, classifying it as inappropriate, she can add a brief description of the content in order to warn any audience sensitive to that kind of content. This brief description, namely the spoiler-text, does not have a maximum length, but deducts characters available for the body of the toot. Spoiler texts are usually very brief as they serve as a warning, while the content is embedded in the body of the toot available to willing readers, only. This trait is confirmed by the distribution of the spoiler-text length, shown in Figure 4b. Here, the probability of a spoiler-text having a length greater than 100 is negligible (more precisely $P(X > 100) = 0.0048$), with a median of 12.

Finally, we focus on the words which characterize the inappropriate text contents only. This analysis represents an initial step helping researchers in the development of more advanced techniques for the automatic identification of inappropriate contents in online social media. In fact, here we limit to a word-level analysis to build a Bag-of-Words representation. Specifically we adopt the following standard text-mining methodology for text preparation:

- HTML tag removal: we use Beautiful Soup library[7] to extract only the text part of the toot, removing any HTML tag;
- tokenization: we split the text into tokens based on white space and punctuation;
- normalization: we convert all the tokens to lowercase;
- stop-words removal: we use the *stopwords* module of NLTK[8] library to discard all tokens which represent stop-words in English language;
- lemmatization: we use the *WordNetLemmatizer* module of NLTK to bring back each word to its base or dictionary form;

[7] Beautiful Soup: https://www.crummy.com/software/BeautifulSoup/

[8] Natural Language Toolkit: https://www.nltk.org/



(a) mastodon.social (b) octodon.social (c) switter.at

Figure 5: Words cloud of inappropriate toots in three of the main Mastodon instances.

- non-English words removal: we use the *Detector* module of the *polyglot*[9] library to filter out all words which are different from English.

A set of words associated to each toot is the outcome of the above pre-processing. Finally, we compute the set of words which are exclusively present in the inappropriate toots. In Figure 5 we visualize this set of words for three of the major instances which produce inappropriate contents, i.e. “mastodon.social”, “octodon.social” and “switter.at”. The first instance is the main European general purpose instance, the second one is focused on technological and geek stuffs, whereas the third one is the place for sexual-related contents. As expected, the sets of words characterizing inappropriate contents are very different from one another. Specifically, in mastodon.social inappropriate contents span a variety of topics from sexual deviance to drugs, using slag words. In octodon.social, sensitive contents fall into two categories: (i) offensive and explicit sexual words, and (ii) spoilers of TV series or movies. As for switter.at instance, inappropriate posts cover the wide spectrum of adult contents. The above findings highlight that the contents and the topics people judge to be inappropriate depend on the community they belong to, meaning that the perception of what may hurt people’s feelings is influenced by the surrounding social context.

Conclusion

Recently, given the latest scandals regarding the violation of privacy and the widespread publication of inappropriate contents in online social media, research on decentralized social networks is experiencing a new and fast growing interest. In fact, decentralized social networks return to individual users or groups of users not only their personal data, but also the control over their contents. The underlying vision is to overcome the tradition paradigm that a central authority decides what to censor, and to give the responsibility of publication and censorship back to the users themselves. In this scenario, Mastodon represents the most interesting decentralized social network currently on the scene, as it is natively organized into communities that impose the rules of publication and censorship. Not only that: Mastodon allows the individual user to decide to publish an alert if she thinks that the toot she is publishing may be perceived as inappropriate by any reader. This feature makes the dataset that we are releasing on the Mastodon’s timelines a useful starting point to investigate what happens when we return responsibility for publishing content to users themselves. But above all, it allows us to investigate what people judge to be inappropriate from a non-authoritative viewpoint, taking into account the real perception of people’s sensibility. Moreover, the current dataset can be easily integrated with its complementary counterpart, i.e. the Mastodon social network, described in (Zignani, Gaito, and Rossi 2018) and released at https://dataverse.mpi-sws.org/dataset.xhtml?persistentId=doi:10.5072/FK2/AMYZGS, to investigate how phenomena as homophily or social pressure act on the perception of the appropriateness of the published contents.

The usage of this dataset empowers researchers to develop new applications as well as to evaluate different machine learning algorithms and methods on different tasks, e.g.:

- *inappropriate text classification*: the schema of the dataset, i.e. each inappropriate content is labeled, facilitates the evaluation of new supervised learning algorithms able to identify whether or not a content is inappropriate. Moreover, since we showed that the semantic of many words depends on the instances, we suppose that word embeddings learnt on different instances may improve the performance of the classifiers, being, at the same time, a method to compare the behaviors of the different communities in Mastodon.
- *automatic policy compliance*: the perception of how a content may hurt the people’s sensibility is person-dependent, even in the case of a common code of conduct. So, the development of a tool which automatically identifies whether or not a text content is compliant with the policy of the instance, and consequently perceived as inappropriate, might represent a useful extension to the

[9] Polyglot:https://polyglot.readthedocs.io/en/latest/

functions of Mastodon.

- *emoji and inappropriate contents*: "emoji" have been widely used to enhance the sentiment, emotion, and sarcasm expressed in social media messages, so much that they often play distinct social and communicative roles in text contents. In many cases, the text contents we released contain "emojis", but a few is known about their semantic and their usage when people express inappropriate contents.
- *users' reactions to inappropriate contents*: how do users react to inappropriate contents? Do they reply with further inappropriate contents or they are more tempered and try to re-establish the original spirit of the code of conduct of the community?

References

Davidson, T.; Warmsley, D.; Macy, M. W.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017.*, 512–515.

Dinakar, K.; Reichart, R.; and Lieberman, H. 2011. Modeling the detection of textual cyberbullying. *The Social Mobile Web* 11(02):11–17.

Founta, A.-M.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *ICWSM*.

Gayo-Avello, D. 2015. Social media, democracy, and democratization. *IEEE MultiMedia* 22(2):10–16.

Gligoric, K.; Anderson, A.; and West, R. 2018. How constraints affect content: The case of twitter's switch from 140 to 280 characters. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018.*, 596–599.

Jha, A., and Mamidi, R. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, 7–16.

O'Keeffe, G. S.; Clarke-Pearson, K.; et al. 2011. Clinical report-the impact of social media on children, adolescents, and families. *Pediatrics* peds–2011.

Ribeiro, M. H.; Calais, P. H.; Santos, Y. A.; Almeida, V. A. F.; and Meira, W. 2017. "like sheep among wolves": Characterizing hateful users on twitter. *CoRR* abs/1801.00317.

Salve, A. D.; Mori, P.; and Ricci, L. 2018. A survey on privacy in decentralized online social networks. *Computer Science Review* 27:154 – 176.

Vigna, F. D.; Cimino, A.; Dell'Orletta, F.; Petrocchi, M.; and Tesconi, M. 2017. Hate me, hate me not: Hate speech detection on facebook. In *ITASEC*.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data* 3.

Xiang, G.; Fan, B.; Wang, L.; Hong, J. I.; and Rosé, C. P. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *CIKM*.

Yenala, H.; Jhanwar, A.; Chinnakotla, M. K.; and Goyal, J. 2017. Deep learning for detecting inappropriate content in text. *International Journal of Data Science and Analytics* 6:273–286.

Zignani, M.; Gaito, S.; and Rossi, G. P. 2018. Follow the "mastodon": Structure and evolution of a decentralized online social network. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018.*, 541–551.