Clim/hate Change – Social Media archiving, AI tools & Hate Speech

Matteo Turrino for Semantic Digital Libraries with Giovanni Colavizza 2022-23

The problem: history and social media

This is an exploratory paper that takes into consideration a technologically-aided approach to the study of history. Digital platforms and social media in particular are, in the 21st century, producing content that is qualitatively and quantitatively challenging; massive amounts of short-form content are being published constantly on private platforms, with much of this content being personal and private in nature, yet accessible to the public. As a result, we are left questioning: how does content get archived? Who does the archiving, and how is it selected? The private nature of companies owning the content determines access in the first place; while companies like Twitter have collaborated with the U.S. Library of Congress, others such as Facebook seldom provide external researchers unrestricted access to their data[1]. There is also a problem of privacy and the right to be forgotten: is social media content ephemeral, or can it be archived as is? Which content should be archived, and which shouldn't? The European Court of Human Rights and UNESCO have promoted the creation of archives covering "digital only" material as an important source of education and research, as well as a safeguard of the right to information[2], yet such archives face various challenges amongst which there is the question of which data is archived. For instance, some social media archives focus on the profiles of important individuals and events, but they do not capture comments[3].

On top of the concerns about archiving content, there is the quantity. The amount being produced is enormous, which begs the question: how will historians be able to study such a wide corpus? While traditional ways of analysing texts will no means lose their importance, AI-tools will become increasingly useful in understanding social patterns in recent history. At the moment, however, NLP training for historians is not usually a priority.

As a case study for this paper, we decided to use the tools available today to see how difficult it is to retrieve social media data and use AI to add a layer of metadata, without having specialized training in AI. We decided to consider a specific social media phenomenon: hate speech directed towards activists. In particular, our **target is comments to Instagram posts from environmental movements**, where activists are seen performing high-visibility, disruptive acts with the end goal of requesting urgent government action with regards to the climate emergency.

For this purpose, we will be using Ultima Generazione's Instagram page. Specifically, we have selected posts where art works or public monuments have been affected (with washable paint

[1] Elisabeth Fondren and Meghan Menard McCune, 'Archiving and Preserving Social Media at the Library of Congress: Institutional and Cultural Challenges to Build a Twitter Archive', *Preservation, Digital Technology & Culture* 47, no. 2 (1 July 2018): 40, https://doi.org/10.1515/pdtc-2018-0011.

[2] Eveline Vlassenroot et al., 'Web-Archiving and Social Media: An Exploratory Analysis', *International Journal of Digital Humanities* 2, no. 1 (1 November 2021): 110, https://doi.org/10.1007/s42803-021-00036-1.

[3] Vlassenroot et al., 114.

or other substances). These posts in particular seem to have attracted a large amount of verbally abusive comments.

Our goal is to see what is possible with the current tools, to explore immediate possibilities and which aspects of this corpus could be opened up to further semantic expansion and exploration through AI tools.

Acquiring the corpus

Steps followed:

1. Data and metadata can be obtained from Instagram in multiple ways. I decided to use the python tool instaloader[4] for ease of use and documentation available.
2. We're only interested in the comments, and not the posts themselves, for simplicity. So we will be selecting posts of interest (see target above), get the post-id, and run the command to exclude videos and pictures:

   ```
   instaloader --comments --login=USERNAME --no-pictures --no-videos -- -"POST-ID"
   ```

   The posts have been selected manually; we could have downloaded all posts and their comments from the "Ultima Generazione" profile, but the Instagram API limits the number of calls made. For this exercise, we decided to limit the number to 9 manually picked posts which match the target.
3. We obtain the comments (data) and the metadata (when was it posted, by whom, in reference to which post..). For this repository, we are only interested in the comment itself, which is the **text** property of each comment. We will however retain the structured metadata as it could be useful for future usages.

The metadata

Instagram provides the following metadata for each comment:

```
 {
   "id": the id of the comment,
   "created_at": the date the comment was posted,
   "text": the content of the post,
   "owner": {
     "id": user id,
     "is_verified": whether the user is verified,
```

[4] 'Instaloader — Download Instagram Photos and Metadata', accessed 1 September 2023, https://instaloader.github.io/index.html.

```
      "profile_pic_url": url of the profile picture,

      "username": username

    },

    "likes_count": how many likes it has received,

    "answers": a list of answers the post has received

  }
```

The model

For the text analysis we looked for a Natural Language Processing model from huggingface.co (a community-built library of deep learning models). We considered using the Italian model (https://huggingface.co/Hate-speech-CNERG/dehatebert-mono-italian)[5] from Hate Alert, a research group at the Kharagpur IIT[6] that specializes in hate speech detection, even though for the purposes of the model, hate speech is a "direct and serious attack on any protected category of people based on their race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability or disease"; and does purposefully exclude toxic and abusive speech, because it is not considered discriminatory on any of those grounds.
So we also took into consideration Citizenlab's toxic comments (https://huggingface.co/citizenlab/distilbert-base-multilingual-cased-toxicity) multilingual model, which has a vaguer definition of what a "toxic" comment is.

Because the model is already pre-trained on a corpus of text, we do not need to provide any training data. While it would be ideal to do so, by using a specific corpus, we want to see if it is possible to use a pre-trained model to a reasonable accuracy.

As a test, we manually selected 30 comments that match our selection criteria. 10 are positive or neutral (cat 1), 10 are non-toxic negative (cat 2) and 10 are verbally abusive (cat 3).
The manual selection criteria, and what we are looking for in the model itself, is not the ability to tell negativity from positive comments. We want the model not to flag respectfully expressed comments (cat 1 and 2) and instead only flag as hate those that contain verbal abuse or threats of physical violence (cat 3) that are not conductive to a conversation.

We ran the comments in both models and calculated the accuracy for each. The Hate Alert model had some errors in categories 1 and 2 (incorrectly marking 20% of them as hate speech), but performed flawlessly on category 3.

The Citizenlab model on the other hand had 100% accuracy on categories 1 and 2, but performed very poorly on hate speech (recognizing only 20%).

As a result, we chose the Hate Alert model, even if it appears that its excellent performance is caused by the presence of profanity. Nonetheless, it is also able to correctly recognize veiled threats that do not contain profanity

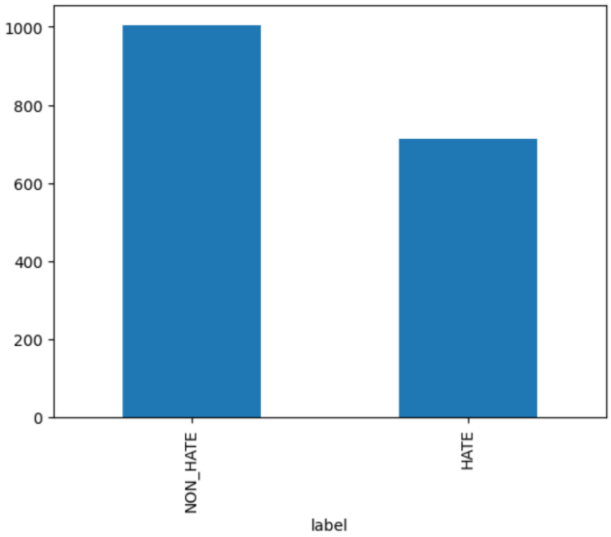
[5] Sai Saketh Aluru et al., 'Deep Learning Models for Multilingual Hate Speech Detection' (arXiv, 9 December 2020), https://doi.org/10.48550/arXiv.2004.06465.
[6] Hate Alert, 'Hate Alert', Hate Alert, accessed 31 August 2023, https://hate-alert.github.io/.

Cat 1 - Positive or neutral comments	HS	Toxicity
Raga grazie per quello che fate :)	NON_HATE	not_toxic
FANTASTICI GRAZIE	NON_HATE	not_toxic
Che gaso raga vvb grazie per quello che fate	NON_HATE	not_toxic
Per pulire la vernice servirà meno acqua di quella che serve per produrre una bistecca...	HATE	not_toxic
Cosa ne pensate del nuovo sistema di recupero plastiche nell oceano joint job francese ?	NON_HATE	not_toxic
io trovo invece che il colore doni bellezza alla statua	NON_HATE	not_toxic
Supporto da Parigi	NON_HATE	not_toxic
Ridicolo e allo stesso tempo agghiacciante arrabbiarsi per un muro sporco di vernice lavabile e far finta di niente con quelli che il pianeta e tutti i suoi ecosistemi stanno distruggendo.. e non con vernice lavabile.	HATE	not_toxic
Disobbedienza sempre! Bravi	NON_HATE	not_toxic
Grandi	NON_HATE	not_toxic
ACCURACY	**80%**	**100%**
Cat 2 - Non-toxic, negative comments		
Vien da piangere	NON_HATE	not_toxic
io non seguirei mai le idee di un gruppo che fa queste cose, vorrei una costruzione, una informazione e formazione. Bocciati	NON_HATE	not_toxic
Grazie internet per la sezione di comedia di oggi	NON_HATE	not_toxic
Sicuramente per risolvere dei problemi non se ne creano degli altri...	NON_HATE	not_toxic
IMBRATTATE CASA VOSTRA.	NON_HATE	not_toxic
Credibilità 0	NON_HATE	not_toxic
Altri modi no eh...??	NON_HATE	not_toxic
Ok che protestate, ma con la vernice stavate beccando i piccioni, non eravate ambientalisti?	HATE	not_toxic
Grandissimi eroi dei giorni nostri, per favore @ultima.generazione rispondi a questa mia semplice domanda, una volta che sporcate,dopo siete voi che pulite?	HATE	not_toxic
Chiedete il buonsenso che non avete. Vergognatevi.	NON_HATE	not_toxic
ACCURACY	**80%**	**100%**
Trigger warning: below are rude, insulting texts that contain profanity.		
Cat 3 - Verbally abusive comments		
Dovevate rimanere nelle palle dei vostri padri ...	HATE	not_toxic
Ve le farei pulire con la lingua sgasandovi con un naftone sotto al naso	HATE	not_toxic
venite qua a taranto a bloccare il traffico vedete che succede	HATE	not_toxic
Dovete morire merde	HATE	toxic
Spero vi brucino vivi	HATE	not_toxic
andate a lavorare INFAMI	HATE	not_toxic

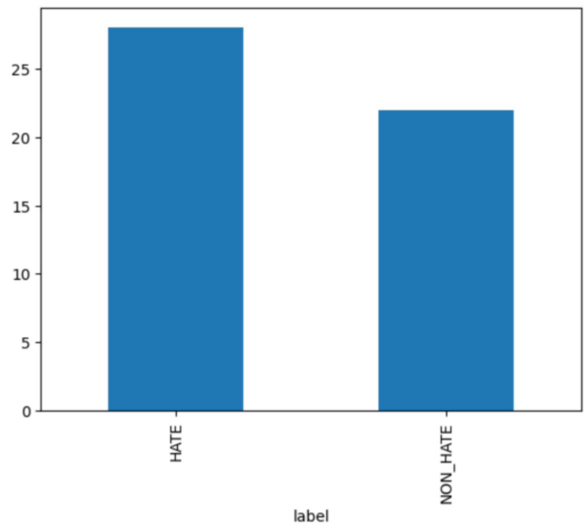
Mentecatti, esseri inutili	HATE	toxic
Ma quale narrazione costruite?? A spaccare le pietre a Vorkuta in Siberia con meno 50 gradi! Vi meritate questo e basta	HATE	not_toxic
Pezzi di merda vi sfido a fare le vostre puttanate a Napoli o Messina o Livorno poi vediamo	HATE	not_toxic
È tornata l'ultima generazione di coglioni	HATE	not_toxic
ACCURACY	**100%**	**20%**

Processing

We can now process all direct comments; Instagram allows for replies to top level comments, but because we are interested in reactions to the main post (not to its replies) and also for simplicity, we will stick to top level comments only. Of 1716 top level comments total, we end up with:



Which is definitely a very high percentage of verbal abuse. If we use the metadata property of “likes_count”, we can check if popular comments show a different ratio; However, in this case, the distribution swaps in favour of hateful comments.



Conclusions

While an oversimplification, this case study was meant to show that the creation of digital libraries with social media data, including comments and associated metadata, could be very useful for social and historical research. In the scenario depicted above, for instance, a longer time frame would allow to see whether hate speech has increased, or diminished over time. Further, with community-provided AI models and tools now being able to perform tasks that only a decade ago required a high degree of technical specialization, new layers of metadata can be created and used to assess public sentiment over specific topics at specific points in recent history. However, as specified in the introduction, this is also problematic from a privacy and archiving point of view: what should be archived? Are all social media conversations a worthwhile historical record, meant to be kept forever, or are they ephemera? Clearly, comments sections are rife for verbal violence. If users knew they will be forever on record, would they behave the same?

Instragram Posts used

- Monumento a Vittorio Emanuele, Milano https://www.instagram.com/p/CpkKotKNGA9/
- Fontana Piazza di Spagna, Roma https://www.instagram.com/p/CqflPtaN9-L/
- Palazzo Vecchio, Firenze https://www.instagram.com/p/Cp41JQdjwg2/
- L.O.V.E a Piazza Affari, Milano https://www.instagram.com/p/CncFoAZtEoB/
- Palazzo Madama, Roma https://www.instagram.com/p/Cm7BGfstK5o/
- BMW, Andy Warhol https://www.instagram.com/p/ClGd5w7DlBT/
- Il seminatore, Vincent Van Gogh https://www.instagram.com/p/CklQ_ddtdMk/
- Primavera, Botticelli https://www.instagram.com/p/CgUF88ojK2w/
- Gruppo del Laocoonte https://www.instagram.com/p/ChaeS1xN16D/

Alert, Hate. ‘Hate Alert’. Hate Alert. Accessed 31 August 2023. https://hate-alert.github.io/.

Aluru, Sai Saketh, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 'Deep Learning Models for Multilingual Hate Speech Detection'. arXiv, 9 December 2020. https://doi.org/10.48550/arXiv.2004.06465.

Fondren, Elisabeth, and Meghan Menard McCune. 'Archiving and Preserving Social Media at the Library of Congress: Institutional and Cultural Challenges to Build a Twitter Archive'. *Preservation, Digital Technology & Culture* 47, no. 2 (1 July 2018): 33–44. https://doi.org/10.1515/pdtc-2018-0011.

'Instaloader — Download Instagram Photos and Metadata'. Accessed 1 September 2023. https://instaloader.github.io/index.html.

Vlassenroot, Eveline, Sally Chambers, Sven Lieber, Alejandra Michel, Friedel Geeraert, Jessica Pranger, Julie Birkholz, and Peter Mechant. 'Web-Archiving and Social Media: An Exploratory Analysis'. *International Journal of Digital Humanities* 2, no. 1 (1 November 2021): 107–28. https://doi.org/10.1007/s42803-021-00036-1.