Hatemeter Project: Analysis of hate speech on twitter A join effort of computer science, humanities and social sciences

Mario Laurent[1]

[1] Université Capitole, Toulouse, France

Abstract. "Hatemeter" project benefits from Europe funding. Its aims to better understand, acknowledge and prevent online hate speech on social media. The project has led to the development of the Hatemeter Platform, which helps in monitoring and analyzing social media data. This paper describes the methodology we developed within the project and gives a specific focus on one of the three countries involved, France. We also present the various platform features and functionalities.

Keywords: Online hate Speech, Anti-Muslims hatred, Twitter monitoring, NLP applied on Social Sciences.

1 Hatemeter Project

1.1 Overall presentation of the project

The 'Hate speech tool for monitoring, analysing and tackling Anti-Muslim hatred online' (Hatemeter) project involves seven partners from three countries. The project received a European Commission grant within the framework of the Rights, Equality and Citizenship Programme (REC Action Grant, project reference: 764583). The project is two years long starting from March 2018. Its main objective is to create an ICT tool (the Hatemeter Platform) to automatically collect data from social media, produce graphs in order to visualize and analyze this data and help users write counter-narratives to the hate speech they encounter.

The team's efforts have focused on the specific case of anti-Muslim hatred and exchanges on the social media Twitter.

1.2 Steps of the project

The project consisted into three main phases:

- each team studied the specific context of its country: state of the art of scientific literature, identification of the laws related to the subject and qualitative interviews with on the ground actors, engaged in the fight against discrimination and hate speech against Muslim.

- the development phase of the Platform: consisted in testing sessions with the NGOs partners in the three countries involved in the project. The beta testers filled in reflexive evaluation questionnaires. Criminologists and sociologists also provided feedback to the development team as new functionalities were added.
- each partner was able to use the platform to carry out its own analyses, with the main objective of producing counter-narratives, for NGOs, and understanding the construction and dynamics of hate speech for researchers [1], [2].

In the rest of the paper, we will focus on the context and results specific to France, but it should be point out that each of the three countries involved faces a different and unique situation.

2 Context of study – France case

2.1 French law context and history

France has a colonial history which has led to greater ethnic and religious mixing, for decades, but which has sometimes given rise to contentious opposing these different populations. France has also suffered directly and repeatedly from terrorist attacks perpetrated by extremists in recent years. This can also fuel discourse of mistrust towards Muslim citizens in general.

In France, several laws play a role in how we can deal with online hate speech.

First, hate speech directed against any person based on their origin or religion is strictly prohibited, both in the case of direct physical or verbal attack and in the case of incitement to hatred, including using electronic devices. This is stated by the Law of 1881 on media freedom [3] and reinforce by a decree of the Penal Code from 2005 [4] which specify that these discourse are also forbidden in a private space.

Then the Law on the Separation of Church and State is often used as a justification for excluding or demanding the exclusion from the public space of any person wearing visible religious symbols. In the case of French Muslim citizens, it is veiled women who are regularly targeted, although this law does indeed concern institutions and not individual citizens, whose freedom of worship is affirmed by the same law [5]. The debates around the interpretation of this law create perpetual controversy between different political groups and sometimes feeds the hate speech on social media.

Finally, the Law on ethnic statistics can be both a protection for citizen’s rights and an additional difficulty to study and understand hate speech. As it prevents to gather any statistics on religion or ethnicity, it ensures a blind and ethical treatment between citizens when processing computerized data on a large scale [6]. However, the lack of official government statistics is often used by groups stigmatizing Muslims to exaggerate their numbers in France and give credence to the thesis of a massive cultural replacement that would threaten traditional French values. Despite this law, UK polling institutes can still carry out surveys on the Muslim population in France and on the estimation of this population by French. According to one study, the estimated rate of

Muslims in France in 2016 was 31% while the actual rate was only 7.5% [7]. In addition, while the Muslim population in France at the end of 2017 was estimated at 5.7 million individuals, some far right groups predicted a number of 20 million [8]. This illustrates the use of the lack of data to play on the fear of a clash of civilizations, theorized by some far right-wing activists.

In July 2019, a new law was debated by the National Assembly, which should make hosts responsible for the hate content published on their platform, including social networks and media (Twitter, Youtube, Facebook) [9]. The legislative field is therefore currently undergoing a transformation around these issues of hate speech, particularly online. It should be noted that the term 'Islamophobia' originally used in the draft law has been replaced by 'anti-Muslim hatred', for fear that there might be confusion between a ban on incitement to hatred against individuals and a ban on criticism of religion itself, which would not be compatible with freedom of opinion [10]. This illustrates the terminological disagreement surrounding incitement to hatred, which is difficult to resolve: many discrimination are indeed on a blurred line between hatred of a person because of their religion, racism, and hatred of religion that extends to the people practicing it. On this point, recent research analyzing the evolution of political discourse in the assembly [11], as well as the discourse of extremist groups online [12], shows a tendency in hate speech directed against "the other", designated as a common enemy of the nation, to abandon biological racial criteria and rely increasingly on cultural and religious criteria.

2.2 Twitter analysis

As any other communication channel, Twitter is a tool that has its limits and rules. A message conveyed on Twitter is therefore, at least, affected by the media and in some cases quite far from the author's thinking, in order to serve their objectives. If it is presented as a tool for sharing information and opinions, Twitter is above all based on a system where access to information depends on the profiles followed by its users. This results in a race for audience and provokes the implementation of communication strategies.

One important characteristic is the maximum size of a Tweet, limited to 280 characters. If many users use the "thread" technique (responding to themselves to increase the possible length of their speech), the mechanics of the media still encourages to write in a concise and punchy way in order to be relayed, and therefore read. Longhi has studied the relationship between the form and content of a message and its effectiveness (number of responses and retweets) [13]. He concluded that this effectiveness seems to be directly related to the controversial nature of its content. Other researchers have studied the propagation in social media of verified news and fake news and have shown that fake news will both reach more people and spread faster [14].

According to our observations, other strategies, such as systematically responding to much more followed profiles by expressing an opposite opinion or exaggerating one's opinion, are widely used to increase one's visibility/number of followers. Finally, the use of subtle and understated discourse, or irony, knowing that the audience will understand the true intentions of the message, is used recurrently in the networks we observed and is a challenge both for automatic analysis and for human analysis, if they are unfamiliar with the community and the subject in question.

2.3 French data sets

Each research team conducted in-depth interviews that allowed us to identify groups that produce anti-Muslim hate speech as well as the hashtags and keywords they use.

In France, interviews and observations allowed us to identify two main forms of anti-Muslim hatred: first, impersonal discourse based on negative stereotypes associated mainly with North Africans descent and culture; secondly, cyber-harassment campaigns that target specific individuals, having appeared in the media as Muslim personalities. The main objective of such campaigns is to silence Muslim citizen and force them out of the public/media sphere. Generally, harassment follows a recurring pattern: a visible religious sign or reference appeared in a media. This sign is associated with a willingness of its bearer to proselytize and immediately seen as an indication of belonging to extremist movements. Finally, the person is quickly seen as an accomplice or sympathizer of terrorist groups. After being identified as a threat by a group of harassers, the victims can be pursued relentlessly, receiving private messages or numerous negatives comments under each of their Tweets. The private messages can go from insults to death threat. Some of the interviewees were exposed to such threat or/and have had their addresses or private photos revealed on harassers group pages.

3 Hatemeter Platform

The Hatemeter Platform [2] serves three main purposes:

- collect data on Twitter,
- store data in a database,
- provide tools to visualize this data, and provide help to write counter narratives to fight against online hate speech.

The Platform collects data on Twitter and Youtube. For Twitter a list of hashtags/ keywords often used by hate groups is used, but also some more "neutral" hashtags that simply talk about anti-muslim hatred (such as "Islamophobia"). In each language, all Tweets displaying the targeted hashtags and keywords are collected daily, along with metadata such as user, date, number of retweetings, number of responses, etc. All this information is then reused by the platform to create graphs allowing the user to analyze the discourse and the networks behind this discourse. These graphs also allow diachronic or synchronic comparisons. All features have been programmed to be as much as possible language-independent, available in Italian, French and English, they could easily be reused for additional languages.

The Platform include five features applied on Twitter.

- 'Recent trends' gives access to a list of all the tweets of the last 24 hours for a selected hashtag or keyword (directly from Twitter).

- 'Hashtags trends' gives access to stored data. It is completed by day by day statistic and display of the most retweeted messages displaying this same hashtag. The co-occurrence graph can be moved or zoomed and the nodes display the names of the hashtags they represent.
- 'Hate speakers' feature focuses on a chosen 10-day period as for 'Hashtags trends' but is applied on users interactions rather than hashtags/keywords.
- 'counter narratives' function allows the user to enter a hate message, either by typing it on the keyboard or by selecting it directly from one of the other features of the platform. The user is then presented with a list of four possible responses, from which they can choose, or decide to edit one. The platform associates each new hate message with one of its initial stereotypical messages and can thus suggest the most relevant answers possible among those available. When the user is satisfied with their counter-narrative, they can copy/paste it to post it with their Twitter account.
- 'alerts' feature allows the user to view the number of Tweets with a selected hashtag/keyword, day by day, over the entire period of time it was present in the list monitored by the platform.

4 Conclusion

Although the project does not introduce major technological or theoretical innovations, its originality and interest lies in the multidisciplinary nature of the research teams, the cooperation with on the field NGOs to develop credible and applicable tools, and the international collaboration that makes it possible to highlight common needs as well as differences between European countries. The project also makes it possible to uncover many questions inviting further research.

Acknowledgement
The consortium is composed of: the University of Trento, Italy, coordinating the project, the Capitole University of Toulouse, France, and the Teeside University, UK, brought together researchers in criminology, social sciences, law and linguistics. The Bruno Kessler Foundation (FBK), a research institute in Trento, was responsible for developing the Platform thanks to a team of NLP (Natural Language Processing) researchers.

References

1. Hatemeter Project Homepage, http://hatemeter.eu/, last accessed 2020/25/02.
2. Di Nicola A, Andreatta D, Martini E, Antonopoulos GA, Baratto G, Bonino S et al.: HATEMETER: Hate speech tool for monitoring, analysing and tackling Anti-Muslim hatred online. eCrime, 71 p. Trento: eCrime (2020).
3. Law on 'liberté de la presse', Article 24 (J.O, 29 July 1881). Available online at: https://www.legifrance.gouv.fr

4. Decret n°2005-284 (J.O, 30 March 2005), Article 1. Available online at : https://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006070719&idArticle=LEGIARTI000006419500
5. Law on ' la séparation des Eglises et de l'Etat', Article 1 (9 december 1905, reinforce on 25 february 2020). Available online at: https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000508749
6. Ordonnance on 'statistiques éthniques', Article 226-19 (12 december 2018). Avalaible online at: https://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000026268247&cidTexte=LEGITEXT000006070719
7. Duffy, B. : Perceptions are not reality: what the world gets wrong. Ipsos Mori (2016). https://www.ipsos.com/ipsos-mori/en-uk/perceptions-are-not-reality-what-world-gets-wrong
8. Banet, R. and Fauchet, B. (2018). 20 millions de musulmans en France ? Ils sont environ 4 fois moins, selon les estimations les plus sérieuses, AFP. Available online at: https://factuel.afp.com/20-millions-de-musulmans-en-france-ils-sont-environ-4-fois-moins-selon-les-estimations-les-plus
9. Law on 'Lutte contre la haine sur internet', still in discussion, see on the official French assembly website: http://www.assemblee-nationale.fr/dyn/15/dossiers/lutte_contre_haine_internet
10. See a news article on that event: https://www.lci.fr/politique/cyber-haine-pourquoi-le-mot-islamophobe-a-t-il-ete-enleve-de-la-proposition-de-loi-2124845.html
11. Tahata, Y.: Définir « les Français », une question de race ? Analyse des formes de racisation dans les débats parlementaires sur la nationalité et l'immigration (1981-2012). Mots, Les langages du politique, 116 (2018). DOI: 10.4000/mots.23050.
12. Froio, C.: Race, Religion, or Culture? Framing Islam between Racism and Neo-Racism in the Online Network of the French Far Right. Perspectives on Politics, 16(3), 696–709(2018). DOI:10.1017/S1537592718001573.
13. Longhi, J.: Tweets politiques: corrélation entre forme linguistique et information véhiculée. #Info. Partager et commenter l'Info sur Twitter et Facebook (2017). https://halshs.archives-ouvertes.fr/halshs-01841132/document 2017
14. Vosoughi, S., Roy, D., Aral,S.: The spread of true and false news online. Science, 359(6380), 1146-1151 (2018). DOI: 10.1126/science.aap9559