European Crime Prevention Award (ECPA)

Annex I

Approved by the EUCPN Management Board in 2018

Please complete the template in English in compliance with the ECPA criteria contained in the Rules and procedures for awarding and presenting the European Crime Prevention Award (Par.2 §3).

General information

1. Please specify your country.

Spain

2. Is this your country's ECPA entry or an additional project?

Yes, it is ECPA's entry

3. What is the title of the project?

SocialHaterBERT: automatic detection and monitoring of hate speech on Twitter through a dichotomous approach based on textual analysis and user profiles

4. Who is responsible for the project? Contact details.

Lara Quijano Sánchez
Position: Assistant Lecturer/ Researcher
Institution: Escuela Politécnica Superior,Universidad Autonóma de Madrid, Madrid, Spain
Email: lara.quijano@uam.es
Address: C. Francisco Tomás y Valiente, 11, 28049 Campus de Cantoblanco, Madrid

5. Start date of the project (dd/mm/yyyy)? Is the project still running (Yes/No)? If not, please provide the end date of the project.

Start date: 01/01/2018. Still Running: Yes

6. Where can we find more information about the project? Please provide links to the project's website or online reports or publications (preferably in English).

Project's website: https://giogia.notion.site/giogia/SocialHaterBERT-Project-9d93e0e5ed4a468fb09b42a50d9a3dd9

Bullying Case study video (a three week evolution of Twitter messages in a high school class): https://www.notion.so/giogia/SocialHaterBERT-Project-

9d93e0e5ed4a468fb09b42a50d9a3dd9#2e40334deb174fd89b2c193a5e9457c3

Papers (in English): [*Pereira et al. 2019*] **Detecting and Monitoring Hate Speech in Twitter**. Sensors, Special Issue on Sensors for Affective Computing and Sentiment Analysis, volume 19, 2019. https://doi.org/10.3390/s19214654. **JCR: 3.275** (Instruments and Instrumentation, 15 de 64, **Q1**). **Referenced 43 times**. https://www.mdpi.com/1424-8220/19/21/4654

Prizes (Spanish):

http://www.fundacionpoliciaespañola.es/index.php/actividades/premios/premio-de-investigacion-de-la-f-p-e/127-resolucion-de-los-premios-de-investigacion-de-la-fundacion-policia-espanola-2018; http://www.interior.gob.es/prensa/noticias/-/asset_publisher/GHU8Ap6ztgsg/content/id/9902695

News (Spanish):
https://elpais.com/tecnologia/2018/11/01/actualidad/1541030256_106965.html

https://www.elespanol.com/omicrono/20181101/mide-odio-redes-sociales/349966008_0.html

https://elpais.com/tecnologia/2019/05/28/actualidad/1559039892_332196.html

Dataset (English): https://zenodo.org/record/2592149#.YWanfBztaUl . **4,279 downloads**

7. Please give a **one page** description of the project (**Max. 600 words)**

Social media platforms have evolved into digital representations of our social interactions. We can use the resources they provide to investigate phenomena that occur within them, such as the creation and spread of offensive and hostile content. The escalating nature of this behaviour in today's polarized world is cause for concern in modern society. This research examines previous efforts and strategies for detecting and preventing hateful content on the social network Twitter, as well as a novel classification approach based on user profiles, related social environments, and generated tweets.

SocialHaterBERT is the result of a multi-stakeholder collaboration between university (the Computer Science Department of the Higher Polytechnic School of the Autónoma University of Madrid) and a public institution (the Spanish National Office Against Hate Crimes of the Spanish State Secretariat for Security of the Ministry of Interior, *ONDOD*). However, this project's outcome is not country or language specific, as it could be easily localized by retraining the designed algorithms with different datasets due to the open science methodology followed, where all source files are freely posted in the project's website.

SocialHaterBERT is comprised of **two modules**:

1. **Bullying Speech Detection**: tweet collection and classification. This results in a revolutionary classification method that outperforms earlier state-of-the-art models.
2. **Social Network Analyzer**: monitoring and visualizing the detected hate content. The literature on hate speech has so far focused solely on providing algorithms for classifying texts as containers of hate speech or not. We take it a step further: the classification offered by SocialHaterBERT can be used to construct networks of concepts and actors based on their link to hate messages, which can then be visually depicted. To the best of the authors' knowledge, this is the first system with these features to be presented in the literature.

The main application of the system is to enhance the pertinent authorities' work on early detecting possible hate content (such as bullying) that may lead to severe consequences (i.e. hate crimes, suicides, etc). Observers can, for example, use SocialHaterBERT to evaluate the evolution of various hate trends, key terms, or actors; research their triggers; and design methods to minimize concrete peaks of hatred (i.e, bullying and violence among minors online). It also gives an interpretable graphical representation of the classification algorithm. Unlike other systems, which are typically black boxes, SocialHaterBERT displays relevant terms, receivers, and emitters discovered inside hate speech messages. This function can be used to investigate more relationships or other components inside hate-related material and gives transparency and explainability to the system which are now hot problems in the field of artificial intelligence technologies.

Scientifically, four contributions are made:

i) A new public dataset that can be used to test, train, and benchmark new developed methods.

ii) An improvement in the performance of the state of the art text based algorithms capable of identifying hate speech on Twitter. To that end, a model based on BERT has been created and tested using our public dataset, providing results that show a significant improvement.

iii) A methodology to create a user database in the form of a relational network to infer textual and centrality features. This contribution has been independently tested with various traditional Machine Learning and Deep Learning, demonstrating the potential usefulness of these features in spotting haters.

iv) A final model, SocialHaterBert, that integrates the previous two approaches by analyzing features other than those inherent in the text. Experiment results reveal that this last contribution greatly improves outcomes, establishing a new field of study that transcends textual boundaries, paving the way for future research in coupled models from a diachronic and dynamic perspective.

I. The project shall focus on prevention and/or reduction of everyday crime and fear of crime within the theme.

8. Which **crime prevention/ reduction mechanisms** were used in this project to contribute to crime prevention and/or the reduction of crime or the fear of crime? Multiple answers are possible.

☒ **Establishing and maintaining normative barriers to committing criminal acts**
e.g. 'Offenders, we are watching you' campaigns

☐**Reducing recruitment** to criminal social environments and activities by eliminating or reducing the social and individual causes and processes that lead to criminality
e.g. social and financial support for disadvantaged families

☐**Deterring** potential perpetrators from committing crimes through the threat of punishment
e.g. decreasing the time between arrest and punishment

☒ **Disrupting** criminal acts by stopping them before they are carried out
e.g. increasing police patrols in vulnerable areas

☒ **Protecting vulnerable targets** by reducing opportunities and make it more demanding to carry out criminal acts
e.g. placing locks and cameras

☐**Reducing the harmful consequences** of criminal acts
e.g. initiatives to recover stolen goods

☐**Reducing the rewards** from criminal acts
e.g. restorative justice programmes

☐**Incapacitating** (or neutralising) perpetrators by denying them the ability (capacity) to carry out new criminal acts
e.g. imprisonment of key gang members

☐**Encouraging** desistance from crime and rehabilitating former offenders so they are able to settle back into a normal life
e.g. prison rehabilitation programs

Explain how this/these crime prevention mechanisms were used ((**Max. 300 words**)

SocialHaterBERT has been used as an observatory of phenomena for significant events and days, such as the International Women's Day, the gay pride parade, or soccer matches, where it has been particularly useful (https://elpais.com/tecnologia/2019/05/28/actualidad/1559039892_332196.html) and to study its impact on LGTBI Communities (http://www.mineco.gob.es/portal/site/mineco/menuitem.ac30f9268750bd56a0b0240e026041a0/?vgnextoid=b1ad544a1929b610VgnVCM1000001d04140aRCRD&vgnextchannel=864e154527515310VgnVCM1000001d04140aRCRD)
The ONDOD has evaluated the results of both modules: (i) the lists of tweets identified as hate containers of the Bullying Speech Detection module, and (ii) the

evolution of the hate cloud and the different tabs of the Social Network Analyzer module. Some of the activities that the use of this system has led to:

• Analysis of tweets tagged by SocialHaterBERT as hate speech containers including their symbology (e.g., emojis).

• Analysis and classification of "tweeter" communities that share messages with toxic content.

• Statistics on relevant events, words and terms, used as a support tool for the police units with Twitter "Trusted Flagger" licenses, for the elimination of hate content.

• **A practical case study where SocialHaterBERT has artificially monitored 3 week Twitter messages of 18 17-year-old students belonging to the same highschool class**. Results:

- Show different communities, synergies and social groups.
- Highlight a peak of hate messages being conducted towards 3 class members.
- Identify emitters, observant and receivers of the bullying episode.
- Highlight the topics, trends and main words being used within the classroom.

The presence and public awareness of this tool inside institutions (such as schools) can serve as a dissuasive motivation as well as a useful instrument in: i) preventing toxic waves of bullying or violent messages, early cutting the problem; ii) understanding dynamics and potential influencers (hate promoters) who should be closely monitored; iii) protecting victims that are being wronged; iv) identifying group synergies, trending topics, and triggers in order to develop mitigation strategies.

II. The project shall have been evaluated and have achieved most or all of its objectives. For more information on evaluation, click here

9. What were the reasons for setting up the project? Was this context analysed before the project was initiated and in what way (How, and by whom? Which data were used?)? In what way did this analysis inform the set-up of the project? (**Max. 150 words**)

The main reason for setting up the project was to contribute to the development of the capacities of State authorities to help in **the identification, monitoring and analysis of online hate speech (particularly on Twitter) to design coordinated actions against cyberbullying**, a big problem that generally affect people's dignity.

We had the previous analysis of the ONDOD, Ministry of the Interior, which oversees recording and monitoring all incidents related to hate crimes and establishing and maintaining contacts with the third sector. This work has even been acknowledged by the EU's Fundamental Rights Agency itself.

With all this previous information, plus the results of the biannual victimization survey developed by the ONDOD, we were perfectly aware of the needs and requirements of the appropriate authorities.

10. What were the objective(s) of the project? Please, if applicable, distinguish between main and secondary objectives. (**Max. 150 words**)

The main goal of this project has been the design of a system, SocialHaterBERT, capable of detecting and classifying bullying and

violence among minors on Twitter, as well as monitoring, analyzing and reporting in a graphical manner hate trends, hateful messages and their emitters and receivers. **This system allows to detect hate waves and triggers, particularly against** minorities or vulnerable individuals like **teenagers who post unfiltered content online**. This tool provides valuable information to security agencies, police departments, and competent authorities (such as school administrators), allowing them to respond more quickly. Thus, the system helps to enhance investigations by taking earlier measures and thus prevent possible extreme negative outcomes (hate crimes, suicides, etc), and protect more effectively vulnerable individuals or groups.

11. Has there been a process evaluation?[1] Who conducted the evaluation (internally or externally?) and what were the main results? Which indicators were used to measure the process? Did you make changes accordingly? (**max. 300 words)**

The project has been subject of several evaluations that have led to pertinent changes both internal: by the ONDOD and law enforcement agents who, after using the tool, have given their feedback for its continuous improvement; and external, where the scientific quality of the contribution has been evaluated as follows:

- Since the beginning of the project (2018) it has been evaluated on numerous occasions:
  - June 2018: By the panel of judges of the Big Data master of the Carlos III University of Madrid where it was presented as a Master Thesis with title "Analysis of Hate Crimes on Twitter" and obtained a grade of 10 with honors.
  - July 2018: Release of the dataset. **Impact**: **4,279 downloads**
  - December 2018: By the panel of judges of the Research Award of the Spanish Police Foundation 2017-2018, Project: "Automatic detection and analysis of hate speech in the social network Twitter
  - September 2019: By a double blind review process carried out by the journal Sensors (JCR indexed Q1), where the work "Detecting and Monitoring Hate Speech in Twitter" was published. This review process involved the changes required by the reviewers. **Quality and impact**: Journal Sensors, **JCR (2019): 3.275**, Instruments and Instrumentation **Q1**). **43 References.**
  - June 2021: By the panel of judges of the degree in computer science of the Autónoma University of Madrid where it was presented as a Bachelor's Degree Final Project with title "Detection of hate messages on Twitter: a study based on profiles within the social network" and obtained a grade of 9.5.
  - July 2021: By a process of blind reviewers carried out by the Proceedings of International Congress "Hate and Discrimination in convulsed times", where the work "Detection of Hate Messages on Twitter: A Bert-Based Model for Classifying Hate Speech in

[1] **Process evaluation:** Also called *implementation evaluation*, or *monitoring*, this process documents **how the activities were implemented** in order to determine any deviations from the original planning. It facilitates finding explanations for when the results of the intervention are not as expected.

Spanish" was published.

12. Has there been an outcome[2] or impact[3] evaluation? Who conducted the evaluation (internally or externally?), which data and evaluation method were used and what were the main results? Which indicators were used to measure the impact? (**Max. 300 words**)

The use of **SocialHaterBERT has enhanced ONDOD monitoring activities**, allowing for a more accurate forecasting of hate crimes. By allocating resources and efforts at specific times and locations, it has aided in the prevention of hate speech's derived effects. The Spanish Action Plan to Combat Hate Crimes calls for an evaluation of the ONDOD performance every 6 months. These evaluations are carried out by members of the State Law Enforcement Agencies, the Ministry of Interior, and other institutions. So far, all evaluations regarding SocialHaterBERT have been very positive. ONDOD will deploy this tool within the two national law enforcement agencies in the near future, publicising its existence and its **public free code nature allowing its reproducibility in other institutions (i.e high schools or other educational institutions).** As a result, the pertinent authorities' response could be even quicker, reducing the likelihood of hate waves or at the very least mitigating their effects.

As for the evaluation method, data and main results of the project's scientific contribution, in our paper *Pereira et al. 2019* we demonstrated that the designed algorithm improved the techniques designed so far. As a result, and following a process of continuously updating our algorithm with emerging techniques, since 2019 we have worked on improving this version and the state of the art proposals that followed it. All of these proposals revolve around the use of the BERT algorithm. The following Table presents the experimental results (soon to be published in a JCR indexed journal) obtained comparing the use of BERT with the improvement proposed in our project. This projects' proposed strategy improves models based solely on BERT by more than 4%, representing a more than 18% improvement over the results obtained in *Pereira et al. 2019*:

Model	**Accuracy**	**F1**	**AUC**	**Precisión**	**Recall**
HaterBERT	0.8343	0.7645	0.7354	0.8506	0.7354
SocialHaterBERT	**0.8472**	**0.8023**	**0.8923**	0.7031	0.7826

III. The project shall, as far as possible, be innovative, involving new methods or new approaches.

[2] **Outcome evaluation:** Measures the **direct effect** (i.e., extent of the changes) **of the intervention on the target group, population, or geographic area**. The information produced by the outcome evaluation determines at what level the **objectives were achieved**.

[3] **Impact evaluation:** Measures **long-term effects** of the intervention on the target group, as well as **indirect effects** on the broader community. The information produced by the impact evaluation determines at what level the **ultimate goals** of the intervention were achieved.

13. How is the project innovative in its methods and/or approaches? (**Max. 150 words**)

Regarding **Bullying Speech Detection:** Following an extensive review of the state of the art, we detected that the best performing methods use a BERT-based classifiers, most of them designed for English texts, three for Spanish, and none combining profiles and relationships between users. This work's scientific potential is to build a multimodal model that extracts certain features from user profiles on Twitter, with the goal of modelling traits alongside the content of the tweet itself and merging them in a novel architecture that combines BERT with Multimodal Transformers, allowing cutting-edge deep learning techniques to reach their full potential. The algorithm presented in this work that combines these two types of attributes outperforms the best base algorithm that only uses textual information by 18%.

As for the **Social Network Analyzer:** this is the first work to visualize, monitor and summarize relevant terms, receivers, and emitters discovered inside hate speech messages.

IV. The project shall be based on cooperation between partners, where possible.

14. Which partners or stakeholders were involved in the project and what was their involvement? (**Max. 200 words**)

The following two partners have been involved in the design and completion of the project:

- Design and implementation: Escuela Politécnica Superior, Autonóma de Madrid University, Madrid, Spain
- Design and dataset generation: Spanish National Office Against Hate Crimes of the Spanish State Secretariat for Security (ONDOD).

Besides, the following stakeholders have been involved as consultants:
Apart from the main Spanish Security Forces, members of the third sector have been carefully considered for the design of the project and will be part in the use and development for improving continuous and collective learning: Movement Against Intolerance (McI); Federation of Gays, Lesbians, Bisexuals and Transsexuals (FGLBT); Fundación Secretariado Gitano, etc.

Also, through the ONDOD there has been coordination with Institutions such as:

- Ayuda a Niños y Adolescentes en Riesgo (ANAR, Help for Children and Adolescents at Risk)
- General Directorate for Equal Treatment and Racial and Ethnic Diversity
- Council on Equal Treatment and Non discrimination
- the Spanish Observatory on Racism and Xenophobia (OBERAXE)

V. The project shall be capable of replication in other Member States.

15. How and by whom is the project funded? (**Max. 150 words**)

No funding

16. What were the costs of the project in terms of finances, material and human resources? (**Max. 150 words**)

0 cost

17. Has a cost-benefit analysis[4] been carried out? If so, describe the analysis, including how and by whom it was carried out and list the main findings of the analysis. (**Max. 150 words**)

The cost has been zero euros. In return, this tool has replaced a full-time person monitoring the network. We can calculate an economic saving of about 50,000 euros, plus the benefit of prevention.

18. Are there adjustments to be made to the project to ensure a successful replication in another Member State?

The Project Code is available on the official website (https://giogia.notion.site/giogia/SocialHaterBERT-Project-9d93e0e5ed4a468fb09b42a50d9a3dd9), where updates are periodically uploaded. Work is being done to give a set of instructions for its reuse and customization along the open computer code.

Likewise, the generated dataset is available for download at https://zenodo.org/record/2592149#.YTnCEd_taUk along with instructions for its use.

The classifier may already be retrained for different languages or domains (for example, texts with examples of cyberbullying among teens), so it can be reused in other member states with just a new dataset that captures similar features.

A use case of how to reuse the tool for a school with the goal of detecting cyberbullying and monitoring students' social network activities is also available on the projects' official website.

19. How is the project relevant for other Member States? Please explain the European dimension of your project.

The ONDOD works closely with European and International Institutions such as OSCE and the FRA and participates in numerous European working groups where best practices are exchanged and shared strategies for monitoring hate speech take place. The aim is transferring knowledge from the SocialHertBERT project to the rest of European colleagues.

Please provide a short general description of the project (abstract for inclusion in the conference booklet – **max. 150 words**).

[4] **Cost-benefit analysis**: A type of economic evaluation that compares the direct and indirect cost of the resources employed in the intervention, with the equivalent economic value of the benefits.

Social media are real-world sensors that can be used to gauge a society's pulse. The vast unfiltered flood of messages is an alarming phenomenon in society, especially if when containing hate speech. In this project, we present SocialHaterBERT, an intelligent system currently being used by the Spanish National Office Against Hate Crimes that identifies and monitors the evolution of hate speech in Twitter. The contributions of this research are many-fold: (1) It introduces the first intelligent system that monitors and visualizes hate speech in social media using social network analysis techniques. (2) It introduces an algorithm that examines features other than those found in the text in an innovative way. Experiments on a case study demonstrate its utility in identifying senders and receivers of hate messages (i.e. bullying) within a high-school class, as well as monitoring group dynamics and toxic peaks in a visual and intuitive manner for authorities.