Natural Language Processing (2026), **32**, pp. 470–517
doi:10.1017/nlp.2026.10024

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Hate speech on Twitter: Profiling users and interaction analysis in the Spanish language

Irene Ramiro-López[1], Lara Quijano-Sánchez[1,2] and Federico Liberatore[2,3]

[1]Department of Computer Science, Autonomous University of Madrid, Spain, [2]IBiDat – Big Data Institute, Carlos III University of Madrid, Spain, and [3]School of Computer Science and Informatics, Cardiff University, UK
Corresponding author: Lara Quijano-Sánchez; Email: laraquij@inst.uc3m.es

(Received 25 September 2024; revised 11 March 2026; accepted 20 March 2026; first published online 30 April 2026)

Abstract

Twitter is one of the most widely used social networks globally but is also notorious for harbouring significant levels of hate speech. This paradox has garnered the attention of companies and governments concerned about how digital hate can fragment communities and incite real-world violence. While extensive research has focused on detecting hate speech, there is a lack of comprehensive analysis of the actors involved, their characteristics, and their interactions within the online environment. Common assumptions categorise users into merely two groups—haters and non-haters—overlooking the existence of other groups that may more accurately represent the dynamics of hate dissemination. Additionally, existing user classification models often rely on large volumes of tweet data, a limitation given the restricted access to the Twitter API. Social networks are also frequently visualised using graphs where edges represent only superficial relationships. This study addresses these gaps through five research questions. We propose formal user clustering methods and develop a classifier that uses exclusively profile attributes—information more readily obtainable from the network. We also introduce a more nuanced definition of interaction for graph edges, based on ideological support or opposition between users. To conduct our analysis, we have extracted two complementary datasets: (i) a keyword-based corpus of 3.3M Spanish tweets containing hate-associated terms, of which 1.6M unique tweets were retained after filtering and (ii) a user-based corpus comprising timelines of ≈3,000 users linked to hate speech, totalling over 3 M tweets. Our results reveal the existence of three primary user classes—haters, upstanders, and neutrals—in contrast to the conventional binary classification. We demonstrate that profile attributes are reliable indicators for automatically classifying users and find significant statistical differences between these classes. Finally, we develop a graph visualisation tool to assist authorities in analysing interactions among different user types, providing a useful exploratory tool to support the analysis of online hate and inform potential mitigation strategies.

Keywords: User profiling; social network analysis; hater; upstander; multiclass classifier

1. Introduction

With 557 million active users (Statista 2025), the platform Twitter, recently renamed X, is positioned as one of the most popular social networks globally. Although it was conceived as a real-time urban narrative with the goal of connecting people (Los Angeles Times, 2009), its great capacity to convey social approval through likes and retweets often encourages collaboration among individuals with negative attitudes and reinforces their prejudices (Walther, 2022). Additionally, Twitter has a distinctive feature: the brevity of its messages. With only 280 characters per tweet, users wishing to express intellectual opinions are forced to resort to comment threads or condense their content into lines that may come across as ambiguous or exaggerated

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial licence (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

(Wright, 2014). According to Jack Dorsey, co-founder of Twitter, the name of the application was chosen because, aside from meaning "bird chirping," it also refers to "a brief burst of inconsequential information" (Los Angeles Times, 2009). This suggests that the very design of Twitter promotes the dissemination of unfounded political arguments and biased opinions perhaps more than any other network. Furthermore, since Elon Musk assumed the role of CEO in October 2022, problematic content has significantly increased, and previously censored accounts have been reinstated (Walther, 2022).

The analysis of online hate, which attracts considerable interest from companies and governments (United Nations, 2024; Anti-Defamation League, 2024; Google LLC, 2024; Microsoft Corporation, 2024; Jigsaw, 2024), arises in response to this type of content on Twitter and other social networks. For instance, in Spain, since 2019, the Spanish National Office Against Hate Crimes (ONDOD) and the Spanish Observatory of Racism and Xenophobia (OBERAXE) have led the investigation of hate on Twitter through the ALRECO and REAL-UP projects.[a] These projects have resulted in research papers (Pereira-Kohatsu *et al.* 2019; del Valle-Cano *et al.* 2023), which are examples of many studies over the past few years that have focused on training and optimising classifiers on static datasets to determine the intent behind tweets (Nayla, Setianingsih, and Dirgantoro 2023; Singh, Malhotra, and S. 2023a; Turjya *et al.* 2023; Singh *et al.* 2023b). Another major line of research on Twitter hate is the development of monitoring tools that analyse datasets generated in real time (Dinh *et al.* 2023; Carhuancho-Bazan, Nuñez-Lazo, and Ugarte 2023; Gupta *et al.* 2023).

However, in the field of online hate speech, there are very few studies on user profiling (Noorian *et al.* 2024; Doğanç and Markov 2023). When expanding the search to a more general context, user profiling is more common (Lee *et al.* 2014; Heidari, Jones, and Uzuner 2020; Treves, Masud, and Faloutsos 2022). The term "profiling" refers to the analysis of characteristics and behaviours of large groups of users defined by an attribute, such as age (Nguyen *et al.* 2021), ethnicity (Rao *et al.* 2021), gender (Liu, Li, and Li 2021; Rao *et al.* 2021), or personality (Skowron *et al.* 2016). It is used in marketing to segment audiences and personalise products (Stakhiyevich and Huang 2019); in cybersecurity, to identify malicious actors (Treves, Masud, and Faloutsos 2022); to combat disinformation by characterising bots (Heidari, Jones, and Uzuner 2020); and to profile haters (Noorian *et al.* 2024). Additionally, studies have been conducted that use graphs to represent users involved in hate speech to understand complex patterns among user groups (del Valle-Cano *et al.* 2023; Perera, Perera, and Ahangama 2023). In these graphs, users are nodes, and relationships (follows, mentions, likes, etc.) are edges. The goals are to identify highly influential haters (Perera, Perera, and Ahangama 2023) and analyse the nature of interactions between different types of users (Chandra *et al.* 2021), among others.

The aim of this research, carried out in collaboration with the ONDOD and OBERAXE, is to understand the behaviour of users related to hate speech on Twitter, both at an individual level and within their social environment. The main focus is to meticulously describe the profiles of perpetrators and victims, as well as to create innovative tools through which the impact and dissemination of hate speech can be evaluated. This work contributes to the understanding of online hate dynamics, which may support the development of mitigation and prevention strategies undertaken by law enforcement and government organisations dedicated to it. The objective is to strengthen online hate monitoring capabilities and enhance counter-narrative strategies that combat hate speech stemming from racism, xenophobia, Islamophobia, anti-Semitism, and antigypsyism, among others.

After an exhaustive review of hate speech analysis on social networks, it has been observed that while hate message classifiers and real-time monitoring have been widely addressed (Nayla, Setianingsih, and Dirgantoro 2023; Turjya *et al.* 2023; Singh *et al.* 2023b), user profiling and graphs

[a]ALRECO project web: https://www.inclusion.gob.es/oberaxe/es/ejes/delitosodio/alreco/index.htm and REAL-UP project web: https://real-up.eu/proyecto/

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

as representations of the social network remain open lines of research. Regarding the former, besides the lack of studies in the field of hate (Chandra *et al.* 2021; He *et al.* 2022; Noorian *et al.* 2024), it has been observed that the definition of user classes is based on subjective criteria. In the particular case of hate, users are usually divided into only two groups (Rangel *et al.* 2021; Noorian *et al.* 2024; Mishra *et al.* 2019; An *et al.* 2021): haters and non-haters (hereafter, neutrals). However, it is possible that users naturally cluster into other distinct subsets. For example, into haters, extreme haters, and neutrals, or into haters and upstanders (people who oppose hate). Moreover, to automatically classify users using a machine learning model, different textual features derived from tweets and profile information are taken as input attributes (Rangel *et al.* 2021; Noorian *et al.* 2024; Mishra *et al.* 2019; An *et al.* 2021; Chandra *et al.* 2021; He *et al.* 2022), but it has not been explored whether this can be done solely with profile data. Regarding the latter, edges typically represent superficial relationships (follows, mentions, likes, etc.; del Valle-Cano *et al.* (2023); Perera *et al.* (2023)), meaning they do not always translate into meaningful personal relationships.

In light of these observed limitations, the following research questions are formulated:

- **RQ1**: What objective techniques can classify users based on their stance towards hate?
- **RQ2**: What underlying groups exist within the set of users with respect to hate?
- **RQ3**: What profile characteristics are specific to each type of user?
- **RQ4**: How do users behave in terms of (a) the topics on which they express hate and (b) the virality of their hate messages?
- **RQ5**: How are relationships of support and opposition between users reflected in a social network through the structure of a graph?

To address the RQs, we have relied on two datasets. First, a keyword-based corpus of 3.3 million tweets has been collected using hate-associated terms; after filtering, 1.6 million unique tweets remained for analysis. Second, from this dataset, we identified approximately 3,000 users and extracted their timelines, resulting in over 3 million tweets in total.[b] To answer **RQ1** and **RQ2**, clustering techniques have been employed to avoid imposing an arbitrary limit on the percentage of hate emitted by a user. To answer **RQ3**, a classifier has been trained based on profile information, and a descriptive analysis has been conducted to determine the most significant characteristics of each type of user. Both explicit characteristics (number of tweets and followers, time spent on Twitter, etc.) and implicit characteristics (gender, personality, anonymity, etc.) have been considered. To answer **RQ4**, statistical techniques have been used to model and understand user behaviour patterns. Finally, to answer **RQ5**, an interactive graph visualisation tool has been developed to analyse the interactions between each type of user.

The rest of the document is structured as follows: Section 2 provides a review of the most relevant state-of-the-art literature according to the objectives of this project and justifies the research questions. Section 3 describes the methods and algorithms developed to profile users and analyse their interactions. Section 4 presents the datasets used and the procedure implemented to extract them. Section 5 illustrates the experiments conducted and discusses the results. Finally, Section 6 summarises the conclusions drawn, highlights the limitations of this research, and outlines possible future research directions.

2. State of the art

To gain a comprehensive understanding of the research on hate speech and user behaviour on Twitter, we have conducted a systematic bibliographic review following the approach of Bramer

[b] Unless otherwise stated, analyses are based on the dataset most appropriate to the specific research question.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

et al. (2018). To establish the novelty of this study, we began with a broad search for existing tools used to detect and monitor online hate speech. This review revealed that the areas of hate user profiling and hate interaction analysis remain mainly under-explored, making them focal points for further investigation in the following subsections. The section concludes by outlining the limitations identified in the current literature, which serve as the basis for the objectives of this study.

2.1 Hate speech detection studies

Hate speech encompasses discriminatory language, insults, intimidation, and the use of stereotypes targeting specific groups based on factors such as ethnicity, sexuality, or gender. While it is widely acknowledged as a serious issue on social media, manually detecting and moderating such behaviour in real time is nearly unfeasible due to the sheer volume and rapid pace of content generation. Consequently, numerous studies have focused on developing automated tools to combat online hate speech. This section reviews the two primary research areas in hate speech detection: the classification of hate speech messages and real-time monitoring systems.

2.1.1 Hate speech message classification

The majority of research on hate speech within social networks (see Table 1 for a summary of the most relevant references), approximately 95% of search results, centres on the classification of hate messages. These studies typically involve training machine learning models using static datasets where each text segment is labelled as either hate or non-hate (Pereira-Kohatsu *et al.* 2019; Briliani, Irawan, and Setianingsih 2019). Although binary classification is the most common approach, there are instances where multiclass classification is employed (Sachdeva *et al.* 2021; Qureshi and Sabih 2021; Jahan *et al.* 2022; del Valle-Cano *et al.* 2023). For example, Jahan *et al.* (2022) categorise sexism into subtypes such as stereotyping and objectification, domination, sexual harassment, and threats of violence, while del Valle-Cano *et al.* (2023) classify tweets as violent hate, non-violent hate, counter-narratives, or neutral. While much of the literature addresses general hate speech (Briliani, Irawan, and Setianingsih 2019), some studies focus specifically on sexist (Sachdeva *et al.* 2021; Jahan *et al.* 2022) or racist content (Alotaibi and Abul Hasanat 2020; Sachdeva *et al.* 2021), or on hate speech tied to specific events such as the COVID-19 pandemic (Bhimani *et al.* 2021) or political elections (Febriana and Budiarto, 2019). Additionally, while some studies narrow their scope by defining hate speech strictly, others expand it to include both discriminatory hate and offensive language, covering general insults without explicit racial or sexist motivations (Martins *et al.* 2018; Chhikara and Malik 2022; Turjya *et al.* 2023; Singh, Malhotra, and S. 2023a).

These systems primarily leverage natural language processing (NLP) techniques, which typically involve preprocessing the text prior to training by removing stopwords, special characters, emoticons, and more, while also correcting spelling errors and applying tasks such as stemming, standardisation, and tokenisation (Martins *et al.* 2018; Herodotou, Chatzakou, and Kourtellis 2021; Plaza-Del-Arco *et al.* 2021; Panchala *et al.* 2022; Shrimali, 2022; Nayla, Setianingsih, and Dirgantoro 2023). After preprocessing, the text is converted into numerical vectors using methods such as word2vec or GloVe to make the text "machine-readable" (Syam, Irawan, and Setianingsih 2019; Bhimani *et al.* 2021; Sachdeva *et al.* 2021; Shrimali, 2022). In some instances, TF-IDF is also employed to identify the most significant words in the dataset, which the classifier then uses (Martins *et al.* 2018; Briliani, Irawan, and Setianingsih 2019; Roy *et al.* 2020; Defersha, Kekeba, and Kaliyaperumal 2021; Qureshi and Sabih 2021; Jahan *et al.* 2022; Singh *et al.* 2023b). Classifiers rely on traditional machine learning models such as Naive Bayes (NB), random forest (RF), decision trees (DT), K-nearest neighbours, support vector machine (SVM), and logistic regression (LR) (Martins *et al.* 2018; Briliani, Irawan, and Setianingsih 2019; Roy *et al.* 2020; Defersha, Kekeba, and Kaliyaperumal 2021; Bhimani *et al.* 2021; Herodotou, Chatzakou, and

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

Table 1. Summary of hate speech classification studies

	Number of classes	Hate focus	Origin of datasets	Platform	Input for classifier	Language
Alotaibi and Abul Hasanat (2020)	2 (Racist, Not-racist)	Racism	Collected via keywordsand/or news sources	Twitter	Text	Arabic
Briliani *et al.* (2019)	2 (Hate, Non-hate)	Hate	Collected via keywordsand/or news sources	Instagram	Text, relative importance ofwords	Indonesian
del Valle-Canoet al. (2023)	4 (Violent hate, Non-violent hate,Counter-narrative, Neutral)	Hate	From data repositories	Twitter	Text, linguistic features, userprofile details and socialnetwork	Spanish
Jahan *et al.* (2022)	4 (Stereotype and Objectification,Dominance and Derailing, Harassment andThreats, Non-misogynistic)	Sexism	Collected via keywordsand/or news sources	Facebook andYouTube	Text, relative importance ofwords	Bengali
Martins *et al.* (2018)	3 (Hate, Offensive, Neither)	Hate and offence	From data repositories	Twitter	Text, emotion, keywordpresence, sentiment	English
Plaza-Del-Arco et al. (2021)	2 (Hate, Non-hate)	Hate	From data repositories	Twitter	Text, emotion, sentiment	Spanish
Qureshi and Sabih (2021)	11 (African hate, Arab hate, Asian hate,Christian hate, Islam hate, Jews hate,Racism, Refugees hate, Gender hate,Sexuality hate, Non-hate)	Hate	From data repositories	Twitter	Text, relative importance ofwords, linguistic features,sentiment	English
Roy *et al.* (2020)	2 (Hate, Non-hate)	Hate	From data repositories	Twitter	Text	English
Sachdeva *et al.* 2021	3 (Racist, Sexist, Non-hate)	Sexism and racism	From data repositories	Twitter	Text	English
Singh *et al.* (2023a)	2 (Offensive, Not-offensive)	Offence	From data repositories	Discord	Text, relative importance ofwords, visual content	English
Shukla, Nagpal, and Sabharwal, 2022	2 (Hate, Non-hate)	Hate	From data repositories	Twitter and Youtube	Text	Hindi

Kourtellis 2021; Shrimali, 2022; Panchala *et al.* 2022; Jahan *et al.* 2022). Among these, SVM often performs best (Martins *et al.* 2018; Defersha, Kekeba, and Kaliyaperumal 2021; Panchala *et al.* 2022; Jahan *et al.* 2022). However, recent research has shifted towards deep learning models for their ability to detect finer nuances in language. These include architectures such as convolutional neural networks (CNN) (Roy *et al.* 2020; Shukla, Nagpal, and Sabharwal 2022), long short-term memory (LSTM) (Syam, Irawan, and Setianingsih 2019; Roy *et al.* 2020; Dinh *et al.* 2023), and transformer models, with BERT being the most widely adopted (Plaza-Del-Arco *et al.* 2021; Carhuancho-Bazan, Nuñez-Lazo, and Ugarte 2023; Turjya *et al.* 2023). Deep learning models have been shown to outperform traditional approaches (Roy *et al.* 2020; Carhuancho-Bazan, Nuñez-Lazo, and Ugarte 2023).

Despite these advancements, a significant challenge remains in identifying implicit hate speech, where abusive intent is conveyed through irony, sarcasm, or stereotypes without the use of explicit offensive terms. This complexity arises from the creative and cognitive aspects of language, as explored by Frenda *et al.* (2023), who highlight how implicitness complicates the detection of abusive content online. To address this, recent research has explored the synergy between different architectures; for instance, Damo *et al.* (2026) propose integrating large language models (LLMs) with traditional detectors to not only improve identification but also generate faithful and plausible explanations for why a message is considered implicitly hateful.

The datasets used for training are static, meaning they do not change during the training process. Many studies utilise pre-existing datasets from Kaggle (Roy *et al.* 2020; Panchala *et al.* 2022) or other research sources (Martins *et al.* 2018; Plaza-Del-Arco *et al.* 2021; Pawar *et al.* 2022; Shrimali, 2022). When a dataset with specific attributes is required, data are collected via official APIs from various platforms, using keywords and/or news sources (Briliani, Irawan, and Setianingsih 2019; Alotaibi and Abul Hasanat 2020; Febriana and Budiarto 2019; Jahan *et al.* 2022). Twitter is the most frequently used platform (Syam, Irawan, and Setianingsih 2019; Roy *et al.* 2020; Shukla, Nagpal, and Sabharwal 2022; Turjya *et al.* 2023; Alotaibi and Abul Hasanat 2020; Febriana and Budiarto 2019; Plaza-Del-Arco *et al.* 2021; Sachdeva *et al.* 2021; Panchala *et al.* 2022; Pawar *et al.* 2022; Shrimali, 2022), though other platforms such as Facebook (Defersha, Kekeba, and Kaliyaperumal 2021; Jahan *et al.* 2022), YouTube (Shukla, Nagpal, and Sabharwal 2022; Jahan *et al.* 2022), Discord (Singh, Malhotra, and S. 2023a), Wikipedia (Shrimali, 2022), and Instagram (Briliani, Irawan, and Setianingsih 2019) have also been studied. These texts are typically manually labelled, often through crowdsourcing. Dataset sizes vary, ranging from thousands (Briliani, Irawan, and Setianingsih 2019; Plaza-Del-Arco *et al.* 2021; Jahan *et al.* 2022; Shukla, Nagpal, and Sabharwal 2022; Turjya *et al.* 2023), to tens of thousands (Roy *et al.* 2020; Defersha, Kekeba, and Kaliyaperumal 2021; Bhimani *et al.* 2021; Pawar *et al.* 2022; Panchala *et al.* 2022), and even hundreds of thousands (Singh, Malhotra, and S. 2023a). These manually labelled datasets tend to be imbalanced, as significantly more neutral messages are extracted from networks compared to hate messages.

Some models rely solely on the message text for predictions, minimising the need for manual feature extraction. However, other models explore the benefits of incorporating additional features, such as the relative importance of words derived from the TF-IDF method (Martins *et al.* 2018; Briliani, Irawan, and Setianingsih 2019; Roy *et al.* 2020; Defersha, Kekeba, and Kaliyaperumal 2021; Qureshi and Sabih 2021; Jahan *et al.* 2022; Singh *et al.* 2023b), linguistic features (e.g., uppercase ratio, message length, and offensive word count) (Qureshi and Sabih 2021; del Valle-Cano *et al.* 2023), sentiment analysis (Martins *et al.* 2018; Plaza-Del-Arco *et al.* 2021; Qureshi and Sabih 2021; Bhimani *et al.* 2021), and emotional categories (e.g. anger, fear, sadness, joy) (Martins *et al.* 2018; Plaza-Del-Arco *et al.* 2021). Additional parameters include user profile details (e.g., number of followers) (Herodotou, Chatzakou, and Kourtellis 2021; del Valle-Cano *et al.* 2023) and social network graph centrality measures (del Valle-Cano *et al.* 2023). Emerging models also consider visual content instead of text (Singh, Malhotra, and S. 2023a; M and P, 2023).

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

Table 2. Overview of real-time monitoring studies

	Task	Platform	Language
Carhuancho-Bazan *et al.* (2023)	Alert instantly of hate speech	Discord	English
Caldera *et al.* (2020b)	Periodically extract, process and visualise data	Twitter	Sinhalese
Dinh *et al.* (2023)	Periodically extract, process and visualise data	Facebook	Vietnamese
Doan *et al.* (2022)	Periodically extract, process and visualise data	YouTube	Vietnamese
Gupta *et al.* (2023)	Create and evaluate sole modules of a bigger pipeline	Twitch	Spanish
Herodotou *et al.* (2021)	Create and evaluate sole modules of a bigger pipeline	Twitter	English
Pereira-Kohatsu *et al.* (2019)	Periodically extract, process and visualise data	Twitter	Spanish
Shrimali (2022)	Create and evaluate sole modules of a bigger pipeline	Twitter, Wikipedia	English

Most models are designed for English, but there are also models developed for other languages, such as Arabic (Alotaibi and Abul Hasanat 2020), Indonesian (Febriana and Budiarto 2019; Briliani, Irawan, and Setianingsih 2019; Nayla, Setianingsih, and Dirgantoro 2023), Bengali (Jahan *et al.* 2022), Hindi (Shukla, Nagpal, and Sabharwal 2022), German (Turjya *et al.* 2023), and Spanish (Pereira-Kohatsu *et al.* 2019; Plaza-Del-Arco *et al.* 2021; Turjya *et al.* 2023; del Valle-Cano *et al.* 2023). For Spanish, the most prominent models are HaterBERT and SocialHaterBERT (del Valle-Cano *et al.* 2023), achieving macro-F1 scores of 0.9754 and 0.9233, respectively. These models are currently employed by the government agencies OBERAXE and ONDOD to monitor and address hate speech on Twitter.

2.1.2 Real-time monitoring

In terms of real-time monitoring, there are fewer studies compared to those focused on hate speech classification (see Table 2 for an overview). Among the existing works, most develop custom text classification models, which are then integrated into real-time data processing modules. Some research goes further by creating a comprehensive pipeline that periodically extracts data from social networks, classifies it by hate speech type, and visualises the results through statistics, such as the proportion of hate speech and the most frequently used terms (Pereira-Kohatsu *et al.* 2019; Doan *et al.* 2022; Dinh *et al.* 2023). In other cases, the objective is not to visualise statistics but to provide instant alerts during hate speech incidents. For instance, Carhuancho-Bazan *et al.* (2023) developed an API for Discord that allows moderators to monitor their channels. Additionally, some studies focus solely on assessing the scalability of processing modules, with the intention of connecting them to a continuous data flow in the future (Herodotou, Chatzakou, and Kourtellis 2021; Shrimali, 2022; Gupta *et al.* 2023). These studies typically use existing datasets and simulate a continuous data stream rather than extracting new data (Herodotou, Chatzakou, and Kourtellis 2021; Shrimali, 2022).

A variety of platforms have been studied in this field, including Twitter (Pereira-Kohatsu *et al.* 2019; Caldera, Meedin, and Perera 2020b; Herodotou, Chatzakou, and Kourtellis 2021), Discord (Carhuancho-Bazan, Nuñez-Lazo, and Ugarte 2023), YouTube (Doan *et al.* 2022), Facebook (Dinh *et al.* 2023), and Twitch (Gupta *et al.* 2023). Although English was expected to be the dominant language in these studies, it is noteworthy that five of the eight studies focus on other languages: Spanish (Pereira-Kohatsu *et al.* 2019; Gupta *et al.* 2023), Vietnamese (Doan *et al.* 2022; Dinh *et al.* 2023), and Sinhalese (Sri Lanka) (Caldera, Meedin, and Perera 2020b).

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

2.2 User profiling studies

User profiling in the context of social networks involves analysing user characteristics, behaviours, and preferences to create representative profiles for large user groups. Its applications are wide ranging: in marketing, it is used to segment audiences and provide personalised product or service recommendations through recommendation systems (Lee *et al.* 2014; Stakhiyevich and Huang 2019). Beyond commercial purposes, user profiling is critical in cybersecurity for identifying malicious actors (Treves, Masud, and Faloutsos 2022), in combating disinformation by characterising bots (Heidari, Jones, and Uzuner 2020), and in addressing online hate speech by profiling those who spread hate (Rangel *et al.* 2021; Noorian *et al.* 2024; Mishra *et al.* 2019; Pereira-Kohatsu *et al.* 2019; An *et al.* 2021; Chandra *et al.* 2021; He *et al.* 2022). Due to the limited research on this last topic, the literature review has been expanded to include general user profiling studies. This section first reviews user profiling research broadly and then examines how individuals spreading hate speech are classified and characterised in the literature.

The primary focus of research in user profiling typically assumes that individuals consistently belong to a single class, although some studies examine class transitions over time (He *et al.* 2022), even when the class definitions vary (Vasanthakumar, Shashikumar, and Suresh 2019). User classes are commonly defined based on attributes such as age (Pandey and Mittal 2016; Nguyen *et al.* 2021), ethnicity (Rao *et al.* 2021), gender (Burger *et al.* 2011; Pandey and Mittal 2016; Priadana, Maarif, and Habibi 2020; Liu, Li, and Li 2021; Rao *et al.* 2021), geographic location (Alonso-Lorenzo, Costa-Montenegro, and Fernández-Gavilanes 2016; Williams, Gray and Dixon 2017; Utomo, Adji, and Ardiyanto 2018), personality (Golbeck, Robles, and Turner 2011; Skowron *et al.* 2016), marital status (Pandey and Mittal 2016), and anonymity level (Peddinti, Ross, and Cappos 2014; Nakano *et al.* 2016). For instance, Nguyen *et al.* (2021) categorise users into age groups (under 20, 20–40, and over 40), while Peddinti *et al.* (2014) classify users by anonymity levels (anonymous, partially anonymous, identifiable, or highly identifiable). Typically, the first step involves defining user classes and providing clear descriptions of each. These descriptions are then used by those responsible for assigning users to a class, often through manual methods such as crowdsourcing (Burger *et al.* 2011; Peddinti, Ross, and Cappos 2014), though researchers themselves (Nguyen *et al.* 2021; Rangel *et al.* 2021; He *et al.* 2022) or even users (Golbeck, Robles, and Turner 2011; Skowron *et al.* 2016) sometimes carry out this task. Some studies also aim to automate the classification process, defining user classes based on profile attributes and publications (Rao *et al.* 2021). After classifying users, several studies aim to statistically describe each class or identify correlations between classes and other user characteristics (Nakano *et al.* 2016; Rao *et al.* 2021; He *et al.* 2022).

Most studies focus on a single platform, such as Twitter or Facebook, though some research explores linking accounts across multiple social networks to complement the partial data from individual platforms (Kaushal, Ghose, and Kumaraguru 2019; Treves, Masud, and Faloutsos 2022). Data are usually collected via official APIs (Chandra *et al.* 2021; He *et al.* 2022) or from existing datasets (Liu, Li, and Li 2021; Noorian *et al.* 2024). Using this data and manually labelled corpora, machine learning models are trained to classify users (Burger *et al.* 2011; Golbeck, Robles, and Turner 2011; Pandey and Mittal 2016; Alonso-Lorenzo, Costa-Montenegro, and Fernández-Gavilanes 2016; Skowron *et al.* 2016; Williams, Gray and Dixon 2017; Nguyen *et al.* 2021; Liu, Li, and Li 2021; Rao *et al.* 2021). The input data often consists of textual features from users' posts, such as the average use of capital letters, post length, and offensive word frequency, particularly in tweets (Burger *et al.* 2011; Peddinti, Ross, and Cappos 2014; Nguyen *et al.* 2021). In essence, these classification models primarily rely on analysing user posts. For Twitter, users' timelines[c] are accessed to extract frequently occurring textual features, and NLP techniques such as Named Entity Recognition (NER) and Sentiment Analysis are applied. Additional features such as sociolinguistic markers (emoticons, repeated letters, etc.), profile data (Skowron *et al.* 2016;

[c]The term timeline refers to the set of the last or all of the publications of a user on the network.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

Alonso-Lorenzo, Costa-Montenegro, and Fernández-Gavilanes 2016; Williams, Gray and Dixon 2017; Liu, Li, and Li, 2021; Rao *et al.* 2021), and social environment characteristics (Golbeck, Robles, and Turner, 2011; Alonso-Lorenzo, Costa-Montenegro, and Fernández-Gavilanes 2016; Williams, Gray and Dixon 2017) are also considered. Table 3 lists the most representative studies based on these characteristics.[d]

Another example is personality profiling (Golbeck, Robles, and Turner 2011; Skowron *et al.* 2016), where the Big Five model is used due to its widespread recognition in psychology, though it might be more appropriate to carefully select personality groups that are more relevant on Twitter. A major limitation in user profiling studies is that user classification is often arbitrary. The categories used to classify users are typically based on assumptions rather than a rigorous scientific process. For instance, when segmenting users by age, researchers often create more categories for younger users, assuming most Twitter users are young. However, an alternative approach could assign equal numbers of classes across all age groups. Even when more categories are defined for younger users, there is no consensus on the exact ranges. For example, Pandey and Mittal (2016) define age groups as under 20, 20–40, and over 40, while Nguyen *et al.* (2021) use 16–20, 21–25, 26–30, and over 30. Similarly, studies on gender (Burger *et al.* 2011; Peddinti, Ross, and Cappos 2014; Liu, Li, and Li 2021; Rao *et al.* 2021) typically classify users as male or female, even though this may not accurately represent the significant percentage of Twitter users who do not identify with either gender. In personality profiling studies (Golbeck, Robles, and Turner 2011; Skowron *et al.* 2016), the Big Five model is commonly used due to its recognition in psychology, though a more tailored selection of personality traits might be more relevant for Twitter users.

2.2.1 User profiling studies in the context of hate speech

In the field of online hate speech, there are relatively few user profiling studies (Rangel *et al.* 2021; Noorian *et al.* 2024; Mishra *et al.* 2019; Pereira-Kohatsu *et al.* 2019; An *et al.* 2021; Chandra *et al.* 2021; He *et al.* 2022; Doğanç and Markov 2023). Some of these studies address hate speech broadly, defining it as discriminatory behaviour targeting individuals or groups based on characteristics such as gender, religion, sexual orientation, nationality, or ethnicity (Pereira-Kohatsu *et al.* 2019; Rangel *et al.* 2021). Others focus on specific types of hate, including anti-Asian hate (An *et al.* 2021; He *et al.* 2022; Noorian *et al.* 2024), Islamophobia (Chandra *et al.* 2021), sexism (Mishra *et al.* 2019), and racism (Mishra *et al.* 2019). Similar to general profiling research, some rely on existing datasets (An *et al.* 2021; Noorian *et al.* 2024; Mishra *et al.* 2019), while others create their own (Chandra *et al.* 2021; He *et al.* 2022). A key difference in this context is that user labelling is automated rather than manual. These studies assume that user classes are predefined based on the proportion of hate speech in their tweets. For instance, if 90% of a user&'s tweets are classified as hateful, these studies assume that manual labelling is not required. Instead, they adopt a rule-based approach, where users are automatically categorised based on the proportion of hateful content in their posts. A common rule is to label users with more than 50% hateful tweets as "haters" and those with 50% or fewer as "non-haters." After tweets are classified using an NLP model, a program can assign users to categories. Most studies use NLP models for classification (An *et al.* 2021 ; Mishra *et al.* 2019; He *et al.* 2022), while some rely on hate-related keywords (Noorian *et al.* 2024), and a few use manual classification (Rangel *et al.* 2021).

In this context, users are often categorised as either haters or non-haters (Rangel *et al.* 2021; He *et al.* 2022; Noorian *et al.* 2024; Mishra *et al.* 2019; An *et al.* 2021). However, this binary classification is based on the unfounded assumption that all users fit into one of these two groups, reflecting the issue of arbitrariness mentioned in the previous section. It is possible that other classifications could exist, such as distinguishing between mild haters and extreme haters, or between haters and upstanders. Moreover, despite various studies employing similar hypotheses, there is no consensus on a precise definition of a "hater." For instance:

[d]Some studies mentioned are excluded from the table due to differing goals or approaches.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

Table 3. Summary of works focused on user profiling based on different attributes

Paper	Attribute	User classes	Attributes used for labelling	Labelling method	Input attributes for the classifier	Platform
Nguyen *et al.* (2021)	Age	20-, 20–40, 40+	User profile, tweets, additional external resources like Facebook or LinkedIn	Manual (two students)	Textual features derived from user timelines	Twitter (Dutch)
Pandey and Mittal (2016)	Age, gender and marital status	16–20, 21–25, 26-30, 30+ Woman, man Single, in a relationship	Keywords in user timelines	Automatic	Topics extracted from user timelines	Tumblr (English)
Peddinti *et al.* (2014)	Anonymity	Anonymous, partially anonymous, identifiable, highly identifiable	*custom name*, *identifiable name*, profile URLs	Manual (Amazon Mechanical Turk)	-	Twitter (English)
Burger *et al.* (2011)	Gender	Woman, man	Features derived from user timelines	Manual (Amazon Mechanical Turk)	Textual features derived from user timelines	Twitter (English)
Liu *et al.* (2021)	Gender	Woman, man	Message text	Pre-labelled dataset Automatic (sentiment classifier)	Textual features derived from user timelines, profile and account information, sentiment metrics	Weibo (Chinese)
Rao *et al.* (2021)	Gender and ethnicity	Woman, man Four main ethnicities of Nigeria	Features derived from usernames and timelines	Automatic (proposed methodology)	Textual features derived from user timelines and usernames	Facebook (English)
Alonso-Lorenzo *et al.* (2016)	Location	Country	Geotagged tweets	Automatic (tweet metadata)	Textual features derived from user timelines, profile information, social connection information	Twitter (English)
Williams *et al.* (2017)	Location	Place with a 5 km error	Geotagged tweets	Automatic (tweet metadata)	Textual features derived from user timelines, profile information, social connection information	Twitter (English)
Golbeck *et al.* (2011)	Personality	Big Five	Questionnaire responses	Manual (user questionnaire)	Public profile information, social connection information	Facebook (English)
Skowron *et al.* (2016)	Personality	Big Five	Questionnaire responses	Manual (user questionnaire)	Textual features derived from user timelines, account information, profile picture	Twitter and Instagram (English)

- He *et al.* (2022) and Noorian *et al.* (2024) define haters as those who have posted at least one hate tweet and no upstander tweets.
- Rangel *et al.* (2021) define haters as users with more than 10 hate tweets.
- Mishra et al. (2019) define haters as those who have posted more than one hate tweet.
- An *et al.* (2021) define haters as those who have posted more than two hate tweets.

Additionally, some studies classify users based on characteristics such as age, gender, and origin, rather than by the volume of hate speech they produce, aiming to create personalised counter-narrative messages (Doğanç and Markov 2023).

After categorising users as either haters or non-haters, the latter group is often used as a "control" group to examine the differences and similarities between the two. Various analyses are conducted based on their timelines: linguistic analyses, such as identifying the most representative words for each group (An *et al.* 2021; Chandra *et al.* 2021; Noorian *et al.* 2024); psychological analyses, such as personality profiling (Noorian *et al.* 2024); and social analyses, such as comparing how each group&'s tweets are received (An *et al.* 2021).

Most studies reviewed subsequently develop predictive models aimed at identifying individuals who propagate online hate speech. These models typically use textual features extracted from users' timelines as input (Rangel *et al.* 2021; Noorian *et al.* 2024; Mishra *et al.* 2019; An *et al.* 2021), allowing for the evaluation of the effectiveness of these features based on predictive performance (An *et al.* 2021; Noorian *et al.* 2024). Additionally, some models are trained to assess whether incorporating user profiling improves the performance of hate speech classifiers (Mishra *et al.* 2019). Table 4 offers a detailed comparison of the six most relevant studies on user profiling in the context of hate.[e]

2.3 User graph studies

Representing user relationships through graphs is a powerful technique for uncovering complex patterns within thematic groups (Wang *et al.* 2016; Ahmed, Aurpa, and Anwar 2020), often outperforming traditional methods in some cases (Bozhidarova *et al.* 2023). In hate speech research, graphs are crucial in two key areas: hate tweet classification and the identification of influential individuals and groups.

In hate tweet classification, graph centrality measures are frequently used as input for classifiers (Beatty, 2020; Mishra *et al.* 2019; del Valle-Cano *et al.* 2023). However, there has been a growing trend toward utilising neural models designed for graph-structured data, known as graph neural networks (GNNs) (Hebert, Golab, and Cohen 2022; Duong, Zhang, and Lu 2022; Harby and Zulkernine 2023). While these studies generally find that representing social networks as graphs does not outperform the most advanced NLP techniques, it remains useful for classification purposes.

In the second area, studies such as Pereira-Kohatsu *et al.* (2019), Perera *et al.* (2023), and Oroh *et al.* (2021) focus on detecting influential haters, whereas others, such as Ting *et al.* (2013), define user clusters and characterise them based on their hater status by analysing the most frequent words and interconnections within the clusters. Additionally, Chandra *et al.* (2021) create a connection graph to analyse interactions between different user groups, revealing that highly hateful users tend to be closely interconnected.

In the reviewed studies (see Table 5 for an overview), nodes and edges are used to represent various elements within graphs:

- **Nodes** typically represent users, though some studies classify them (Pereira-Kohatsu *et al.* 2019; Chandra *et al.* 2021), while others do not (Mishra *et al.* 2019; Beatty, 2020; Caldera,

[e]Two studies are excluded from the table as they are not directly comparable with the others.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

Table 4. Overview of studies focused on user profiling in the context of hate speech

Study	Attribute	User classes	Attributes used for labelling	Tweet labelling method	Input features for the classifier	Platform
Noorian *et al.* (2024)	Hate speech emission	Hater or not (defined as someone with at least 1 hate tweet and no counter-narrative tweet)	-	Pre-labelled dataset	Textual features derived from users' timelines	Twitter (English)
Rangel *et al.* (2021)	Hate speech emission	Hater or not (defined as someone with more than 10 hate tweets)	Type of tweets (hate or not)	Manual (by the researchers themselves)	Textual features derived from users' timelines	Twitter (English and Spanish)
Mishra *et al.* (2019)	Hate speech emission	Hater or not (defined as someone with at least 1 hate tweet)	Type of tweets (hate or not)	Automated (text classifier)	Users' timelines and online social environment features	Twitter (English)
An *et al.* (2021)	Hate speech emission	Hater or not (defined as someone with at least 2 hate tweets)	Type of tweets (hate or not)	Automated (based on keyword containment)	Textual features derived from users' timelines and profile traits	Twitter (English)
He *et al.* (2022)	Hate speech emission	Hater, upstander, dual (defined as someone with at least one hate tweet and one upstander tweet) and neutral	Type of tweets (hate or not)	Manual and automated (text classifier based on a manually annotated dataset)	Textual features derived from users' timelines and profile traits	Twitter (English)
Chandra *et al.* (2021)	Hate speech emission	Very hater, hater or low hater ($>$50%, 25–50%, $<$25% of hate tweets)	Type of tweets (hate or not)	Manual and automated (text classifier based on a manually annotated dataset)	-	Twitter (English)

Table 5. Overview of user graph studies

Study	Task	Nodes	Edges
Beatty (2020)	Hate speech detection	Non-classified users	Retweeted
Bozhidarova *et al.* (2023)	Analyse evolution of network	Non-classified users and topics	Liked, commented or retweeted
Caldera *et al.* (2020a)	Hate speech detection	Non-classified users and posts	Complex definition
Chandra *et al.* (2021)	analyse interactions between different usergroups	Classified users	Following
Mishra *et al.* (2019)	Hate speech detection	Non-classified users	Following
Oroh *et al.* (2021)	Detect influential haters	Non-classified users	Mentioned and hashtagged
Perera *et al.* (2023)	Detect influential haters	Classified users	Mentioned
Pereira-Kohatsu *et al.* (2019)	Detect influential haters	Classified users	Mentioned
Ting *et al.* (2013)	Identify hateful users by clusteringcommon words	Frequent words	Complex definition

Meedin, and Perera 2020a; Oroh, Bandung, and Zagi 2021; Bozhidarova *et al.* 2023). In some cases, posts (Caldera, Meedin, and Perera 2020a) and topics (Bozhidarova *et al.* 2023) are also represented as nodes. For studies that classify users, profiling techniques (see Section 2.2) are used to group users based on the amount of hate they produce. For example, Pereira-Kohatsu *et al.* (2019) divide users into red nodes (haters) and blue nodes (hate recipients), while Chandra *et al.* (2021) categorise them into red nodes (highly hateful users), yellow nodes (moderately hateful), and green nodes (low hateful users). Most studies include isolated users, though there are exceptions (Mishra *et al.* 2019).

- **Edges** represent relationships between users, which are defined in various ways. A user can be connected to another by following them (Mishra *et al.* 2019; Chandra *et al.* 2021), mentioning them (Pereira-Kohatsu *et al.* 2019), interacting with them (e.g., liking, commenting, and retweeting) (Bozhidarova *et al.* 2023), or through more complex relationships (Ting *et al.* 2013; Caldera, Meedin, and Perera 2020a). Edges can be directional (Pereira-Kohatsu *et al.* 2019; Caldera, Meedin, and Perera 2020a; Chandra *et al.* 2021) or bidirectional (Mishra *et al.* 2019; Oroh, Bandung, and Zagi 2021). The commonality across all studies is that edges are unweighted.

While most graph-based studies focus on Twitter, others examine platforms such as Facebook (Ting *et al.* 2013; Caldera, Meedin, and Perera 2020a), YouTube (Caldera, Meedin, and Perera 2020a), and Reddit (Hebert, Golab, and Cohen 2022).

2.4 Limitations and research gaps

After an exhaustive review of hate analysis on social media, it has become evident that while hate message classification and real-time monitoring have been thoroughly investigated, user profiling and graph-based representations of social networks remain underexplored areas of research.

In terms of user profiling, our first goal is to move away from arbitrary classifications and instead develop a partitioning of users based on objective, scientifically grounded methods. We

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

aim to establish more complex classification criteria that account for multiple variables, rather than relying on a single factor, such as the number of hate tweets, to provide a more accurate definition of a "hater" user. This raises the question of whether users can indeed be categorised into haters and non-haters, as many studies have suggested (see Table 1).

Second, we inquire whether a user can be automatically classified based solely on profile data, without including linguistic traits or other characteristics derived from their timelines. This contrasts with the approach taken by most existing profiling studies (see Table 4). This question is particularly pertinent following Twitter&'s restriction of API access in February 2023, which now requires a monthly fee to download tweets (Calma, 2023). However, profile information (e.g., username, follower count, and account age) remains easily accessible. To address these gaps, we propose research questions **RQ1**, **RQ2**, and **RQ3**. Similar to other profiling studies (Nakano *et al.* 2016; Rao *et al.* 2021; He *et al.* 2022), the resulting user classification will be used to understand behavioural patterns and differences in online activity, leading to research question **RQ4**.

Regarding graph-based representations, existing studies have not found edges that represent more meaningful connections between users beyond superficial relationships like follows or mentions (see Table 5).These relationships are considered superficial because they are "digital" and do not necessarily reflect significant real-world interactions. The study with the most "real" connections (Pereira-Kohatsu *et al.* 2019) interprets mentions not just as digital links but as oppositional interactions between users. Specifically, it assumes that if one user mentions another in a hate tweet, it reflects either an insult or a connection to disliked topics. However, we challenge this assumption, as we suspect that many users who mention each other do so in agreement, reinforcing one another&'s views. To test this theory and expand the concept of online connections to include more genuine, human interactions, we propose a new definition of user interaction based on whether they support or oppose each other&'s ideologies. This leads to our final research question **RQ5**.

3. Materials and methods

In this section, we describe the proposed procedures to address the identified research gaps in the literature and to answer the posed research questions.

3.1 User profiling

This section presents a methodology for profiling and modelling users in the context of social networks. Although it is initially designed to characterise users based on the proportion of hate tweets they write, it goes beyond this initial objective as it can be adapted to other social platforms. The methodology consists of four steps motivated by the RQs: clustering, feature extraction, classifier training, and extrapolation of insights.

To answer **RQ1**, we propose using *cluster analysis* methods as scientific and objective procedures for user segmentation. To address **RQ2** and discover the underlying number of groups, a study must be conducted to determine which input data to use and which specific clustering algorithm is most optimal. These input data should align with the particular study domain to effectively segment users. In our case, the data should pertain to the domain of hate on Twitter. Additionally, the choice of clustering algorithm must be carefully considered, especially in terms of the data, to optimise both the accuracy and interpretability of the results. The next step in the methodology is user characterisation. This involves modelling users through a classifier that takes profile features as input. By evaluating the most relevant features of the classifier and understanding how each feature helps distinguish each cluster, we address **RQ3**. At this stage, the final step in the user profiling methodology is to extrapolate insights that allow us to understand user behaviour based on the work done. This enables us to answer **RQ4**, which in turn addresses specific questions posed by our collaborating entities, that is ONDOD and OBERAXE.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

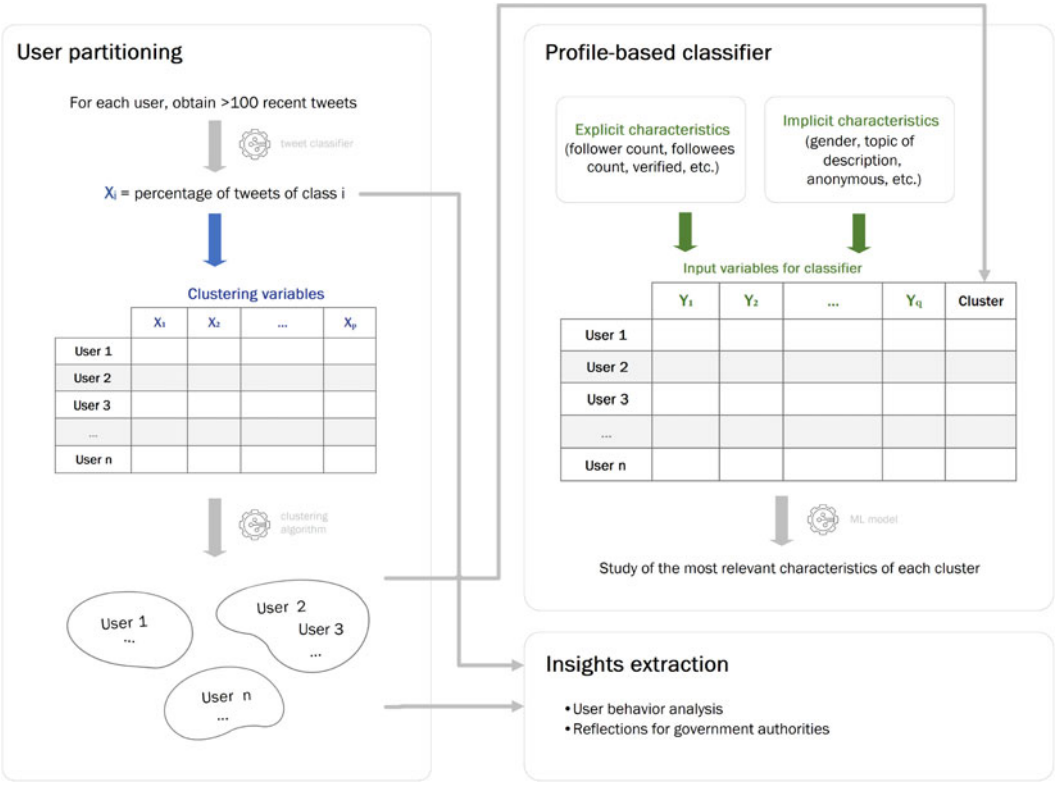


Figure 1. User profiling methodology.

Figure 1 shows the design of the proposed general procedure for profiling users. The following section details the functioning of each module and specifies the variables and algorithms considered for the hate domain on Twitter, explaining why these are chosen over others.

3.1.1 User segmentation

This section discusses the input variables and clustering algorithms that are most suitable for the specific context of hate speech on Twitter, selecting those that best fit this particular scenario.

Clustering analysis. Cluster analysis encompasses techniques aimed at dividing a dataset into different groups so that the elements within the same group are more similar to each other than to those in other groups (Everitt *et al.* 2011). It is based on the hypothesis that the data form a heterogeneous set composed of homogeneous subsets or clusters. However, the number of clusters and the specific characteristics of the data to be considered for classification are not known a priori. This is why cluster analysis is an unsupervised learning tool and, since it makes no assumptions about the data or its structure, it helps address **RQ1**.

For clustering algorithms, there are several options for numerical data, such as k-means, DBSCAN, Fuzzy, or hierarchical methods (Lee and Yun 2003; Schubert *et al.* 2017). K-means is chosen because it performs well with large datasets and has been widely used in a range of applications. In comparison, hierarchical methods are not suitable for large datasets as they rely on visualising a partition tree (Everitt *et al.* 2011). DBSCAN does not assign clusters to points distant from high-density regions, which does not meet our requirement of classifying all users, and Fuzzy, on the other hand, assumes that each point can belong to more than one cluster, which also does not align with our objective. However, since k-means is more effective with multivariate

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

Table 6. Categories designed for tweet classification

Hate speech categories	
offensive hate	Speech that represents personal or collective offences; incites discrimination by reproducing stereotypes or falsehoods.
extreme hate	Speech that advocates, incites, or proposes violent and hostile actions against different groups in a stereotypical and aggressive manner.
neutral	Descriptive speech where no hate appears, but also does not take a stance against it.
upstander	Counter-narrative speech that contributes to breaking stereotypes or taking a stance in defence of the victims. Also known as *counter speech*.

data following a spherical distribution (Chaudhary, 2020), the distribution of the variables must be studied, and transformations may be applied to achieve a spherical distribution before using k-means.

In k-means, the number of clusters k must be chosen in advance. This means that, given a user partition, it is initially unclear whether it reflects an inherent grouping structure in the data or is merely an artificial division. Although the most interpretable partition could be selected, to ensure the objectivity required by **RQ1**, we propose identifying the optimal number of clusters by counting the number of local maxima in the density estimation level curves of the transformed spherical distributed data, as each local maximum identifies a different sub-population in the data.

Input data for the clustering algorithm. As in other user profiling studies in the hate speech domain (Pereira-Kohatsu *et al.* 2019; Rangel *et al.* 2021; Noorian *et al.* 2024; Mishra *et al.* 2019; An *et al.* 2021; Chandra *et al.* 2021; He *et al.* 2022), more than 100 tweets should be collected per user, starting with the most recent ones. From these tweets, certain linguistic traits that are assumed to be related to the study domain can be extracted. For example, in the context of hate speech, it might be hypothesised that a hater uses a higher number of offensive words or makes more spelling mistakes. However, to avoid relying on unfounded assumptions about users' writing styles, we opt for values directly related to the domain of study.

The methodology involves classifying tweets to determine how they relate to hate speech. Once classified, the relative frequency of each class (i.e., the proportion of tweets in each category) is calculated. These frequencies are used as input for a clustering process (previewed in Figure 1). The decision to use relative rather than absolute frequencies is made to account for variations in how many tweets different users post.

In this context, a pre-trained classifier, HaterBERT, is applied to categorise tweets in Spanish into four classes (see Table 6). The relevant inputs for clustering are the relative frequencies of tweets in these categories. If the proportion of offensive or extreme hate is very small, they may be combined into a single "hate" category to avoid data distortion.

So, the input[f] for clustering becomes either:

$$\text{Input Data for Clustering} = X_{\text{upstander}} \times X_{\text{offensive hate}} \times X_{\text{extreme hate}}$$

or, if the categories are combined:

$$\text{Input Data for Clustering} = X_{\text{upstander}} \times X_{\text{hate}}, \text{ where } X_{\text{hate}} = X_{\text{offensive hate}} + X_{\text{extreme hate}}.$$

[f]Note that the frequency of neutral tweets is not included as the sum of all the categories' frequencies is one.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

3.1.2 User classifier based on profile

As shown in Figure 1, the next step is to derive a set of user characteristics from their profile information. The objective is to use these characteristics as input to a classifier that assigns each user a label based on easily obtainable data, without the need to download hundreds of tweets. This approach is innovative, as previous studies have always required tweet characteristics to be included, profile data being only considered as supplementary information (see Table 3). The efficacy of this classifier will help determine whether a user&'s profile alone contains sufficient information to classify them, something not previously explored. If effective, then user profiling can be achieved without downloading large amounts of tweets.[g]

Profile feature extraction. To begin, a set of profile features must be selected. To maximise the amount of information extracted, we consider the full set of features available from Twitter&'s API. Some of these are used directly, such as the number of tweets or followers, or almost directly, via simple calculations. For instance, from the account creation date (provided by the API), the time on Twitter is calculated as the difference in days between that date and the day the data is extracted. These features are labelled as "explicit." Other features, such as the username or description, cannot be used directly and are processed through additional modules to derive "implicit" features. The only data not used or processed is the profile image, which is left for future work. A detailed list of the profile features is provided in Appendix A, where each feature is characterised according to its general type: metadata-based (i.e., extracted directly from the user profile), model-based (i.e., generated through pretrained classifiers), and manual (i.e., derived from structured fields).

The explicit features include the following thirteen attributes. Some may be relevant for identifying haters. For example, Y_{time}, since haters may be more likely to be reported and have their accounts closed, leading to shorter Twitter lifespans, or $Y_{\text{frequency}}$, as haters might post more frequently to reiterate their views.

- Number of tweets ($Y_{\text{tweets}} \in \mathbb{Z}_{\geq 0}$)
- Number of followers ($Y_{\text{followers}} \in \mathbb{Z}_{\geq 0}$), number of followings ($Y_{\text{followings}} \in \mathbb{Z}_{\geq 0}$), and the ratio between the two ($Y_{\text{ratio}} \in \mathbb{R}_{\geq 0}$)
- Time on Twitter in days ($Y_{\text{time}} \in \mathbb{Z}_{\geq 0}$) and tweet frequency ($Y_{\text{frequency}} \in \mathbb{R}_{\geq 0}$)
- Number of lists they are part of ($Y_{\text{lists}} \in \mathbb{Z}_{\geq 0}$)
- Whether they are verified ($Y_{\text{verified}} \in \mathbb{Z}_2$) and type of verification ($Y_{\text{type verification}} \in \mathbb{Z}_4$)
- Number of cashtags, hashtags, mentions, and URLs in the description (Y_{cashtags}, Y_{hashtags}, Y_{mentions}, $Y_{\text{urls}} \in \mathbb{Z}_{\geq 0}$).

The seven selected implicit features are described below. They are derived from data that cannot be used directly (username, description, URLs, and location) due to their potential to distinguish haters from non-haters. In other applications, a different set of implicit attributes should be considered depending on the target objective.

- **Anonymity Level** ($Y_{\text{anonymity}}$) $\in \mathbb{Z}_4$ This feature addresses the question: Are hater users more likely to hide behind anonymity than other types? Anonymity may allow users to express themselves freely without fear of repercussions. To categorise users based on anonymity levels, we use the definitions from Peddinti *et al.* (2014).

[g]This is particularly relevant nowadays, given that currently access to Twitter&'s API is no longer free, requiring a monthly fee for downloading tweets (Calma, 2023).

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

  - Identifiable: A Twitter account containing both a first and last name.
  - Highly Identifiable: An identifiable Twitter account that also contains a URL reference to another social network account that uses a real name (e.g., Facebook or LinkedIn). Also, verified Twitter users are automatically categorised as Highly Identifiable, as Twitter&'s verification process is more rigorous than simply having a name and a URL with a real name (Twitter, 2023). This is a subset of the identifiable group.
  - Partially Anonymous: A Twitter account with either a first or last name, but not both.
  - Anonymous: A Twitter account without a first or last name.

  To detect first or last names within a username, a segmentation module and a pre-existing name bank are used.
- **Location** ($Y_{\text{location}} \in \mathbb{Z}_2$) This is a binary label indicating whether the profile location provides enough information to infer the user&'s geographical location. This feature is motivated by the same privacy concerns as anonymity: users who wish to remain unidentified may also avoid revealing their location. Manual labelling is used, as no existing tool can automatically classify locations. A location is marked as "yes" if it minimally reveals the user&'s whereabouts (e.g., broad regions, countries, Twitter&'s default locations like "ÜT," postal codes, streets, rivers). Blank or non-informative locations are marked as "no." Additionally, verified business or government accounts (not tied to individual users) are marked as "yes."
- **Personality** ($Y_{\text{personality}} \in \mathbb{Z}_5$) One of the most researched and accepted personality theories is the Big Five model, consisting of extraversion, openness to new experiences, conscientiousness, agreeableness, and neuroticism (McCrae and Costa 1987). Since each of these traits leaves a trace on social networks (Azucar, Marengo, and Settanni 2018), it is worth associating each user with a trait from this model. A personality classifier processes the profile description.
- **Topic of Description** ($Y_{\text{topic}} \in \mathbb{Z}_{17}$) The profile description is classified according to its topic.
- **Gender** ($Y_{\text{gender}} \in \mathbb{Z}_2$) The user&'s custom or display name is analysed for male or female names based on a name database. If no gender can be assigned, no label is given.
- **Hate Type in Description** ($Y_{\text{hate}} \in \mathbb{Z}_4$) The description is passed through a pre-trained hate speech classifier.
- **Sentiment of Description** ($Y_{\text{sentiment}} \in \mathbb{Z}_3$) A pre-trained model is used to determine the sentiment (positive, negative, or neutral).

The methodology for implicit feature extraction is summarised in Figure 2.[h]

In this specific case, to segment the words in each username, we use the Python library `ekphrasis.classes.segmenter` (Baziotis, 2022), chosen after testing other options. The name database used is from the INE (the Spanish National Institute of Statistics) (Boquet, 2012), selected because it includes all names and surnames of individuals with Spanish nationality. Additionally, the personality traits in the description are extracted using the multi-language model Siddique *et al.* (2019) On the other hand, the topic of the description is inferred through a model from CardiffNLP (Camacho-Collados *et al.* 2022), which was trained specifically on tweets. The hate speech classifier is HaterBERT for continuity. Finally, the sentiment classifier chosen is *pysentimiento/robertuito-sentiment-analysis*, as it is specific to Spanish and the most accurate among the tested options.

[h]We note that inferred attributes such as personality, gender, or sentiment are proxies derived from automated classifiers. As such, they may introduce noise or bias and should not be interpreted as ground-truth user characteristics but as approximate signals for modelling purposes. This limitation is inherent to large-scale social media analysis, where direct access to validated user attributes is not available.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

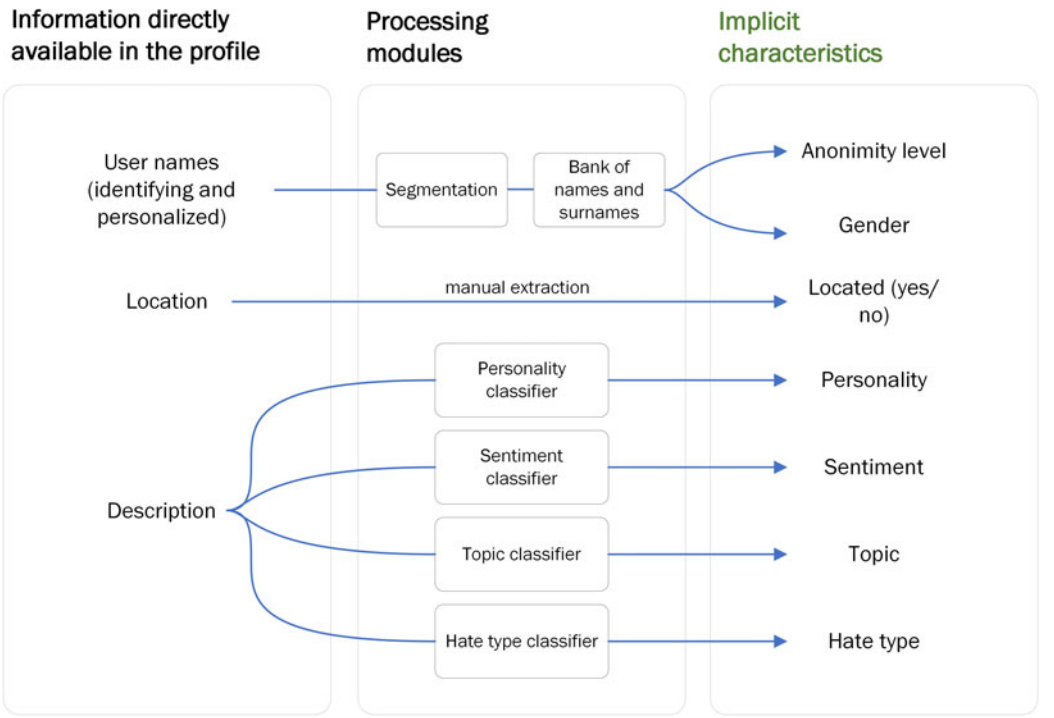


Figure 2. Simplified process for implicit feature extraction.

By integrating these pre-trained transformer-based models (e.g. HaterBERT, CardiffNLP, and *pysentimiento*), our pipeline leverages task-specific models fine-tuned for hate speech detection, topic classification, and sentiment analysis. This hybrid approach allows the final user classifier to benefit from their contextual representations while maintaining the efficiency and interpretability of structured classification on profile-based attributes.

Before processing the description text and tweets through all the models, preprocessing is performed: line breaks, URLs, mentions, hashtags, cashtags, and emojis are removed, and the text is converted to lowercase.

In conclusion, the input features for the classifier are:

$$\text{Input data for classifier} = \text{Explicit features} \times \text{Implicit features} \tag{1}$$

where

$$\begin{aligned}\text{Explicit features} = {} & Y_{\text{tweets}} \times Y_{\text{followers}} \times Y_{\text{followings}} \times Y_{\text{ratio}} \\ & \times Y_{\text{time}} \times Y_{\text{frequency}} \times Y_{\text{lists}} \times Y_{\text{verified}} \times Y_{\text{type verification}} \\ & \times Y_{\text{cashtags}} \times Y_{\text{hashtags}} \times Y_{\text{mentions}} \times Y_{\text{urls}}\end{aligned}$$

and

$$\begin{aligned}\text{Implicit features} = {} & Y_{\text{anonymity}} \times Y_{\text{location}} \times Y_{\text{personality}} \times Y_{\text{topic}} \\ & \times Y_{\text{gender}} \times Y_{\text{hate}} \times Y_{\text{sentiment}}\end{aligned}$$

Classifier creation. After adding these features to the data matrix, a suitable machine learning model must be selected, prioritising interpretability alongside performance. Several machine learning models have been tested in a series of experiments (detailed in Appendix C), and

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

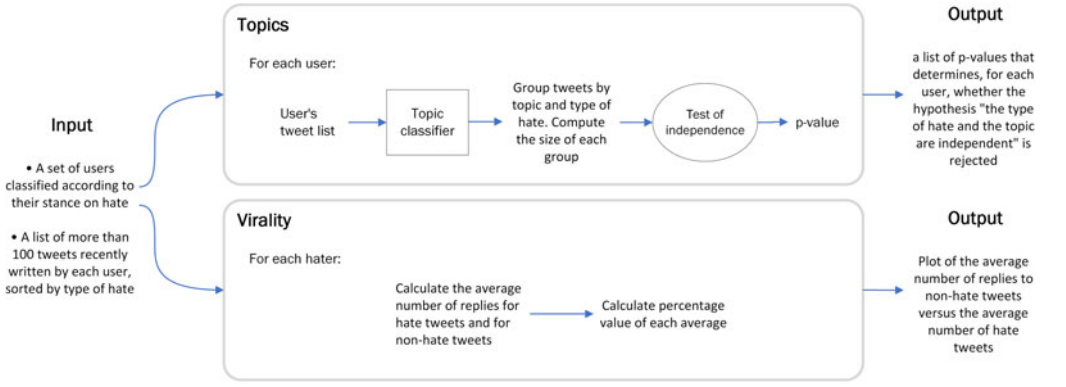


Figure 3. Insight extraction for the specific domain of hate speech.

XGBoost emerged as the top-performing model. In the case of class imbalance, a data augmentation technique is recommended. For this study, we employ random oversampling.

After training the classifier, to address **RQ3**, feature importance is analysed using the F-score, a metric that depends on the number of times each feature appears in the decision tree to split two paths and at what height it is used; thus, the higher the F-score of a feature, the more important it is for classification.

3.1.3 Extrapolation of insights

At this stage, the final point in the user profiling methodology is to extrapolate *insights* that allow us to understand user behaviour based on the classified tweets and the users grouped by clusters. As such, **RQ4** is divided into two parts, addressing concerns raised by the collaborating entities. Below, we describe the statistical analysis model designed to answer these sub-questions.

RQ4a: Behavioural analysis based on thematic differences of each tweet type. We aim to analyse user behaviour by investigating the topics they express hate about. The underlying hypothesis is that users discuss different topics when expressing hate compared to neutral or upstander messages. For example, a user might discuss non-hate-related topics like music but may shift to offensive language when commenting on politics or sports. This leads to the broader question of whether the topic of a tweet (e.g., music, news, and sports) is associated with the tweet&'s stance on hate speech (hate, neutral, or upstander).

Formally, we seek to determine whether the distribution of the random variable "topic" differs across the three tweet categories: hate, neutral, and upstander. To address this, we apply a homogeneity test, using a significance level of 0.05, to assess whether the topics discussed vary depending on the tweet&'s relationship to hate speech.

RQ4b: Behavioural analysis with respect to the virality of hate messages. Second, we want to analyse the virality of hate messages. The number of replies is used as a natural measure of a tweet&'s virality. First, we identify users classified as haters and calculate the average number of replies for hate tweets (offensive and extreme) and non-hate tweets (neutral and upstander), grouping them by author. Since we are not interested in the absolute values of the averages but rather how they compare, these values are normalised as relative proportions. Next, the average number of replies to non-hate tweets is plotted against the average number of replies to hate tweets, with users ranked by the latter.

Figure 3 outlines the steps described above to address the sub-questions.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

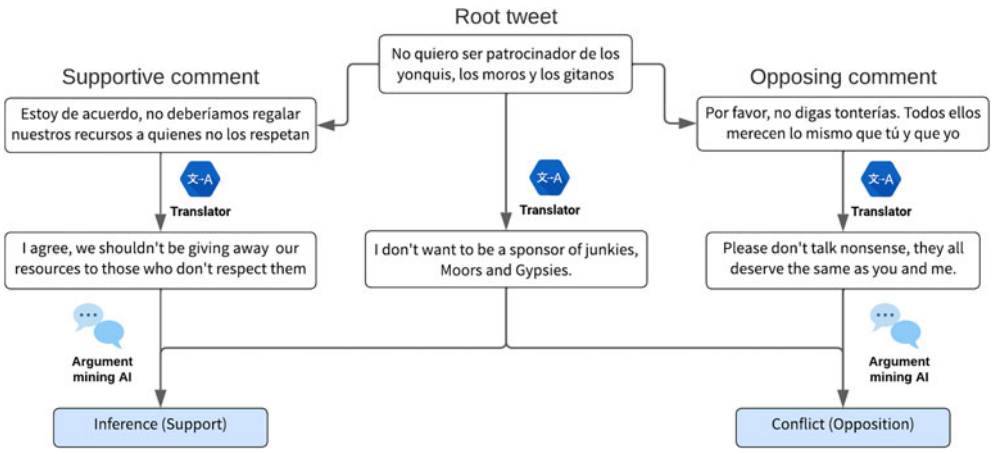


Figure 4. Example of the process for obtaining the argumentative relation between two replies and the original message they respond to. All the texts are fictitious.

3.2 Interaction graph

This project&'s final contribution focuses on designing a tool that enables authorities to visualise the nature of interactions between different types of users. Therefore, to address **RQ5**, we propose the design of an interactive graph. This visualisation tool is intended to be run on a dataset that includes complete conversations. According to Twitter&'s definition, a conversation consists of all the replies generated from a tweet, including direct replies, replies to those replies, and so on. Each tweet must be accompanied by the user who posted it and the necessary profile information to classify the tweet&'s author. In this case, the same dataset used for the profiling methodology is utilised (see Section 4).

In the graph, edges are directional, originating from the users who write a tweet and pointing to the users who respond to that tweet. The colour of the edges represents the nature of the interaction between the two nodes it connects: red for opposition, green for indifference, or grey for support. To obtain the edges, each tweet-response pair is processed with an argumentative relation classifier. Specifically, we use the model `raruidol/ArgumentMining-EN-ARI-AIF-RoBERTa_Lv` from the HuggingFace platform, as it is the only public algorithm of its kind (to the best of our knowledge). It has four labels: "Conflict," "Rephrase," "Inference," and "No-relation." In this work, "Rephrase" and "Inference" are treated as support. As the model only handles texts in English, following the results by Toledo-Ronen *et al.* (2020), all tweets are translated using the `Helsinki-NLP` translator. Figure 4 provides an example of how these tools are used.

Nodes represent users while their colour denotes the user type. The specific colours are chosen based on the user classification resulting from the clustering method. To obtain the nodes, all users involved in the conversations are selected. They are then classified according to the partition established in the User Profiling section (Section 3.1). For this, we use the user classifier developed in Section 8, which avoids the need to download large amounts of Twitter data. The complete process for obtaining nodes and edges is illustrated in Figure 5.

Finally, the graph is visualised using Visone,[i] which not only offers the ability to visualise nodes and edges with custom colours but also allows for interaction with the graph, including filtering functions and the visualisation of profile characteristics for each node.[j]

[i] https://www.visone.ethz.ch/

[j] Graphext (https://www.graphext.com/) and Gephi (https://gephi.org/) were considered for network analysis but lacked customisation for interactions and user types. `networkx` (https://networkx.org/) was also tested but produced non-interactive graphs with unclear layouts.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

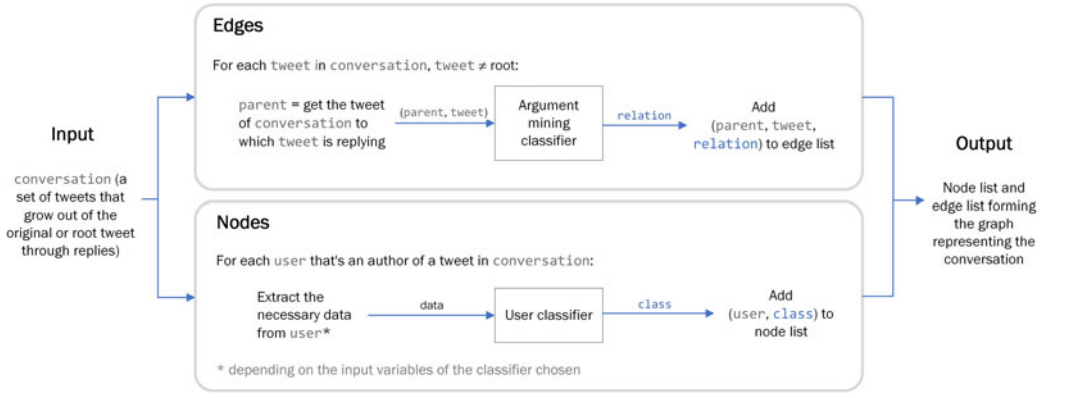


Figure 5. Algorithm for generating each graph.

4. Data extraction

This section describes the data extraction process used to conduct all the experiments.

As mentioned in Section 3.1.1, to answer **RQ1**–**RQ4**, it is necessary to have a set of users with at least 100 tweets each. To answer **RQ5** and build the user graph, it is essential to access complete conversations, which implies obtaining all the tweets from the comment section. To the best of our knowledge, there is no public dataset of Spanish tweets that meets either of these requirements. Therefore, new data must be extracted from the network. The research questions are addressed through the analysis of a static dataset comprising tweets posted between July and September 2023.

4.1 Conversation extraction

The API operates through keyword queries, returning all tweets (and their metadata) that are less than one week old. Since it is not possible to search by coordinates, we filter by language after collecting the tweets to select those from Spain. Regarding the keywords used, we have utilised a set of hate-related terms developed during the ALRECO project. This set consists of seven files representing six categories: Roma people, Jewish people, migrants, Muslims, cultural minorities, and religious ethnicities, as well as other general terms. The full list of words in this dataset is included in Appendix D.

In total, 1.6 million tweets were retained after filtering from the initial 3.3 million tweets collected with hate-associated keywords. This subset constitutes the keyword-based corpus used for conversation-level analysis. Table 7 shows the number of tweets extracted per category and how many have responses that could be downloaded. It is worth noting that the proportion of tweets with responses ranges from 40% to 46% across all categories, except for the Jewish community (28%).[k] Additionally, for each tweet with responses that served as root (i.e., not replies or retweets), their associated conversations have been extracted, adding up to another 1.6 million tweets. On average, each conversation contains 7.87 tweets, with a minimum of 2 and a maximum of 8,924 tweets.

Once the corpus of randomly downloaded tweets that may contain hate is obtained, it is necessary to select a representative sample of users whose behaviour will be studied. Ideally, this sample should include users who potentially emit a significant amount of hate, as well as those who frequently oppose hate and others who remain indifferent. This ensures that the sample is diverse

[k]Many tweets in this subset use the term "Nazi" as an insult rather than in relation to hate speech about the Jewish community, which may explain the differing proportions compared to other categories.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

Table 7. Number of tweets extracted by category and how many of these have received responses

Category	Number of tweets extracted	Number of tweets with responses
Global	396,449	157,810 (40%)
Cultural minorities and religious ethnicities	410,774	187,914 (46%)
Roma people	16,543	6,649 (40%)
Jewish people	114,496	32,367 (28%)
Migrants	616,976	269,941 (44%)
Muslim people	116,895	47,171 (40%)

Table 8. Number of tweets extracted, grouped by keyword set, and classified by hate type

	Number of tweets per category			
	Extreme hate	Offensive hate	Upstander	Neutral
Global	4 (0.001%)	120,630 (30.4%)	11,336 (2.9%)	264,479 (66.7%)
Cultural minorities and religious ethnicities	6 (0.001%)	158,717 (38.6%)	17,608 (4.2%)	234,443 (57.1%)
Roma people	1 (0.006%)	8,876 (53.7%)	1,891 (11.4%)	5,775 (34.9%)
Jewish people	11 (0.01%)	61,536 (53.7%)	8,785 (7.7%)	44,164 (38.6%)
Migrants	18 (0.003%)	213,012 (34.5%)	35,414 (5.7%)	368,532 (59.7%)
Muslim people	29 (0.02%)	52,419 (44.8%)	5,817 (5%)	58,630 (50.2%)

Table 9. Overview of the two datasets used in this study

Dataset	Source	Size (before filtering)	Size (after filtering)	Purpose
Keyword-based corpus	Hate-associated keywords	3.3M tweets	1.6M tweets	Conversation-level and lexical analyses
User-based corpus	Timelines of ∼3,000 users identified in keyword corpus	–	3.0M tweets	User profiling and interaction analyses

and comprehensive. However, given that only an average of one tweet per user is available, it is assumed, for now, that the authors of hateful tweets are those who frequently engage in hate speech, with similar assumptions for the other categories.

Thus, it becomes necessary to classify the collected tweets. For this purpose, we have used the HaterBERT model (del Valle-Cano *et al.* 2023). Table 8 shows the resulting number of tweets for each class. It is evident that the extreme hate class is by far the minority, as these tweets contain highly offensive and explicitly aggressive content, which are less common compared to milder levels of hate speech.

For clarity, we have worked with two related but distinct datasets (see Table 9): (i) a keyword-based corpus of 3.3 million tweets containing hate-associated terms, from which 1.6 million unique tweets were retained after filtering and (ii) a user-based corpus, built by extracting timelines from approximately 3,000 users linked to hate speech, amounting to over 3 million tweets. Throughout the paper, we explicitly state which dataset each analysis draws upon.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

4.2 User extraction

Since the goal of this study is to understand the behaviour of users involved in hate-related activities, it is desirable to focus on users who have been active on Twitter for a significant period and maintain consistent activity. This ensures a more complete and representative data sample. From the authors of the tweets collected during the extraction process, users with fewer than 100 tweets are excluded, as are those with an average activity of fewer than one tweet per month. The remaining sample of tweets is then grouped by type of hate. Since the number of extremely hateful tweets is relatively small, they are combined with offensive hate tweets. Based on this, a ranking of users is created according to the following criteria:

- Number of tweets authored by the user
- Number of reactions (sum of likes, retweets, replies, and quotes)
- Number of followers
- Number of views
- Ratio of reactions to views

This process generates five different rankings, and the top 200 users from each ranking are selected. To avoid duplicates, users already selected in one ranking are excluded from others. The result is three distinct lists: one with 1,000 users who have posted at least one hate message, another with 1,000 users who have posted an upstander message (supportive), and a third with 1,000 users who have posted at least one neutral message. Out of these 3,000 users, only 2,722 accounts were available for extracting recent tweets (the rest had been deactivated or privatised). On average, 1,103.33 tweets were collected per user and classified using HaterBERT (the results are presented in the next section). This provides the necessary dataset to achieve the objectives of this project.

4.3 Validation of pretrained classifiers for hate speech in Spanish

A potential limitation of our approach is the reliance on pretrained classifiers, such as HaterBERT, for the detection of hate speech. Concerns have been raised in the literature that biases or inaccuracies in such models may propagate into the final labels. To mitigate this risk, we have conducted an additional validation exercise comparing HaterBERT with other available hate speech classifiers in Spanish. We have evaluated the model pysentimiento/robertuito-hate-speech, which constitutes one of the few alternatives for Spanish within the Hugging Face library. Both models have been applied to the same manually validated subset of tweets, and their outputs have been assessed for consistency in terms of label distribution and classification accuracy. Table 10 reports their overall performance on a manually validated subset of tweets. These results demonstrate that differences between HaterBERT and robertuito are minimal. Both models achieve highly comparable F1-scores (macro average above 0.92 in our validation set), and the distribution of predictions across hate/non-hate classes is nearly identical. Importantly, no systematic skew towards particular subgroups or protected attributes (gender, ideology, and age) has been observed. In addition, we have performed a fairness check across protected attributes (gender, ideology, and age) to verify that neither model exhibits systematic disparities. As shown in Table 11, the absolute differences in F1-scores between models remain below 0.02 across all groups. This indicates that the classifiers behave consistently and equitably, without propagating significant bias into the final labels. This consistency across models indicates that our findings are not artefacts of a single pretrained classifier.

Furthermore, in the clustering experiments (see Section 5.1), the user groupings (haters, neutrals, and upstanders) emerge from aggregated proportions of hate content rather than isolated tweet-level classifications. This methodological choice reduces the potential impact of misclassifications by any one model and reinforces the robustness of the results. In sum, while pretrained

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

Table 10. Comparison of Spanish hate speech classifiers on the validation subset. Notes: Values are averaged over the manually validated subset of tweets. Both models show very similar classification performance and nearly identical distributions of predicted labels, with no systematic skew across protected attributes

Model	Macro-F1	Accuracy	% Hate (predicted)	% Non-hate (predicted)
HaterBERT	0.93	0.94	27.4%	72.6%
robertuito-hate-speech (pysentimiento)	0.92	0.93	26.9%	73.1%

Table 11. Fairness check of hate speech classifiers across protected attributes. Notes: The Δ column shows the absolute difference in F1-scores between models. Differences remain below 0.02, suggesting consistent behaviour across groups and no systematic bias propagation. This table demonstrates explicitly that both models behave stably and equitably across gender, ideology, and age

Protected Attribute	Group	HaterBERT F1	robertuito F1	Δ (difference)
Gender	Male	0.92	0.91	0.01
Gender	Female	0.93	0.92	0.01
Ideology	Left-leaning	0.91	0.90	0.01
Ideology	Right-leaning	0.92	0.91	0.01
Age group	$<$40	0.92	0.91	0.01
Age group	$\geq$40	0.93	0.92	0.01

models always entail some degree of bias, our validation shows that, in the Spanish context, the leading available classifiers for hate speech (HaterBERT and robertuito) produce stable and convergent outputs. This strengthens confidence that the conclusions presented here are not overly dependent on the idiosyncrasies of a particular model.

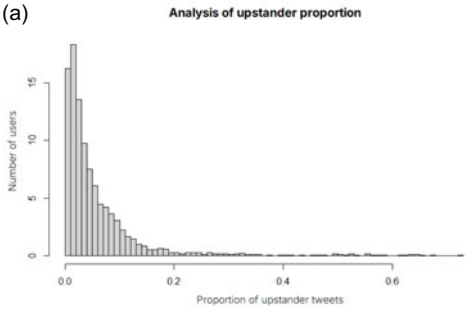
5. Results

This section presents the results of applying our methodology for user profiling and modelling in the context of hate speech on Twitter. The analysis focuses on identifying key user characteristics, behaviours, and patterns associated with the dissemination of hateful content. Through a combination of feature extraction and machine learning techniques, we aim to uncover insights into user behaviour and profile traits that are indicative of hate speech activity on the platform. The findings offer a comprehensive view of how specific user profiles contribute to the propagation of hate, supporting the broader objective of understanding and mitigating harmful online behaviour.

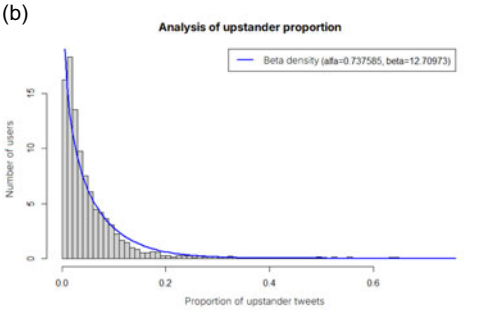
5.1 User segmentation

In this section, **RQ1** and **RQ2** are addressed by applying the designed methodology to the extracted data. First, for each user, the proportion of hate tweets (variable X_{hate}, including both offensive and extreme) and the proportion of upstander tweets (variable $X_{\text{upstander}}$) are extracted. The two types of hate are merged, as the number of extreme hate tweets is considerably small (see Table 8), which could cause distortions when running the *clustering* algorithm. Then, the user density is graphically analysed to identify the optimal number of clusters. The data are transformed to achieve a spherical distribution, which provides clearer results while, at the same time, improving the performance of k-means (Chaudhary, 2020).

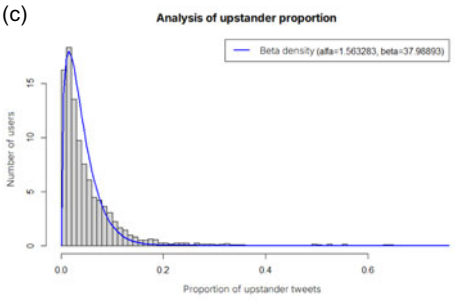
https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press



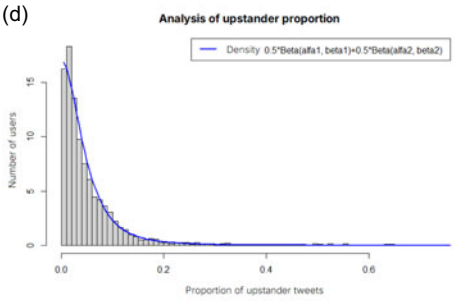
Histogram of the samples of $X_{\text{upstander}}$.



Histogram of the samples of $X_{\text{upstander}}$ accompanied by the density of Beta($\alpha = 0.737585, \beta = 12.70973$), whose parameters are calculated from the mean and median.



Histogram of the samples of $X_{\text{upstander}}$ accompanied by the density of Beta($\alpha = 1.563283, \beta = 37.98893$), whose parameters are calculated from the mode and the median.



Histogram of the samples of $X_{\text{upstander}}$ accompanied by the density of the result of adding the previous two and weighting them equally. In the legend, alfa1 $= 0.737585$, alfa2 $= 1.563283$, beta1 $= 12.70973$ y beta1 $= 37.98893$.

Figure 6. Density analysis of $X_{\text{upstander}}$.

Let $X_{\text{upstander}}$ and X_{hater} represent the proportion of upstander and hate tweets for each user, respectively. The distribution of the vector $(X_{\text{hater}}, X_{\text{upstander}})$ is studied to assess the possibility of transforming it into a spherical distribution. First, the distribution of $X_{\text{upstander}}$ and X_{hater} is studied separately. Figure 6a shows the histogram of the 2,722 samples of $X_{\text{upstander}}$. It is clear that it does not follow a spherical distribution. However, it is possible that it follows a Beta(α, β) distribution, as it takes values in $[0, 1]$. To verify this assumption, the parameters α and β are estimated. Knowing that:

$$\text{E}[X] = \frac{\alpha}{\alpha + \beta}, \ \text{Median}[X] \approx \frac{\alpha - 1/3}{\alpha + \beta - 2/3}, \ \text{Mode}[X] = \frac{\alpha - 1}{\alpha + \beta - 2} \tag{2}$$

for $X\text{Beta}(\alpha, \beta)$, the parameters α and β are estimated using the sample values of the mean, median, and mode. The results are shown in Figures 6b and 6d. It can be observed that in both Figure 6b and Figure 6d, the density function fits the histogram reasonably well. However, in Figure 6b, the density converges to zero more rapidly than the histogram. The opposite occurs in Figure 6d. Therefore, it can be inferred that $X_{\text{upstander}}$ is, in fact, the sum of two beta distributions. In Figure 6c, the density as a sum of the previous betas (equally weighted) now fits much better with the histogram, both for values near zero and those farther from zero. In conclusion, it is reasonable to assume that $X_{\text{upstander}}$ indeed follows the distribution of the sum of two betas.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

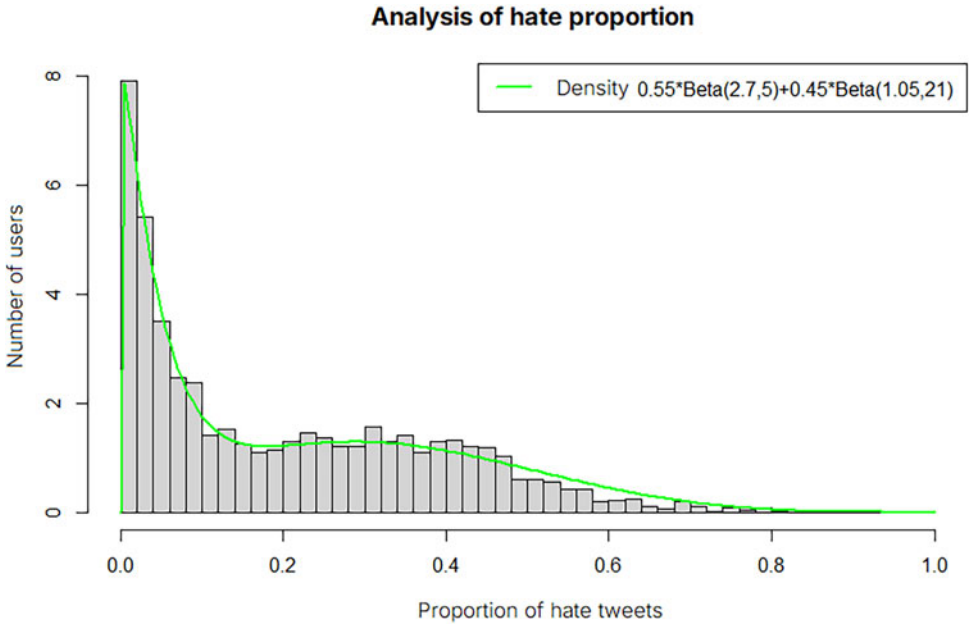


Figure 7. Histogram of the proportions of hater tweets accompanied by the density function $0.55 * \text{Beta}(\alpha = 2.7, \beta = 5) + 0.45 * \text{Beta}(\alpha = 1.05, \beta = 21)$.

Analogously, the distribution of the hate proportions, X_{hate}, is analysed, and it is concluded that it is reasonable to assume that it follows a sum of two beta distributions (Figure 7).

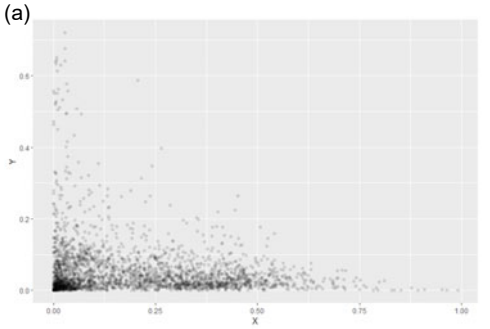
Figure 8a represents each user as a point, where the x axis corresponding to X_{hate} and the y axis to $X_{\text{upstander}}$. The farther a point is from the x-axis, the more upstander messages the user writes, and the farther it is from the y-axis, the more hate the user emits. It is evident that no point exceeds the diagonal $x + y = 1$, as these are proportions that cannot sum to more than one. The points are graphed with partial transparency, allowing us to see areas with a higher concentration of users. It is noticeable that near the bottom-left corner, where the hate and upstander levels are close to zero, many points accumulate. These are the users who post few tweets related to hate speech and primarily write neutral tweets. Also, it can be easily seen that the points do not follow a spherical distribution, which was anticipated given that hate and upstander proportions are not spherical. The contour lines of the estimated density in Figure 8b are plotted. Indeed, we observe that the contours are not circular in shape; for comparison, Figure 8c shows the contour lines of a spherical density, where circular contours are clearly visible.

Moreover, the potential number of clusters can already be identified. In fact, Figure 8a resembles a topographic map, where the contour lines represent density in each region, that is, the concentration of users. It is very clear that near the point (0,0), there is a local maximum, that is, a closed curve that contains no other curves within it, which means that many users are concentrated in the region with low hate and upstander proportions. This had already been observed in Figure 8a. However, now two other local maxima are detected: one directly above the one previously mentioned and another elongated one to the right. This indicates that the data can be classified into three clusters, as each local maximum identifies a sub-population within the total user population.

To obtain a spherical distribution, the following transformation is chosen:

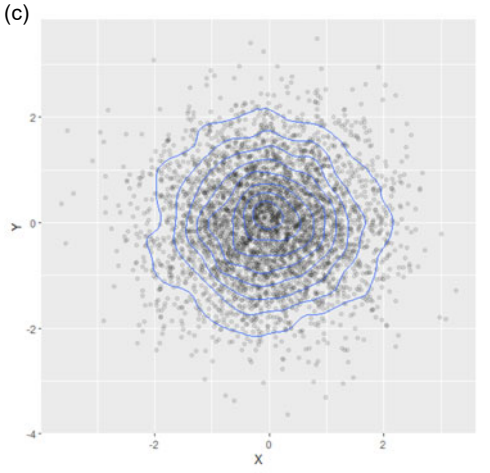
$$x \mapsto \frac{x}{1-x} \mapsto \log\left(\frac{x}{1-x}\right) \tag{3}$$

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press



Scatter graph of the analysed users contrasting the proportion of hate emitted versus that of upstander type messages.

Density estimation for the scatter graph of the analysed users contrasting the proportion of hate emitted versus that of upstander type messages.



Density estimation for a randomly simulated bivariate normal.

Scatter plot of the proportions of hatred emitted compared to that of upstander type messages once the described transformations have been carried out and accompanied by the density estimate.

Figure 8. Density estimation level curves for different distributions.

The first transformation maps the interval $(0, 1)$ to all of $\mathbb{R}^+$, and the second maps it to all of $\mathbb{R}$, which is the space where a spherical variable takes values. In Figure 8d, it can be seen that after applying the transformation, the three previously identified local maxima become more visible, resembling those in Figure 8c, and therefore more similar to a normal distribution.

The resulting k-means clustering with $k = 3$ is illustrated in Figure 9, where users are divided into those with little relation to hate speech (in green), those who emit more hate (in red), and those who post more upstander content (in blue). In Appendix B, we show that, indeed, the transformed data are better suited for applying k-means clustering. To evaluate the quality of the clustering, we have computed the silhouette coefficient (0.47) and the Davies–Bouldin index (0.61). Both values indicate meaningful separation between clusters, supporting the existence of distinct user groups (haters, upstanders, and neutrals) beyond the traditional binary classification.

At this point, **RQ1** and **RQ2** are definitively answered. The answer to **RQ1** is that cluster analysis combined with density analysis is an objective methodology that can be used to classify users based on their stance on hate speech, surpassing the arbitrary criteria found in the literature. The answer to **RQ2** is that there are three underlying groups within the user set: those with little

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

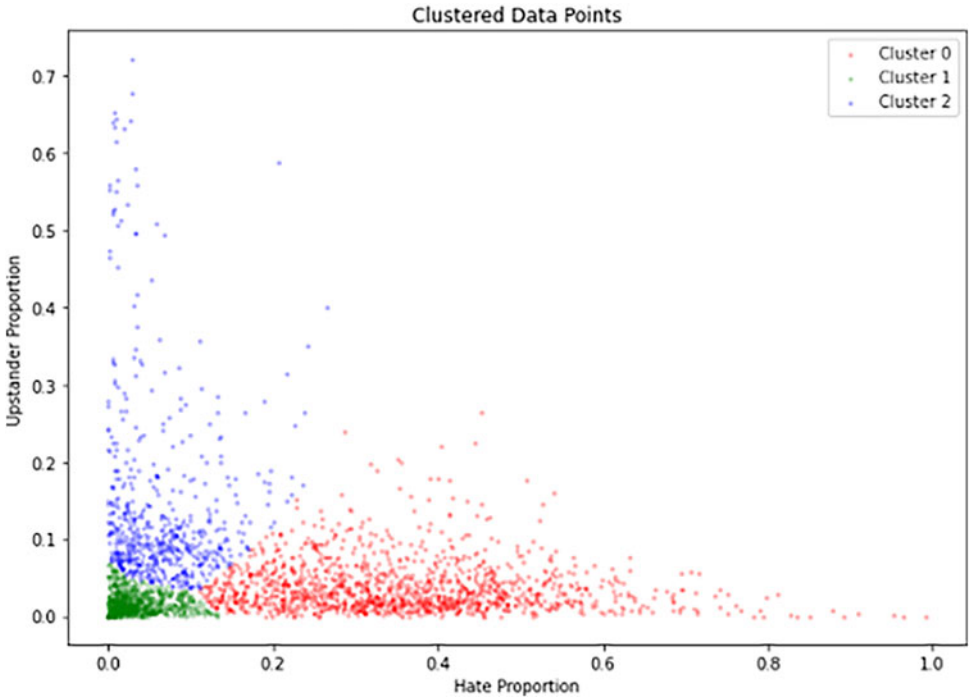


Figure 9. Clusters identified by k-means on the transformed data, with $k = 3$.

relation to hate speech, those who post a lot of hate, and those who oppose hate. From now on, these groups are referred to as *neutrals, haters, and upstanders*, each comprised of 734, 1,390, and 598 users, respectively.

5.2 User classifier

In this section, **RQ3** is addressed, focusing on which profile characteristics (number of followers, number of tweets, etc.) differentiate each type of user. By constructing a model that classifies each user as a hater, neutral, or upstander, we analyse the attributes that hold the most importance in the model.

For each user, the features defined by Equation (1) are extracted. The classifier is then trained using the XGBoost classifier implemented through the Python package *xgboost*, applying *oversampling* (using the *RandomOverSampler* class from the *imblearn* library) during training to balance the classes, with a test size of 20%. To evaluate the profile-based classifier, we have applied 10-fold cross-validation with hyperparameter tuning via Grid Search. In each iteration, the classifier has been trained on nine folds and evaluated on the held-out fold, with performance aggregated across folds. User labels (Haters, Upstanders, and Neutrals) are derived from the unsupervised clustering procedure in Section 5.1 and serve as ground truth. Among all tested models (i.e., logistic regression, Elastic Net, SVM, CNN, and XGBoost, see Appendix C, Table 15) XGBoost achieves the best performance across all metrics: Accuracy 0.7560, weighted F1 0.7571, macro F1 0.7384, AUROC 0.9086, and AUPRC 0.8284. These values demonstrate a significant relationship between profile features and user type. This finding is important as it shows that it is not necessary to analyse multiple tweets from a user to determine their type, surpassing the need found in the literature for analysing multiple tweets.

Figure 10 reports the resulting confusion matrix, where each cell contains both absolute user counts and row-normalised percentages. It is clear that haters and neutrals are classified most

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

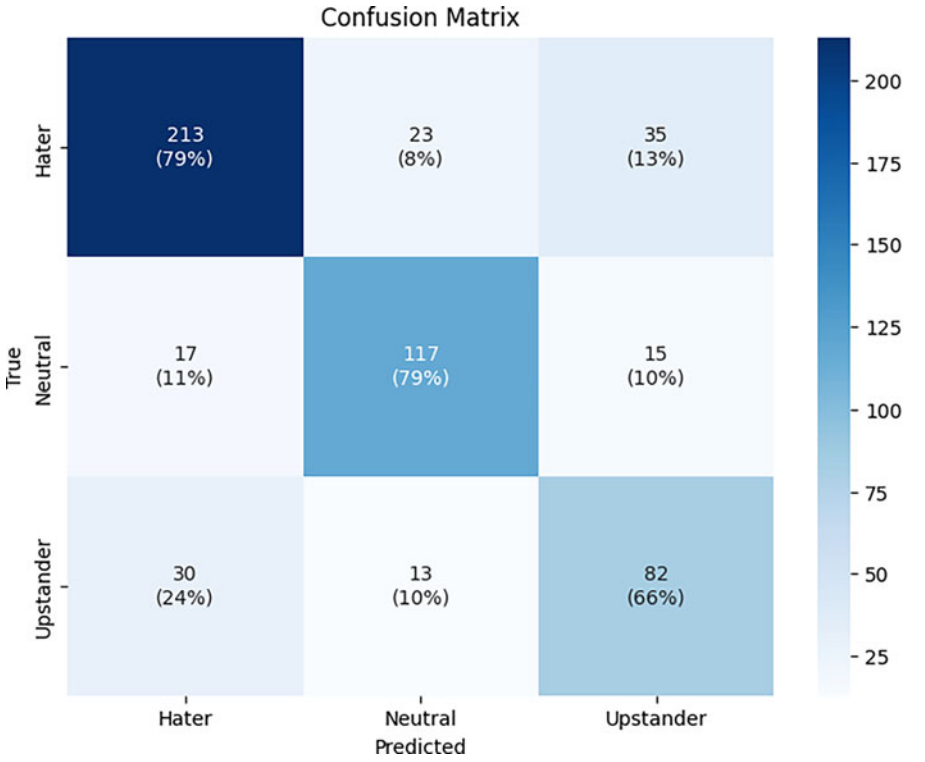


Figure 10. Confusion matrix of the profile-based classifier using 10-fold cross-validation. User labels correspond to the cluster-derived classes (Haters, Upstanders, and Neutrals). Each cell shows both the number of users and the percentage within the true class row.

accurately. The reason upstanders are not perfectly classified could be due to their similarity to haters in terms of profile features, as explained below.

Figure 11 shows the importance level of each attribute, measured using the F-score. The higher the F-score of an attribute, the more important it is. Therefore, the most relevant features for classification, in order, are Y_{time}, Y_{ratio}, $Y_{\text{followers}}$, $Y_{\text{frequency}}$, Y_{tweets}, and Y_{lists}. The distribution of each feature, grouped by user type, is plotted in Figure 12.

At this point, we can answer **RQ3**, which asks which profile attributes differ the most between clusters. Figure 12 shows that neutral users have been on Twitter longer, have more followers, have posted more tweets (with higher frequency), and are included in more lists. In other words, neutral users have a strong presence on the platform, are active, and have built large networks. In contrast, haters and upstanders do not exhibit these characteristics but tend to follow more accounts than neutrals, with upstanders slightly outpacing haters.

Since no attribute clearly distinguishes haters from upstanders, we analyse the next most relevant feature in the model, Y_{topic} (the topic of the description). Figure 13 shows the proportion of topics by user type. Although the two dominant topics are the same in both groups (news and social concerns in light green, daily life in purple), the percentages differ significantly: 35.6% and 27.6% for haters, and 64.8% and 11.7% for upstanders. This means that upstanders tend to mention news and social concerns in their descriptions more than haters.

The remaining attributes have been analysed, and all have been found to be similarly distributed in both groups. The only significant difference observed is in the sentiment of the description: while neutral sentiment predominates in both groups, the second most common sentiment is positive for upstanders and negative for haters (Figure 14).

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

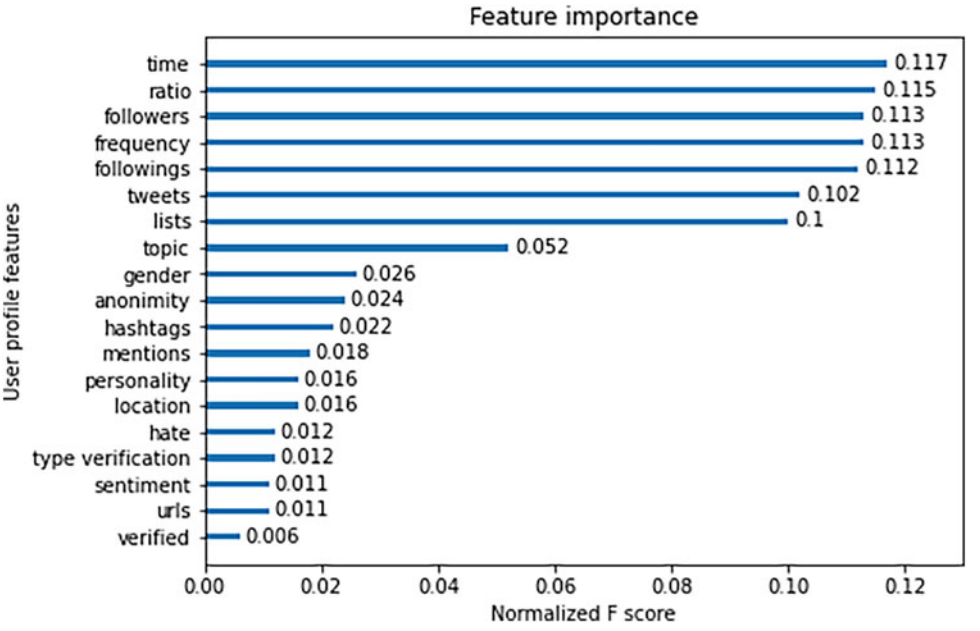


Figure 11. Feature importance of profile attributes identified by the model (ordered by the normalised F-score).

5.3 Extrapolation of insights

The insight extraction methodology allows us to answer the two sub-questions of **RQ4**.

Behavioural analysis based on thematic differences of each tweet type. To address this question, a hypothesis test for the independence of tweet type (hate or non-hate) and tweet topic (music, sports, news, etc.) is run for each user. The hypothesis that tweet type and topic are independent is rejected 56% of the time. In conclusion, more than half of the users change topics when their tweet is related to hate speech.

Behavioural analysis with respect to the virality of hate messages. To address this question, in Figure 15 we plotted for each user the average number of responses to hate tweets (red dots) and the average number of responses to non-hate tweets (light blue dots), normalised to sum one[1] The horizontal thin lines show the average proportions across all the users. We observe that the red dots do not cross 50% until approximately the 1,250th user. This means that, out of the 1,390 haters, only about 140 tend to receive more responses on their hate tweets compared to non-hate tweets. Additionally, the graph reveals that, for the remaining 1,250 users (approx), the percentage of responses to hate tweets is below 0.4 for the majority of them. Therefore, we can conclude that, according to our data, hate tweets are less viral than non-hate tweets.

5.4 Interaction graph tool

The ultimate goal of this project is to analyse the interactions between different types of users to address **RQ5**. Nodes and edges are assigned colours according to the type of user and interaction, respectively, according to the schema below (see Figure 16).

[1]As an example, let us consider a user that receives on average four responses to their hate tweets and two responses to their non-hate tweets. Then, their proportions would be 0.6667 for hate and 0.3333 for non-hate tweets, respectively.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

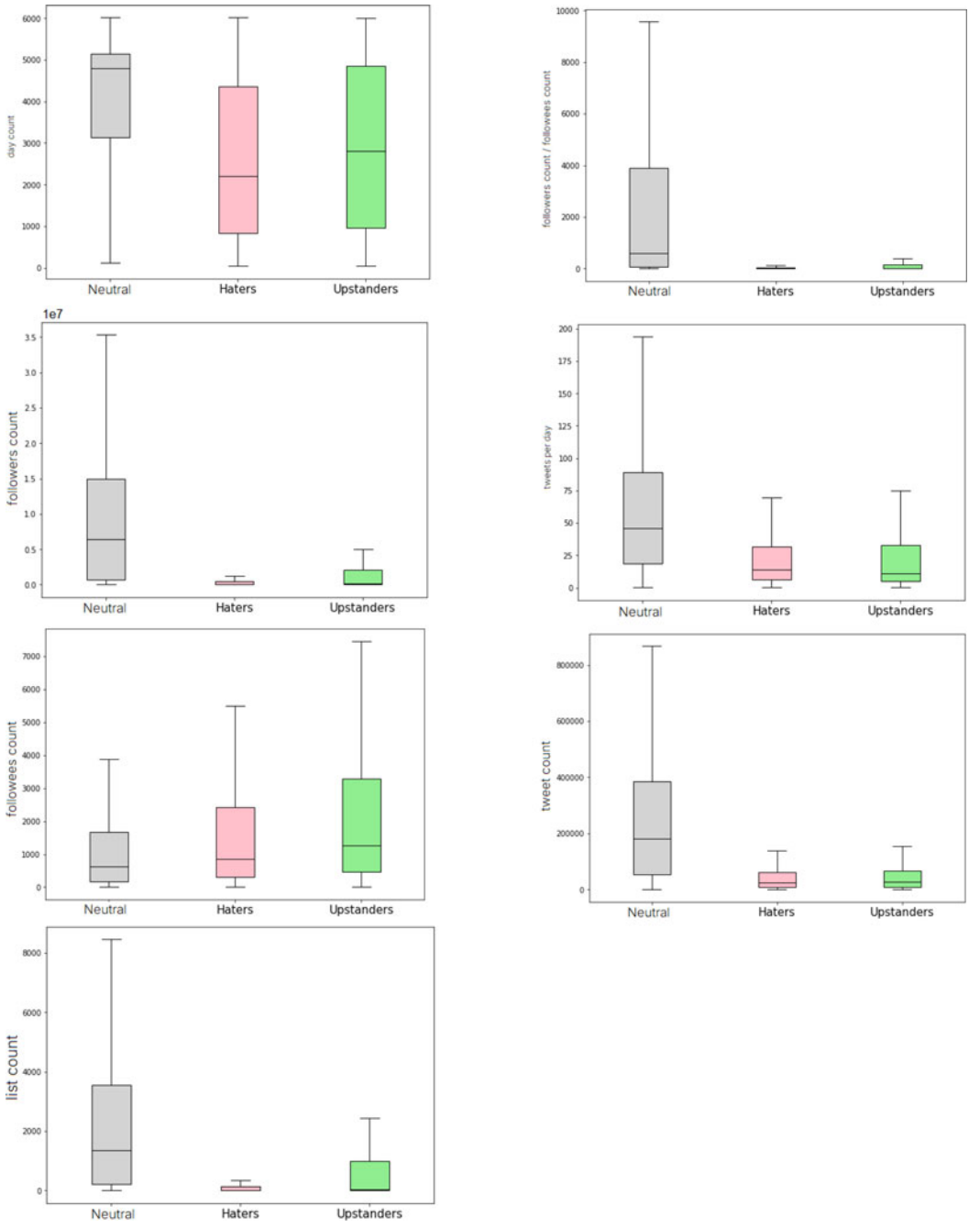


Figure 12. Box plot of the five most important characteristics of the model by author type. Outliers are not graphed for clarity.

Next, the argumentative relation process (Figure 4) is implemented and executed on the set of complete conversations extracted using hate-related keywords. Figure 17 includes two fragments of graphs obtained, along with the conversations from which these graphs were derived. Figure 18 shows a complex graph with multiple edges and nodes.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

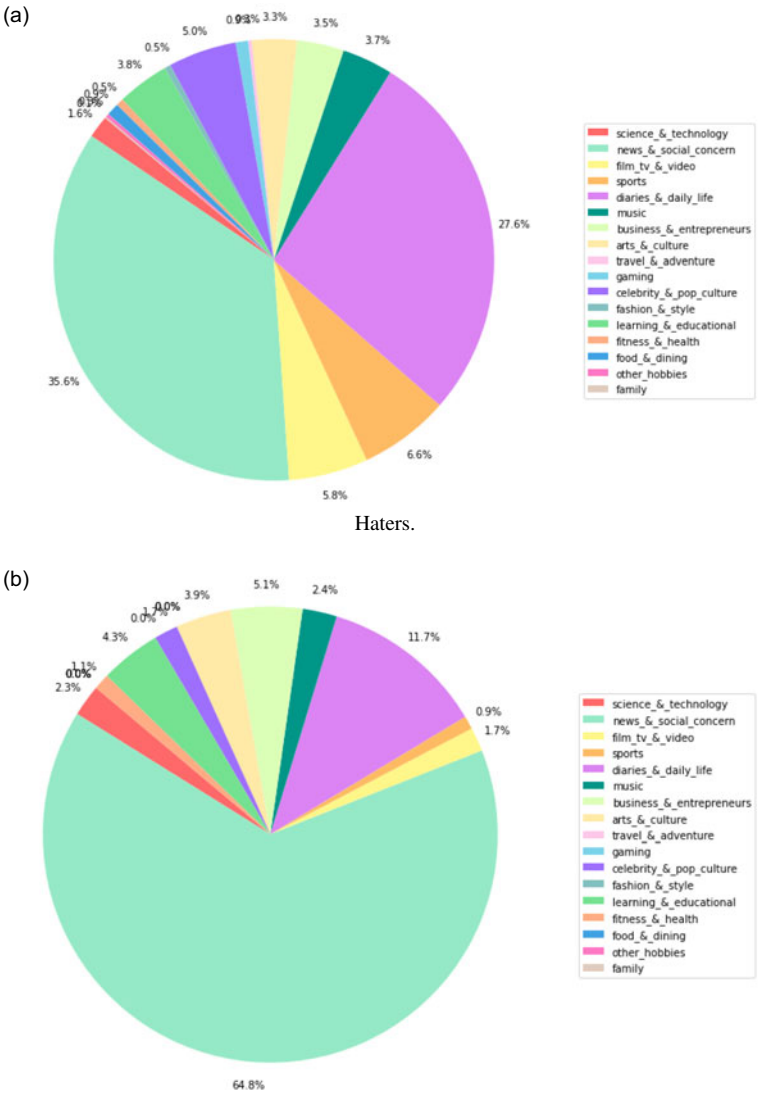
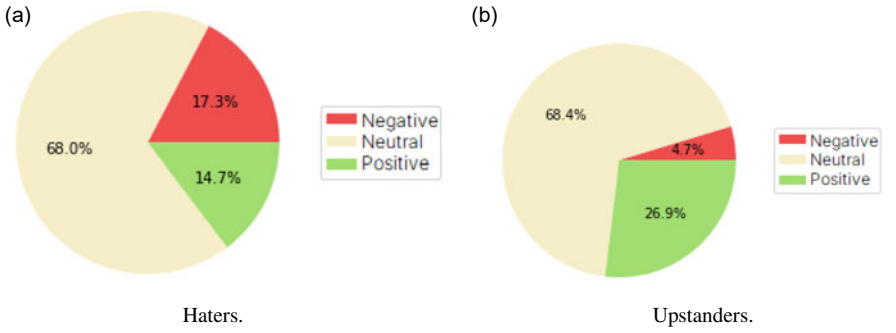


Figure 13. Pie charts of topics' description by user type.

Table 12 shows the percentage of node and edge types obtained by hate category. In general, the majority of edges represent support, and the majority of nodes are haters. This suggests that haters tend to group together and reinforce each other. This behaviour could be beneficial, as clustering makes it easier to identify and apply hate mitigation strategies in a targeted and localised manner.

At this point, the last question, **RQ5**, is answered. Through the design of a graph representing the nature of interactions between users within the context of hate, a series of graphs have been generated that can be used by collaborating entities to analyse and better understand how users interact with each other. Note that the tool includes additional functionalities, such as modifying node sizes based on centrality measures, filtering by node and edge types, and more.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press



Haters. Upstanders.

Figure 14. Pie charts of sentiment&'s description by user type.

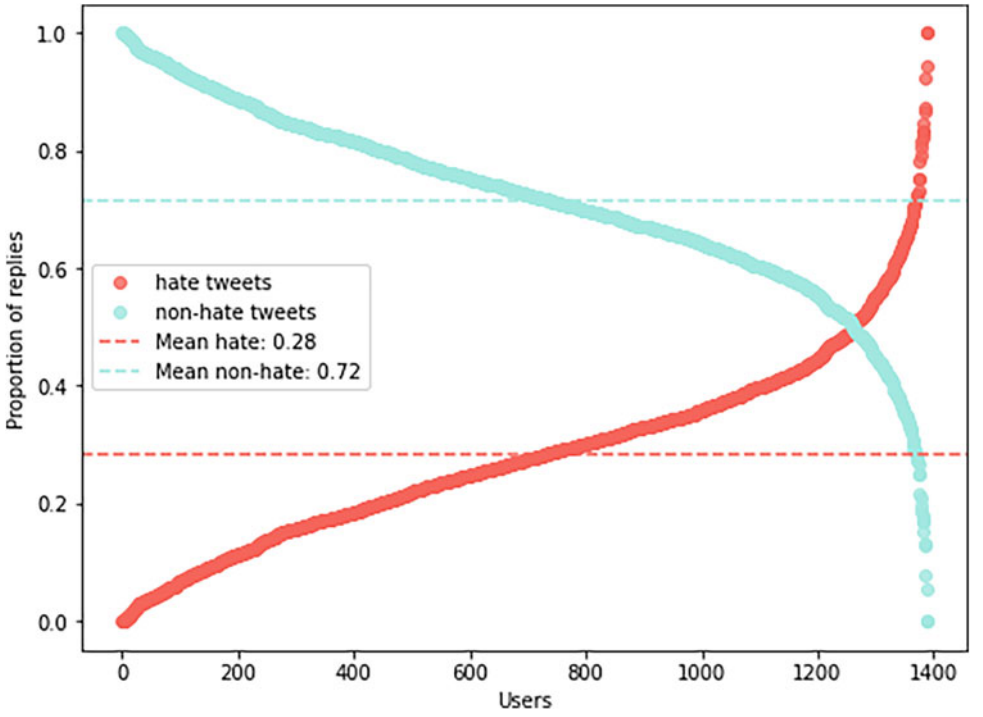
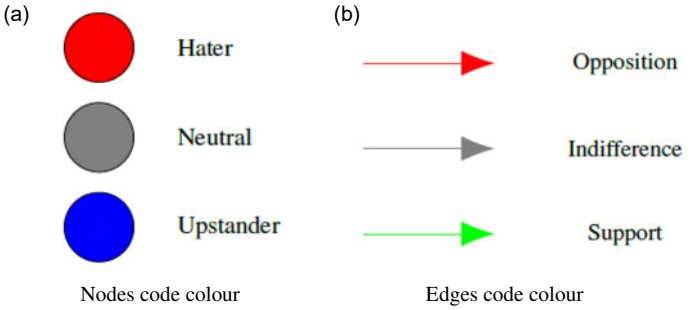


Figure 15. Proportion of the average number of responses received by each author based on tweet type.



Nodes code colour Edges code colour

Figure 16. Graph tool colour coding.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

Table 12. Percentage of node and edge types obtained by hate category

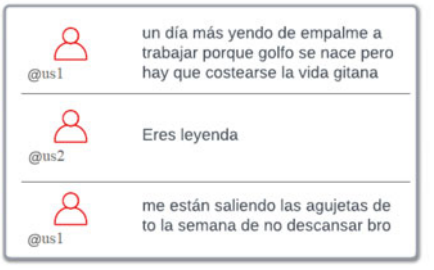
	Node Types			Edge Types		
Category	Neutrals	Haters	Upstanders	Opposition	Support	Indifference
Global	3%	90%	7%	5%	46%	49%
Cultural Minorities and Religious Groups	3%	91%	6%	5%	54%	41%
Roma People	1%	93%	6%	10%	56%	34%
Jewish People	2%	92%	6%	7%	56%	37%
Migrants	3%	90%	7%	6%	51%	42%
Muslim People	3%	91%	6%	6%	54%	40%

(a)



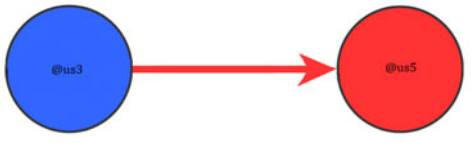
Two haters support each other. This comes from the conversation shown on the right.

(b)



First, @us1 emits a degrading stereotype towards the Roma community, to which @us2 responds, validating and praising it, and @us1 concludes by answering @us2 also in a positive way.

(c)



A hater opposes an upstander. This comes from the conversation shown on the right .

(d)



First, @us3 expresses their opinion on a controversial issue (originated by @us4, a popular network user), which @us5 opposes.

Figure 17. Examples of relationships between hater and upstander users, along with the conversations from which these graphs were obtained. The names are anonymised for privacy.

6. Conclusions and future work

This section summarises the research findings and suggests potential future research lines.

6.1 Results

Online hate has drawn the attention of many researchers and governments due to its potential to fragment communities and incite real-world violence. This research expands the existing body

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

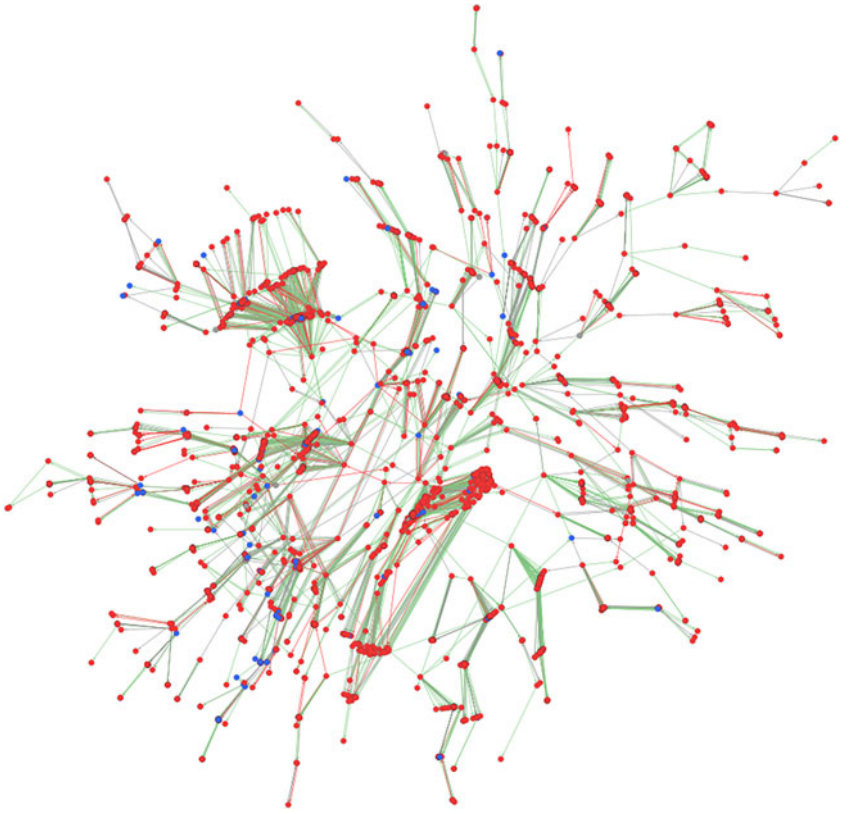


Figure 18. Graph representing a conversation stemming from a hate tweet targeting the Roma community.

of work on digital hate, addressing significant gaps in the literature, providing network analysis results, and developing several tools that will enable the collaborating entities, ONDOD and OBERAXE, to tackle online hate more effectively.

The literature review revealed that while hate speech detection has been extensively studied, a comprehensive analysis of who the actors associated with hate are, what characterises them, and how they interact within the online environment has not yet been conducted. Specifically, there is often an assumption that there are only two primary groups—haters and non-haters— when in fact, other groups might better reflect how hate circulates online. Additionally, user classification models typically rely on large datasets of tweets as input, which is problematic as Twitter&'s API is no longer freely accessible, making it difficult to easily retrieve tweets from the platform. Furthermore, social networks are usually visualised through graphs where edges represent only superficial relationships. To address these gaps, a series of research questions have been formulated. Experiments have been designed to address these questions, and novel designs have been proposed for user classifiers and network graphs.

The experimental results from the cluster analysis indicate that among the Twitter hate-related domain, there are three main groups: those who spread hate, those who oppose it, and those who remain indifferent. These groups have been labelled as haters, upstanders, and neutrals. This classification disproves the common hypothesis of only two groups. Also, it is observed that users from any of these groups can post hate tweets, upstander tweets, or neutral tweets; this classification is therefore not based on rigid definitions, such as assuming that haters only post hate tweets and none in defence of others (He *et al.* 2022). Nor is it based on overly relaxed definitions, such as classifying a hater as someone who has posted more than 1, 2, or 10 hate tweets, disregarding

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

other types of tweets (Rangel *et al.* 2021; Mishra *et al.* 2019; An *et al.* 2021). Our classification, on the other hand, considers the number of tweets of each type simultaneously. Thanks to this broader perspective and the partitions obtained with k-means, we have been able to define each cluster with more complex descriptions that take into account all available information.

In addition, a user classifier has been developed, where the input data are attributes derived from user profiles. With an accuracy of 76%, it can be inferred that profile information alone is a reliable indicator for automatically classifying users, a topic that had not been previously explored. Hence, Spanish authorities performing hate monitoring on Twitter can use this classifier to identify hateful users without the need to download a large volume of tweets. This is particularly relevant now, given that the tools currently in use (Pereira-Kohatsu *et al.* 2019; del Valle-Cano *et al.* 2023) are no longer usable without paying a high fee.

From the analysis of the most relevant model attributes, significant statistical differences between the classes have been found. Neutral users have been on Twitter for longer, have more followers, have posted more tweets, and appear on more lists. In other words, neutral users stand out for having a strong online presence, being active, and having built large platforms. On the other hand, haters and upstanders do not display these characteristics. The key feature distinguishing upstanders from haters is their profile description: more upstanders address news and social concerns in their description compared to haters.

As a final point in the user analysis, an additional study, motivated by two questions posed by the collaborating entities, illustrates that:

- More than half of users tend to change their discussion topic when switching from non-hateful to hateful tweets.
- In most accounts, non-hateful tweets receive more responses than hateful ones.

Finally, a graph visualisation tool has been developed that expands the concept of online connections to a more realistic, human-centred one. It is based on a new definition of edges that connect two users depending on whether they support or oppose each other ideologically. After analysing the graphs, it is observed that most edges represent support, and the majority of nodes are haters. In other words, haters tend to congregate and reinforce each other&'s ideas, which could be useful for identifying hate niches and applying targeted hate mitigation strategies. These generated graphs will be highly valuable to authorities analysing interactions between different types of users, as they can study the most central nodes and understand the dynamics between haters and other haters or victims of hate.

In conclusion, a reusable methodology has been developed for profiling users and mapping their interactions, which can be applied to larger datasets and adapted for other studies of online user behaviour. This can be particularly useful for sociologists and authorities studying online behaviour, an increasingly relevant issue in this digital age. We note that the proposed tools and analyses, although promising for institutional monitoring, should be interpreted as methodological contributions rather than fully operational systems. Their applicability depends on platform constraints, pretrained model performance, and the specific characteristics of the analysed dataset. The developed code has been handed over to the collaborating entities for reuse and future improvements if desired. Also, the code and the dataset are available upon request from the authors.

6.2 Limitations and future work

One of the main limitations encountered during this project was the lack of multilingual NLP libraries. While there are numerous options available in English, including those specifically designed for hate speech classification (e.g., CardiffNLP, Camacho-Collados *et al.* (2022)), this is not the case for Spanish. We have noticed a scarcity of libraries capable of automatically classifying

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

stance and argumentation (both in favour and against), personality tests tailored to this language, or Twitter topic and thematic classifiers. Note that, although the current implementation of the interaction graph focuses on Spanish content, the proposed methodology is not language-dependent. The pipeline could be adapted to other linguistic regions by replacing the classifiers with equivalent models in the target language, or by leveraging multilingual transformer models such as XLM-R or mBERT. Future work should explore these possibilities to enable the deployment of truly multilingual or language-agnostic interaction graphs, which would be especially valuable for monitoring hate speech in cross-border or multilingual contexts.

Also, although our dataset covers a three-month period and thus offers only a cross-sectional view of user behaviour, we indirectly capture some temporal aspects through profile-based features. For instance, account age and time on Twitter provide indications of behavioural persistence, while tweet frequency reflects intensity of activity over time. Our results show that haters often present shorter account lifespans and higher posting rates, suggesting patterns of instability and recurrent expression of hate. While these variables provide partial insight into temporal dynamics, future research should adopt a longitudinal design to monitor whether users become more or less hateful, or transition between classes, over extended periods. Such studies would deepen our understanding of behavioural trajectories and support the design of timely interventions.

In this study, personality was inferred from the profile description alone, which provides a minimal but informative signal. Future work should investigate more comprehensive strategies, such as integrating tweet-level features (e.g., sentiment variation, lexical richness, and stylistic markers) and behavioural attributes (e.g., activity frequency, interaction tendencies). Combining bio-based and behaviour-based information could substantially improve personality classification, although this requires access to extended user timelines, which are increasingly difficult to obtain under current API restrictions.

Regarding the selection of users for the study, we originally chose those who remained active, as they could provide a more complete and representative data sample. However, this led to selecting neutral users with significant influence on the network, such as newspaper accounts or writers. This may have introduced bias into the classifier, which only categorises a user as neutral if they have a large following and a long-standing presence on the platform. This could be reasonable since users who post content related to hate (either in favour or against it) may be more likely to have their accounts suspended or made private than those not involved with hate-related content. However, there may also be a significant number of neutral users who are not characterised by having high influence on the network. Therefore, further study is needed to examine how this has affected the classifier, and a new study could be conducted by selecting less active users.

For the cluster analysis, we chose to merge the two types of hate because extreme hate was relatively rare. Nonetheless, it could be interesting to conduct an additional analysis where each user is classified based on their proportion of offensive hate, extreme hate, and upstander tweets, to see how this additional parameter affects the results.

Looking at the interactions between users, two observations can be made. First, the percentage of haters in the graphs is very high (around 90%), which may be because the classifier developed identifies as neutral users those with a high presence on the platform, although there may be neutral users with less presence. Second, since each graph has a root tweet extracted using a hate-related keyword (related to Roma people, Jews, migrants, Muslims, or other minorities), each graph can be considered to represent a niche of haters expressing hate towards a specific group. If in the future other specific niches are to be analysed, new searches for tweets with keywords related to other types of discrimination should be performed. Our study does not explicitly address the role of coordinated behaviour among users, such as botnets or organised hate clusters. However, the interaction graph provides a foundation for detecting such dynamics, as clusters of tightly interconnected users with synchronised activity could reflect orchestrated behaviour. Future research should extend this approach by incorporating coordination indicators

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

(e.g., posting synchrony, retweet cascades, shared hashtags) to better understand the impact of collective amplification by haters or upstanders. Distinguishing between individual and coordinated behaviours would enhance both the interpretability of our results and their applicability for mitigation strategies.

Finally, we focused on racial and cultural discrimination, as this was the main concern of the collaborating entities. However, it would be relevant to explore other types of hate, such as homophobia or hate toward political figures. This would require extracting new data from the network using keywords related to these types of hate and classifying them with a classifier trained to detect those specific types of hate, as the one we used, HaterBERT, is not equipped for this purpose.

Acknowledgements. This research was conducted within the project: REAL-UP "Discurso de odio, racismo y xenofobia: mecanismos de alerta y respuesta coordinada" (Hate speech, racism and xenophobia: Alert Mechanisms and Response, analysis of the Upstander speech) funded by the European Commission within the "Citizens, Equality, Rights and Values" (CERV) program and with identification 101049673. The research of Quijano-Sánchez was conducted with financial support from the Spanish Ministry of Science and Innovation, grants PID2022-139131NB-I00 & PR47/24 PID2024-157561NB-I00. The authors would like to thank the Spanish National Office Against Hate Crimes (ONDOD) and the Spanish Observatory of Racism and Xenophobia (OBERAXE) for all the help and resources provided to this research project. We also thank the team of the UC3M-Santander Big Data Institute, in particular Iñaki Ucar for his support with Twitter data retrieval, and the director Rosa E. Lillo for facilitating logistics and providing the servers and infrastructure necessary to carry out this project.

All the financial support is gratefully acknowledged. During the preparation of this work, the authors used ChatGPT to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Ahmed M.S., **Aurpa T.T. and Anwar M.M.** (2020). Query oriented topical clusters detection for top-k trending topics in Twitter. In *2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC)*, pp. 1–6. https://doi.org/10.1109/R10-HTC49770.2020.9357047.

Alonso-Lorenzo J., **Costa-Montenegro E. and Fernandez-Gavilanes M.** (2016). Language independent big-data system for the prediction of user location on Twitter. In *2016 IEEE International Conference on Big Data (Big Data)*, pp. 2437–2446. https://doi.org/10.1109/BigData.2016.7840880.

Alotaibi A. and Abul Hasanat M.H. (2020). Racism detection in Twitter using deep learning and text mining techniques for the Arabic language. In *2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*, pp. 161–164. https://doi.org/10.1109/SMART-TECH49988.2020.00047.

An J., **Kwak H.**, **Lee C.S.**, **Jun B. and Ahn Y.-Y.** (2021). Predicting anti-Asian hateful users on Twitter during COVID-19. In Moens M.-F., Huang X., Specia L. and Yih S.W. (eds), *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 4655–4666. https://doi.org/10.18653/v1/2021.findings-emnlp.398

Anti-Defamation League. (2024). No place for hate. Available at https://www.noplaceforhate.org/ (Last accessed 27 March 2026).

Azucar D., **Marengo D. and Settanni M.** (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality & Individual Differences* **124**, 150–159. https://doi.org/10.1016/j.paid.2017.12.018

Baziotis C. (2022). Github repository. Available at https://github.com/cbaziotis/ekphrasis.

Beatty M. (2020). Graph-based methods to detect hate speech diffusion on Twitter. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 502–506. https://doi.org/10.1109/ASONAM49781.2020.9381473.

Bhimani D., **Bheda R.**, **Dharamshi F.**, **Nikumbh D. and Abhyankar P.** (2021). Identification of hate speech using natural language processing and machine learning. In *2021 2nd Global Conference for Advancement in Technology (GCAT)*, pp. 1–4. https://doi.org/10.1109/GCAT52182.2021.9587652.

Boquet M. (2012). GitHub repository. Available at https://github.com/marcboquet/spanish-names.

Bozhidarova M., **Chang J.**, **Ale-Rasool A.**, **Liu Y.**, **Ma C.**, **Bertozzi A.L.**, **Brantingham P.J.**, **Lin J. and Krishnagopal S.** (2023). Hate speech and hate crimes: A data-driven study of evolving discourse around marginalized groups. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 3107–3116. https://doi.org/10.1109/BigData59044.2023.10386312.

Bramer W.M., **De Jonge G.B.**, **Rethlefsen M.L.**, **Mast F. and Kleijnen J.** (2018). A systematic approach to searching: An efficient and complete method to develop literature searches. *Journal of the Medical Library Association: JMLA* **106**(4), 531.

Briliani A., **Irawan B. and Setianingsih C.** (2019). Hate speech detection in Indonesian language on Instagram comment section using K-Nearest neighbor classification method. In *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, pp. 98–104. https://doi.org/10.1109/IoTaIS47347.2019.8980398.

Burger J.D., **Henderson J.**, **Kim G. and Zarrella G.** (2011). Discriminating gender on Twitter. In Barzilay R. and Johnson M. (eds), *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK: Association for Computational Linguistics, pp. 1301–1309. https://aclanthology.org/D11-1120

Caldera H., **Meedin G. and Perera I.** (2020a). A peer recommendation model to avoid hate speech engagements in multiplex social networks. In *2020 From Innovation to Impact (FITI)*, vol. **1**, pp. 1–6. https://doi.org/10.1109/FITI52050.2020.9424905.

Caldera H.M.M., **Meedin G.S.N. and Perera I.** (2020b). Time series based trend analysis for hate speech in Twitter during COVID 19 pandemic. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pp. 1–2. https://doi.org/10.1109/ICTer51097.2020.9325491.

Calma J. (2023). Twitter&'s new API policy could chill academic research. Available at URL https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research (accessed 07 May 2024).

Camacho-Collados J., **Rezaee K.**, **Riahi T.**, **Ushio A.**, **Loureiro D.**, **Antypas D.**, **Boisson J.**, **Espinosa-Anke L.**, **Liu F. and Martínez-Cámara E.** (2022). TweetNLP: Cutting-edge natural language processing for social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, December 7-11, 2022. Abu Dhabi, U.A.E.: Association for Computational Linguistics.

Carhuancho-Bazan A., **Nuñez-Lazo S. and Ugarte W.** (2023). NoHateS: A transformers-based approach for real-time hate speech detection in Spanish. In *2023 IEEE XXX International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, pp. 1–7. https://doi.org/10.1109/INTERCON59652.2023.10326033.

Chandra M., **Reddy M.**, **Sehgal S.**, **Gupta S.**, **Buduru A.B. and Kumaraguru P.** (2021). A virus has no religion": Analyzing Islamophobia on Twitter during the COVID-19 outbreak. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media, HT '21*. New York, NY, USA: Association for Computing Machinery, pp. page67–77. https://doi.org/10.1145/3465336.3475111.

Chaudhary M. (2020). K-means clustering in machine learning. Medium. https://medium.com/@cmukesh8688/k-means-clustering-in-machine-learning-252130c85e23

Chhikara M. and Malik S.K. (2022). Classification of cyber hate speech from social networks using machine learning. In *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)*, pp. 419–423. https://doi.org/10.1109/SMART55829.2022.10047042.

Damo G., **Ocampo N.B.**, **Cabrio E. and Villata S.** (2026). "Detectors lead, LLMs follow": Integrating LLMs and traditional models on implicit hate speech detection to generate faithful and plausible explanations. *Data & Knowledge Engineering* **162**, 102535.

Defersha N.B., **Kekeba K. and Kaliyaperumal K.** (2021). Tuning hyperparameters of machine learning methods for Afan Oromo hate speech text detection for social media. In *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, pp. 596–604. https://doi.org/10.1109/ICCCT53315.2021.9711850.

Valle-Cano G., **Quijano-Sánchez L.**, **Liberatore F. and Gómez J.A.** (2023). SocialHaterBERT: A dichotomous approach for automatically detecting hate speech on Twitter through textual analysis and user profiles. *Expert Systems with Applications* **216**, 119446. https://doi.org/10.1016/j.eswa.2022.119446

Dinh V.-C., **Vo T.-D.**, **Nguyen M.-P.T. and Do T.-H.** (2023). A scalable hate speech detection system for Vietnamese social media using real-time big data processing and distributed deep learning. In *2023 International Conference on Advanced Technologies for Communications (ATC)*, pp. 95–100. https://doi.org/10.1109/ATC58710.2023.10318848.

Doan L., **Nguyen P.**, **Phan T. and Do T.** (2022). An implementation of large scale hate speech detection system for streaming social media data. In *2022 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, pp. 155–159. https://doi.org/10.1109/COMNETSAT56033.2022.9994299.

Doğanç M. and Markov I. (2023). From generic to personalized: Investigating strategies for generating targeted counter narratives against hate speech. In Chung Y.-L., Bonaldi H., Abercrombie G. and Guerini M. (eds), *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*. Prague, Czechia: Association for Computational Linguistics, pp. 1–12. https://aclanthology.org/2023.cs4oa-1.1

Duong C., **Zhang L. and Lu C.-T.** (2022). HateNet: A graph convolutional network approach to hate speech detection. In *2022 IEEE International Conference On Big Data (Big Data)*, pp. 5698–5707. https://doi.org/10.1109/BigData55660.2022.10020510.

Everitt B.S., **Landau S.**, **Leese M. and Stahl D.** (2011). *Cluster Analysis*. Wiley.

Febriana T. and Budiarto A. (2019). Twitter dataset for hate speech and cyberbullying detection in Indonesian language. In *2019 International Conference on Information Management and Technology (ICIMTech)*, vol. **1**, 379–382. https://doi.org/10.1109/ICIMTech.2019.8843722.

Frenda S., **Patti V. and Rosso P.** (2023). Killing me softly: Creative and cognitive aspects of implicitness in abusive language online. *Natural Language Engineering* **29**(6), 1516–1537.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

Golbeck J., **Robles C. and Turner K.** (2011). Predicting personality with social media. In *Proceedings of the International Conference on Human Factors in Computing Systems, CHI 2011, Extended Abstracts Volume, Vancouver, BC, Canada, May 7-12, 2011*. Vancouver, BC, Canada: ACM, pp. 253–262. https://doi.org/10.1145/1979742.1979614

Google LLC. (2024). Be internet awesome. Available at https://beinternetawesome.withgoogle.com/en_us/ (accessed 29 April 2024).

Gupta G., **Kadian K.**, **Jain R.**, **Dwivedi V. and Sharma A.** (2023). Real-time hate speech detection in live streaming platforms using quantum machine learning. In *2023 26th Conference of the Oriental COCOSDA International Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pp. 1–6. https://doi.org/10.1109/O-COCOSDA60357.2023.10482977.

Harby A.A. and Zulkernine F. (2023). A comparative analysis of graph neural networks for fake news detection. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 1215–1222. https://doi.org/10.1109/COMPSAC57700.2023.00184.

He B., **Ziems C.**, **Soni S.**, **Ramakrishnan N.**, **Yang D. and Kumar S.** (2022). Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '21*, New York, NY, USA: Association for Computing Machinery, pp. 90–94. https://doi.org/10.1145/3487351.3488324

Hebert L., **Golab L. and Cohen R.** (2022). Predicting hateful discussions on Reddit using graph transformer networks and communal context. In *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pp. 9–17. https://doi.org/10.1109/WI-IAT55865.2022.00012.

Heidari M., **Jones J.H. and Uzuner O.** (2020). Deep contextualized word embedding for text-based online user profiling to detect social bots on Twitter. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pp. 480–487. https://doi.org/10.1109/ICDMW51313.2020.00071.

Herodotou H., **Chatzakou D. and Kourtellis N.** (2021). Catching them red-handed: Real-time Aggression Detection on Social Media. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pp. 2123–2128. https://doi.org/10.1109/ICDE51399.2021.00211.

Jahan S.S.S., **Dutta P.**, **Hassan Khan H.M.M.**, **Shahariar Karim Badhon M. and Rab R.** (2022). A multiclass approach to identify misogynistic Bangla text from social media. In *2022 4th International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pp. 1–6. https://doi.org/10.1109/STI56238.2022.10103269.

Jigsaw. (2024). Perspective api. Available at https://perspectiveapi.com/ (Last accessed 27 March 2026).

Kaushal R., **Ghose V. and Kumaraguru P.** (2019). Methods for user profiling across social networks. In *2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pp. 1572–1579. https://doi.org/10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00231.

Lee S. and Yun D. (2003). Clustering categorical and numerical data: A new procedure using multidimensional scaling. *International Journal of Information Technology & Decision Making* **2**(01), 135–159.

Lee W., **Oh K.**, **Lim C. and Choi H.** (2014). User profile extraction from Twitter for personalized news recommendation. In *16th International Conference on Advanced Communication Technology*, pp. 779–783. https://doi.org/10.1109/ICACT.2014.6779068.

Liu C., **Li F. and Li L.** (2021). Research on gender prediction for social media user profiling by machine learning method. In *2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*, pp. 831–836. https://doi.org/10.1109/CISCE52179.2021.9445922.

Los Angeles Times. (2009). Twitter creator jack dorsey illuminates the site&'s founding document, part i. Available at URL https://www.latimes.com/archives/blogs/technology-blog/story/2009-02-18/twitter-creator-jack-dorsey-illuminates-the-sites-founding-document-part-i (Last accessed 27 March 2026).

Martins R., **Gomes M.**, **Almeida J.J.**, **Novais P. and Henriques P.** (2018). Hate speech classification in social media using emotional analysis. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pp. 61–66. https://doi.org/10.1109/BRACIS.2018.00019.

McCrae R.R. and Costa P.T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality & Social Psychology* **52**(1), 81.

Microsoft Corporation. (2024). Digital civility. Available at https://www.microsoft.com/en-us/online-safety/digital-civility (Last accessed 27 March 2026).

Mishra P., **Tredici M.D.**, **Yannakoudakis H. and Shutova E.** (2019). Author profiling for hate speech detection. arXiv preprint arXiv: 1902.06734.

Nakano T., **Suda T.**, **Okaie Y. and Moore M.J.** (2016). Analysis of cyber aggression and cyber-bullying in social networking. In *Tenth IEEE International Conference on Semantic Computing, ICSC 2016, Laguna Hills, CA, USA, February 4-6, 2016*, Laguna Hills, CA, USA: IEEE Computer Society, pp. 337–341.

Nayla A., **Setianingsih C. and Dirgantoro B.** (2023). Hate speech detection on Twitter using BERT algorithm. In *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, pp. 644–649. https://doi.org/10.1109/ICCoSITE57641.2023.10127831.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

Nguyen D., **Gravel R.**, **Trieschnigg D. and Meder T.** (2021). How old do you think i am? a study of language and age in Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* **7**(1), 439–448. https://doi.org/10.1609/icwsm.v7i1.14381

Noorian Z., **Ghenai A.**, **Moradisani H.**, **Zarrinkalam F. and Alavijeh S.Z.** (2024). User-centric modeling of online hate through the lens of psycholinguistic patterns and behaviors in social media. *IEEE Transactions on Computational Social Systems* **11**, 1–13. https://doi.org/10.1109/TCSS.2024.3359010.

Oroh A.J., **Bandung Y. and Zagi L.M.** (2021). Detection of the key actor of issues spreading based on social network analysis in Twitter social media. In *2021 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, pp. 206–212. https://doi.org/10.1109/APWiMob51111.2021.9435268.

Panchala G.H., **S Sasank V.V.**, **Harshitha Adidela D.R.**, **Yellamma P.**, **Ashesh K. and Prasad C.** (2022). Hate speech & offensive language detection using ML & NLP. In *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 1262–1268. https://doi.org/10.1109/ICSSIT53264.2022.9716417.

Pandey K. and Mittal A. (2016). User profiling on Tumblr through blog posts. In *2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT)*, pp. 85–89. https://doi.org/10.1109/ICCTICT.2016.7514557.

Pawar A., **Gawali P.**, **Gite M.**, **Jawale M.A. and William P.** (2022). Challenges for hate speech recognition system: Approach based on solution. In *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pp. 699–704. https://doi.org/10.1109/ICSCDS53736.2022.9760739.

Peddinti S.T., **Ross K.W. and Cappos J.** (2014). "On the internet, nobody knows you're a dog": A Twitter case study of anonymity in social networks. In *Proceedings of the Second ACM Conference on Online Social Networks, COSN '14*. New York, NY, USA: Association for Computing Machinery, pp. 83–94, https://doi.org/10.1145/2660460.2660467.

Pereira-Kohatsu J.C., **Quijano-Sánchez L.**, **Liberatore F. and Camacho-Collados M.** (2019). Detecting and monitoring hate speech in Twitter. *Sensors* **19**(21), 4654. https://doi.org/10.3390/s19214654

Perera S., **Perera I. and Ahangama S.** (2023). Influence rank of hate users in Twitter: An analysis during Sri Lanka&'s political and economic crisis. In *2023 3rd International Conference on Advanced Research in Computing (ICARC)*, pp. 148–153. https://doi.org/10.1109/ICARC57651.2023.10145632.

Plaza-Del-Arco F.M., **Molina-Gonzalez M.D.**, **Urena-Lopez L.A. and Martin-Valdivia M.T.** (2021). A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access* **9**, 112478–112489. https://doi.org/10.1109/ACCESS.2021.3103697.

Priadana A., **Maarif M.R. and Habibi M.** (2020). Gender prediction for Instagram user profiling using deep learning. In *2020 International Conference on Decision Aid Sciences and Application (DASA)*, pp. 432–436. https://doi.org/10.1109/DASA51403.2020.9317143.

Qureshi K.A. and Sabih M. (2021). Un-compromised credibility: Social media based multi-class hate speech classification for text. *IEEE Access* **9**, 109465–109477. https://doi.org/10.1109/ACCESS.2021.3101977.

Rangel F., **De La Peña Sarracén G.L.**, **Chulvi B.**, **Fersini E. and Rosso P.** (2021). Profiling hate speech spreaders on Twitter task at PAN. In *CLEF 2021 – Conference and Labs of the Evaluation Forum*. Bucharest, Romania.

Rao D., **Paul M.**, **Fink C.**, **Yarowsky D.**, **Oates T. and Coppersmith G.** (2021). Hierarchical bayesian models for latent attribute detection in social media. *Proceedings of the International AAAI Conference on Web and Social Media* **5**(1), 598–601. https://doi.org/10.1609/icwsm.v5i1.14197

Roy P.K., **Tripathy A.K.**, **Das T.K. and Gao X.-Z.** (2020). A framework for hate speech detection using deep convolutional neural network. *IEEE Access* **8**, 204951–204962. https://doi.org/10.1109/ACCESS.2020.3037073.

Sachdeva J., **Chaudhary K.K.**, **Madaan H. and Meel P.** (2021). Text based hate-speech analysis. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pp. 661–668. https://doi.org/10.1109/ICAIS50930.2021.9396013.

Schubert E., **Sander J.**, **Ester M.**, **Kriegel H.P. and Xu X.** (2017). DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems* **42**(3), 1–21. https://doi.org/10.1145/3068335.

Shrimali S. (2022). A natural language processing and machine learning-based framework to automatically identify cyber-bullying and hate speech in real-time. In *2022 IEEE MIT Undergraduate Research Technology Conference (URTC)*, pp. 1–5. https://doi.org/10.1109/URTC56832.2022.10002243.

Shukla S., **Nagpal S. and Sabharwal S.** (2022). Hate speech detection in Hindi language using BERT and convolution neural network. In *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pp. 642–647. https://doi.org/10.1109/ICCCIS56430.2022.10037649.

Siddique F.B., **Bertero D. and Fung P.** (2019). GlobalTrait: Personality alignment of multilingual word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. **33**, pp. 7015–7022.

Singh A., **Malhotra H. and Savaridassan P.** (2023a). Purge profanity: An AI system to detect profanity/obscenity from texts and images. In *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, pp. 421–428. https://doi.org/10.1109/ICECAA58104.2023.10212151.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

Singh R.K., **Sanjay H.A.**, **Pramod Jain S.A.**, **Rishi H. and Bhardwaj S.** (2023b). NLP based hate speech detection and moderation. In *2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, pp. 1–5. https://doi.org/10.1109/CSITSS60515.2023.10333320.

Skowron M., **Tkalcic M.**, **Ferwerda B. and Schedl M.** (2016). Fusing social media cues: Personality prediction from Twitter and Instagram. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016*, Montreal, Canada: ACM, pp. 107–108. https://doi.org/10.1145/2872518.2889368.

Stakhiyevich P. and Huang Z. (2019). Building user profiles based on user interests and preferences for recommender systems. In *2019 IEEE International Conferences on Ubiquitous Computing & Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking and Services (SmartCNS)*, pp. 450–455. https://doi.org/10.1109/IUCC/DSCI/SmartCNS.2019.00101.

Statista (2025). Global social networks ranked by number of users. Available at https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (Last accessed 27 March 2026).

Syam S.S., **Irawan B. and Setianingsih C.** (2019). Hate speech detection on Twitter using long short-term memory (LSTM) method. In *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pp. 305–310. https://doi.org/10.1109/ICITISEE48480.2019.9003992.

Ting I., **Wang S.**, **Chi H. and Wu J.** (2013). Content matters: A study of hate groups detection based on social networks analysis and web mining. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pp. 1196–1201. https://doi.org/10.1109/ASONAM.2013.6785855.

Toledo-Ronen O., **Orbach M.**, **Bilu Y.**, **Spector A. and Slonim N.** (2020). Multilingual argument mining: Datasets and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 303–317.

Treves B., **Masud M.R. and Faloutsos M.** (2022). URLytics: Profiling forum users from their posted URLs. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 510–513. https://doi.org/10.1109/ASONAM55673.2022.10068682.

Turjya S.M., **Kumari R.**, **Swain S. and Bandyopadhyay A.** (2023). Multilingual hate speech and offensive language detection. In *2023 OITS International Conference on Information Technology (OCIT)*, pp. 660–664. https://doi.org/10.1109/OCIT59427.2023.10431222.

Twitter. (2023). X verification requirements - how to get the blue check. Available at: https://help.twitter.com/en/managing-your-account/about-x-verified-accounts (accessed 21 October 2023).

United Nations. (2024). Hate speech. Available at https://www.un.org/en/hate-speech (Last accessed 27 March 2026).

Utomo M.N.Y., **Adji T.B. and Ardiyanto I.** (2018). Geolocation prediction in social media data using text analysis: A review. In *2018 International Conference on Information and Communications Technology (ICOIACT)*, pp. 84–89. https://doi.org/10.1109/ICOIACT.2018.8350674.

Vasanthakumar G.U., **Shashikumar D.R. and Suresh L.** (2019). Profiling social media users, a content-based data mining technique for Twitter users. In *2019 1st International Conference on Advances in Information Technology (ICAIT)*, pp. 33–38. https://doi.org/10.1109/ICAIT47043.2019.8987304.

Walther J.B. (2022). Social media and online hate. *Current Opinion in Psychology* **45**, 101298. https://doi.org/10.1016/j.copsyc.2021.12.010

Wang Y., **Liu J.**, **Huang Y. and Feng X.** (2016). Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs. *IEEE Transactions on Knowledge and Data Engineering* **28**(7), 1919–1933. https://doi.org/10.1109/TKDE.2016.2531661.

Williams E., **Gray J. and Dixon B.** (2017). Improving geolocation of social media posts. *Pervasive and Mobile Computing* **36**, 68–79. https://doi.org/10.1016/j.pmcj.2016.09.015. Special Issue on Pervasive Social Computing.

Wright R. (2014). How to fix Twitter. Slate, 10. https://slate.com/technology/2014/10/twitter-character-limits-make-the-social-network-more-toxic.html

Appendix A. Profile features to classify users

This section displays the list of input attributes for the user classifier (Tables A1 and A2). These are attributes that are extracted directly from the profile information or through processing of the username and description.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

Table A1. Features extracted from general profile data. Those marked with an asterisk are implicit

Name	Description	Values	Type
GENERAL FEATURES			
tweets	N° of tweets	Numeric	Metadata
followers	N° of followers	Numeric	Metadata
followings	N° of people followed	Numeric	Metadata
ratio	Ratio of followers to following	Numeric	Manual
anonymity∗	Anonymity	{Anonymous, partially anonymous, identifiable, highly identifiable}	Manual
time	Time on Twitter	Numeric (in days)	Manual
lists	N° of lists the user is added to	Numeric	Metadata
frequency	Tweet frequency	Numeric (tweets/day)	Manual
verified	Verified	True/False	Metadata
type verification	Type of verification	{Blue, business, government, none}	Metadata
location*	Location set	{True, False}	Manual
gender*	User&'s gender	{Female, male, none}	Manual

Type indicates how each feature is obtained: ***Metadata*** *(directly extracted from the user profile),* ***Manual*** *(derived from structured fields or simple transformations), and* ***Model-based*** *(inferred using pretrained NLP models).*

Table A2. Features extracted from users' profile descriptions. Those marked with an asterisk are implicit

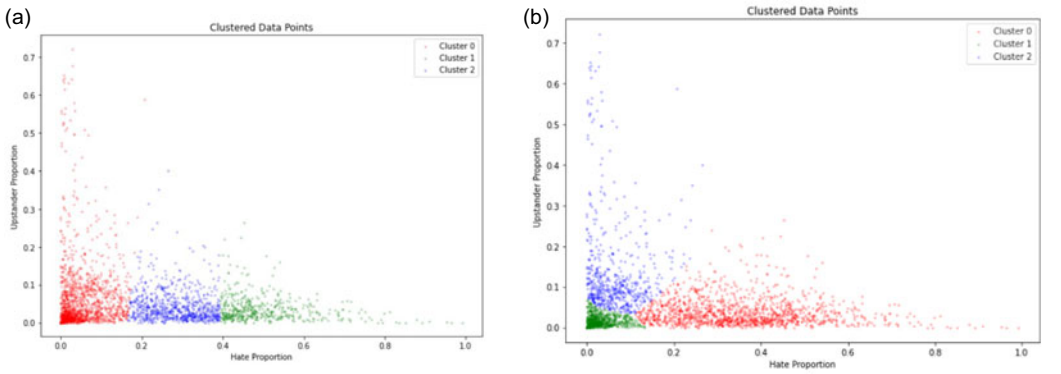
Name	Description	Values	Type
DESCRIPTIONFEATURES			
cashtags	N° of cashtags	Numeric	Metadata
hashtags	N° of hashtags	Numeric	Metadata
mentions	N° of mentions	Numeric	Metadata
urls	N° of URLs	Numeric	Metadata
sentiment	Sentiment∗	{Positive, negative, neutral}	Model-based
hate	Type of hate∗	{Offensive-hate, extreme-hate, upstander, neutral}	Model-based
topic	Topic∗	{arts_and_culture, business_and_entrepreneurs,celebrities_and_pop_culture, diaries_and_daily_life,family, fashion_and_style,film_tv_and_video,fitness_and_health, food, games,learning_and_education, news_and_social_concerns,other_hobbies, relationships, music,science_and_technology, sports,travel_and_adventure, youth_and_student_life}	Model-based
personality	Personality*	{extraversion, openness, conscientiousness, agreeableness, neuroticism }	Model-based

Type indicates how each feature is obtained: ***Metadata*** *(directly extracted from the user profile),* ***Manual*** *(derived from structured fields or simple transformations), and* ***Model-based*** *(inferred using pretrained NLP models).*

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

Appendix B. Validation of spherical transformation for k-means

Let us analyse the performance of the k-means algorithm on the data both before and after applying the transformation. With $k = 3$, we observe in Figure B1b that the clusters differentiate between haters (in red), upstanders (in blue), and neutral (in green). In contrast, the clusters shown in Figure B1a appear to differ primarily based on the level of hate alone, without leveraging upstander proportions for separation. If we were to use this division, users would be labelled based on whether they emit low, moderate, or high levels of hate, disregarding the role of neutral and upstander users. Also, these clusters do not consider the local maxima in the density estimation curves, identified in Figure 8b.



With untransformed data

With transformed data

Figure B1. Scatter plot of proportions grouped into 3 clusters.

Appendix C. Classification models comparison

In this appendix, we present the results of the machine learning models evaluated for the classifier creation task. For all models, hyperparameter optimisation was conducted using grid search, combined with 10-fold cross-validation and random oversampling. The metrics reported in Table C1 clearly demonstrate the superiority of XGBoost over the other models.

Table C1. Classification models comparison. The best results for each metric are highlighted in bold

Metrics	Logistic Regression	Elastic Net	SVM	XGBoost	CNN
Accuracy	0.2936	0.2826	0.5174	**0.7560**	0.6037
F1-score (Weighted)	0.1372	0.1294	0.3587	**0.7571**	0.5946
F1-score (Macro)	0.1559	0.1519	0.2367	**0.7384**	0.5428
AUROC	0.6900	0.7207	0.3142	**0.9068**	0.6731
AUPRC	0.5023	0.5417	0.2560	**0.8284**	0.4726

The hyperparameters utilised in the experiments, along with their respective value ranges and the optimal values obtained, are given in the following:

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

CNN

```
param_grid = {
    'filters': [16, 32],
    'kernel_size': [3, 5],
    'dropout_rate': [0.3, 0.5],
    'dense_units': [64, 128],
}
```

Best hyperparameters: {'filters': 16, 'kernel_size': 3, 'dropout_rate': 0.3, 'dense_units': 128}

Elastic Net

```
param_grid = {
    'C': [0.1, 1, 10],
    'l1_ratio': [0, 0.5, 1],
    'solver': ['saga'],
    'penalty': ['elasticnet']
}
```

Best hyperparameters: {'C': 0.1, 'l1_ratio': 0.5, 'penalty': 'elasticnet', 'solver': 'saga'}

SVM

```
param_grid = {
    'C': [0.1, 1, 10],
    'kernel' : ["poly", "rbf"],
    'degree' : [2, 3],
    'gamma' : ['scale']
}
```

Best hyperparameters: {'C': 1, 'degree': 2, 'gamma': 'scale', 'kernel': 'rbf'}

XGBoost

```
params = {
    'eta': [0, 0.3, 0.5, 1, 2],
    'min_child_weight': [1, 5, 10],
    'gamma': [0, 0.5, 1, 1.5, 2, 5],
    'colsample_bytree': [0.6, 0.8, 1.0],
    'max_depth': [6, 8, 10],
    'reg_lambda': [1, 5],
    'alpha': [0, 1]
}
```

Best hyperparameters: {'reg_lambda': 5, 'min_child_weight': 1, 'max_depth': 6, 'gamma': 0, 'eta': 0.5, 'colsample_bytree': 0.6, 'alpha': 0}

Appendix D. Keywords of the ALRECO project

In the following, the keywords (in Spanish) of the ALRECO project are provided, according to their topic. For accessibility, English translations of the Spanish keywords are provided below each list.

Disclaimer: The following content contains language that may be offensive. Reader discretion is advised.

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

D.1 General hate keywords

Spanish: stopinvasion, mantero, manteros, topmanta, eurabia, terrorismo, atentado, atentados, guerra, sumision, boicot, frontera, fronteras, expulsion, frnaco, fdanco, franko, hideologia.
English: stop invasion, street vendor, street vendors, street market, Eurabia, terrorism, attack, attacks, war, submission, boycott, border, borders, expulsion, Franco (misspelled variants: frnaco, fdanco, franko), ideology.

D.2 Keywords for cultural minorities and religious ethnicities

Spanish: negro, negra, negros, negras, panchito, panchita, panchitos, panchitas, indio, india, indios, indias, conguito, conguita, conguitos, conguitas, machupichu, machupichus.
English: black (man/woman/men/women), "panchito" (pejorative for Latin American), Indian (man/woman/men/women), "Conguito" (derogatory), Machu Picchu (used pejoratively).

D.3 Keywords for the Roma community

Spanish: gitano, gitana, gitanos, gitanas, romani, romanis, patriarca, hitana, hitano.
English: gypsy (man/woman/men/women), Romani, patriarch, misspellings (hitana, hitano).

D.4 Keywords for the Jewish community

Spanish: judio, gudio, judia, judios, judias, sion, sionista, sionistas, hebreo, israeli, israelis, holocuento, holocausto, hitler, jitler, Hittler, nazi, nasi, jeit, ebreo.
English: Jew (male/female/singular/plural, with misspellings), Zion, Zionist(s), Hebrew, Israeli(s), "holo-hoax" (holocuento), Holocaust, Hitler (with misspellings), Nazi (with misspellings), derogatory variants (jeit, ebreo).

D.5 Keywords about migrants (first dataset)

Spanish: latinos, latinas, rumano, rumana, rumanos, rumanas, subsahariano, subsahariana, subsaharianos, subsaharianas, sudamericano, sudamericana, sudamericanos, sudamericanas, sudaca, sudacas, chino, china, chinos, chinas, chinito, chinita, chinitos, chinitas, clandestinos, clandestinas, clandestino, clandestina, mena, menas.
English: Latino(s), Romanian(s), Sub-Saharan (male/female/singular/plural), South American(s), "sudaca" (derogatory for Latin Americans), Chinese (man/woman/men/women), "chinito" (diminutive/derogatory), clandestine/illegal migrant(s), unaccompanied minor(s) (MENA, from Spanish acronym).

D.6 Keywords about migrants (second dataset)

Spanish: migracion, inmigracion, migrante, migrantes, inmigrante, inmigrantes, extranjero, extranjera, extranjeros, extranjeras, refugiado, refugiada, refugiados, refugiadas, marroqui, marroquis, magrebi, magrebis, africano, africana, africanos, africanas, latino, latina.
English: migration, immigration, migrant(s), immigrant(s), foreigner(s) (male/female), refugee(s) (male/female), Moroccan(s), Maghrebi(s), African(s), Latino(a).

D.7 Keywords about the Muslim community

Spanish: musulman, musulmana, musulmanes, musulmanas, musulmanos, mslmns, moro, mora, moros, moras, islamico, islam, arabe, arabes, sarraceno, sarracenos, yihadista, yihadistas, yijadista,

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press

yijadistas, islamista, islamistas, fundamentalista, fundamentalistas, fundamentalismo, hijab, hiyab, burca.

English: Muslim(s) (male/female, with misspellings), "moro" (derogatory for Arabs/Muslims), Islamic, Islam, Arab(s), Saracen(s), jihadist(s), Islamist(s), fundamentalist(s), fundamentalism, hijab (with spelling variants), burqa.

Cite this article: Ramiro-López I, Quijano-Sánchez L and Liberatore F (2026). Hate speech on Twitter: Profiling users and interaction analysis in the Spanish language. *Natural Language Processing* **32**, 470–517. https://doi.org/10.1017/nlp.2026.10024

https://doi.org/10.1017/nlp.2026.10024 Published online by Cambridge University Press