PROTOCOL AND SYSTEM OF INDICATORS

to detect hate speech on social media

ALRECO

GOBIERNO DE ESPAÑA | MINISTERIO DE INCLUSIÓN, SEGURIDAD SOCIAL Y MIGRACIONES

oberaxe
Observatorio Español del Racismo y la Xenofobia

Catalogue of publications of the General State Administration
https://cpage.mpr.gob.es

CONTENTS

CONCEPTUAL FRAMEWORK AND GENDER APPROACH

CONCEPTUAL
FRAMEWORK AND
GENDER APPROACH

1

Introduction

PROTOCOL AND
SYSTEM OF INDICATORS
**to detect hate speech on
social media**

1. Introduction

The AL-RE-CO project (Hate speech, racism and xenophobia: alert and coordinated response mechanisms) **aims to help government authorities to better identify, analyse, monitor and assess online hate speech with a view to designing shared strategies to contend with discourse motivated by racism, xenophobia, Islamophobia and antisemitism. The project is co-financed by the EU Commission's Directorate-General for Justice under the European Union's Rights, Equality and Citizenship Programme and is implemented by a consortium of four institutions and entities coordinated by the Spanish Observatory on Racism and Xenophobia (OBERAXE) under the auspices of the Directorate-General for Integration and Humanitarian Aid of the State Secretariat for Migration at the Ministry of Labour, Migration and Social Security.**

The project engages in three main lines of action:

1. To devise a protocol and system of indicators, including a Best Practices Report;
2. To develop an IT tool to detect and monitor hate speech (on Twitter);
3. To design shared strategies between key actors, institutions and agents, including groups affected by hate speech.

In phase one of the work, we identified experiences and best practices developed mainly in the European Union enabling the development of a protocol and system of indicators. From this review of experiences at international level (projects, reports, platforms, tools and scientific literature), we compiled the **"Report of best practices or similar experiences developed in the EU to identify, analyse, monitor and assess online racist, xenophobic, Islamophobic and anti-Semitic hate speech"**, the project's first deliverable (D 2.1).

The second deliverable, the "**Protocol and System of indicators to detect online hate speech**" is presented below. It aims to develop an action protocol with a system of indicators and search criteria to identify speech that promotes online racism, xenophobia and hatred. The system also includes **early warning indicators** with which to assess the intensity, severity, distribution and potential impact of hate speech, the objective being to devise initiatives to prevent possible discriminatory incidents and hate crimes.

2

Conceptual framework: hate speech

2. Conceptual framework: hate speech

The Internet has been used to reinforce and magnify hate speech and, to a certain extent, this has normalised it. Due to online dissemination, the general public (especially young people) are being exposed to hate speech on a regular basis. These novel tools are used by organizations, groups and individuals to spread their racist, xenophobic, Islamophobic and anti-Semitic ideas thus increasing their impact. It is fair to say that that the online dissemination of hate speech has grown exponentially in all European countries, especially through social media.

According to the European Commission's fourth Monitoring Report on the Code of Conduct on countering illegal hate speech online[1] -January 2019-, racist and xenophobic hate speech, mainly targeting immigrants, refugees and ethnic minorities, continues to be predominant and the object of the greatest number of social media complaints. For example, 17% of the complaints filed in 2018 were for hate speech targeting immigrants (xenophobia), 15.6% for sexual orientation, 13% for Islamophobia, 12.2% for anti-Gypsyism and 10.1% for anti-Semitic speech.

This situation is compounded by the broad debate on how to define the legal, conceptual, ideological and social impact of hate speech. People commonly speak of 'hate speech', 'hate crime', 'hate incident', etc. and terms such as 'racism', 'xenophobia', 'Islamophobia' and 'ant-Semitism' are likewise subject to discussion and interpretation. Therefore, for the purposes of this project, a **conceptual framework** had to be established prior to developing the protocol and the indicators included in this document. We based that framework on a set of definitions established and agreed upon in laws, regulations and national and international commitments, and on the Agreement signed between the General Council of the Judiciary, the Office of the Prosecutor General, the Ministry of Justice, the Ministry of the Interior, the Ministry of Education and Vocational Training, the Ministry of Labour, Migration and Social Security, the Ministry of the Presidency, Relations with the Courts and Equality, the Ministry of Culture and Sports and the Centre for Legal Studies, to cooperate at institutional level in the fight against

---

1 Accessible at **https://ec.europa.eu/info/sites/info/files/code_of_conduct_factsheet_7_web.pdf**

racism, xenophobia, LGBTIphobia and other forms of intolerance[2]. First of all, Recommendation No. R (97) 20 of 30 October 1997 of the Committee of Ministers of the Council of Europe on hate speech which defines the latter as follows: "... the term 'hate speech' shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, antisemitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin."

General Policy Recommendation No. 15 on combating hate speech and its explanatory memorandum, of the Council of Europe's European Commission against Racism and Intolerance (ECRI)[3]. According to ECRI, **hate speech** is to be understood as "the advocacy, promotion or incitement, in any form, of the denigration, hatred or vilification of a person or group of persons, as well as any harassment, insult, negative stereotyping, stigmatization or threat in respect of such a person or group of persons and the justification of all the preceding types of expression, on the ground of race, colour, descent, national or ethnic origin, age, disability, language, religion or belief, sex, gender, gender identity, sexual orientation and other personal characteristics or status". Since all human beings belong to the same species, ECRI rejects theories based on the existence of different races. However, in this Recommendation ECRI uses the term "race" in order to ensure that those persons who are generally and erroneously perceived as belonging to another race are not excluded from the protection provided for under the Recommendation.

---

2 Joint agreement signed by the General Council of the Judiciary, the Office of the Prosecutor General, the Ministry of Justice, the Ministry of the Interior, the Ministry of Education and Vocational Training, the Ministry of Labour, Migration and Social Security, the Ministry of the Presidency, Relations with the Courts and Equality, the Ministry of Culture and Sports and the Centre for Legal Studies, to cooperate at institutional level in the fight against racism, xenophobia, LGBTIphobia and other forms of intolerance. Available at:
http://www.inclusion.gob.es/oberaxe/ficheros/ejes/cooperacion/Acuerdo_insterinsticuional_original.pdf

3 ECRI GPR No. 15 on combating hate speech and its explanatory memorandum. Available at
http://www.inclusion.gob.es/oberaxe/ficheros/documentos/2016_12_21-Recomendacion_ECRI_NO_15_Discurso_odio-ES.pdf

For its part, the **OSCE** Ministerial Council (December 2003) invited States to incorporate hate crimes in one way or another and to produce reliable information and statistics (Decision on Tolerance and Non-Discrimination No. 4/3). It agreed on the concept **of Hate Crime** to refer to **crimes motivated by intolerance**, i.e. by **prejudice or animosity** towards the victim, OSCE (2003) defining it as: "Any criminal offence, including offences against persons and property where the victim, place or object of the prejudice are selected because of their connection, affiliation, support and actual and supposed membership to a group that can be based on "race," national or ethnic origin, language, colour, religion, age, physical or mental disability, sexual orientation or other similar factors, whether real or imagined."

As already noted, the AL-RE-CO project focuses on speech motivated by racism, xenophobia, Islamophobia and antisemitism. However, for practical purposes, the reference to the concept of anti-Gypsyism[4] specifically indicated has been considered within the framework of racism. In that connection we present the following concepts:

---

4   Available at **https://rm.coe.int/ecri-general-policy-recommendation-no-13-on-combating-anti-gypsyism-an/16808b5aef**

CONCEPTUAL FRAMEWORK AND GENDER APPROACH

Racism

shall mean the belief that a ground such as race, colour, language, religion, nationality or national or ethnic origin justifies contempt for a person or a group of persons, or the notion of superiority of a person or a group of persons *(GPR No 7). Although religion is not included in the definition of racial discrimination in Article 1 of the **International Convention on the Elimination of All Forms of Racial Discrimination**[5], the Committee on the Elimination of Racial Discrimination recognises, in the light of the principle of intersectionality, that racist hate speech extends to speech "targeting persons belonging to certain ethnic groups who profess or practice a religion different from the majority, including expressions of Islamophobia, antisemitism and other similar manifestations of hatred against ethno-religious groups, as well as extreme manifestations of hatred such as incitement to genocide and to terrorism". General Recommendation No. 35 on Combating racist hate speech, CERD/C/GC/35, 26 September 2013, para. 6.*

Antigypsyism [6]

is a specific form of racism against the Roma/Gypsy population and is defined as "an ideology founded on racial superiority, a form of dehumanisation and institutional racism nurtured by historical discrimination, which is expressed, among others, by violence, hate speech, exploitation and the most blatant kind of discrimination"[7].

6 ECRI GPR No. 13 on combating anti-Gypsyism and discrimination against Roma adopted on 24 June 2011 p 4.

7 "The European Commission's High Level Group on Combating Racism and Xenophobia added anti-Gypsyism as a category" in the monitoring of hate speech in the EU
https://www.gitanos.org/actualidad/archivo/125684.html

5 International Convention on the Elimination of all Forms of Racial Discrimination adopted and open for signing and ratification by the General Assembly through Resolution 2106 A (XX) of 21 December 1965 and in force since 21 December 1969. Available at **https://www.ohchr.org/SP/ProfessionalInterest/Pages/CERD.aspx**

Xenophobia

Prejudice against or hatred or fear of people from other countries or cultures. Sentiment, attitude or behaviour of hostility, rejection or hatred towards foreigners or those perceived as such. It is an ethnocentric prejudice towards the culture, values and traditions of foreign nationals, its manifestations ranging from rejection, contempt and threats to segregation, deprivation of rights, assault and murder. In the framework of cooperation, point 15 of the recent UN Global Compact for Safe, Orderly and Regular Migration of 8 December 2018[8] asserts that the Global Compact is people-centred with an important human dimension, and paragraph f) states that "The Global Compact is based on international human rights law and upholds the principles of non-regression and non-discrimination. By implementing the Global Compact, we ensure effective respect for and protection and fulfilment of the human rights of all migrants, regardless of their migration status, across all stages of the migration cycle. We also reaffirm the commitment to eliminate all forms of discrimination, including racism, xenophobia and intolerance". Point 16 of this Global Compact establishes that the New York Declaration for Refugees and Migrants approved a political declaration and a set of commitments which set up the framework of cooperation on which the Global Compact is based consisting of 23 objectives with their implementing measures. Among those objectives, number 17 establishes: "Eliminate all forms of discrimination and promote evidence-based public discourse to shape perceptions of migration".

---

8 Intergovernmental Conference entrusted with approving the Global Compact for Safe, Orderly and Regular Migration. Marrakesh (Morocco), 10-11 December 2018. Outcome document of the conference. A/CONF.231/3 Available at **https://undocs.org/es/A/CONF.231/3**

CONCEPTUAL FRAMEWORK AND GENDER APPROACH

Antisemitism

On 26 May 2016, 31 member countries of the International Holocaust Remembrance Alliance, IHRA[9], adopted the following definition of antisemitism:

"Antisemitism is a certain perception of Jews, which may be expressed as hatred toward Jews. Rhetorical and physical manifestations of antisemitism are directed toward Jewish or non-Jewish individuals and/or their property, toward Jewish community institutions and religious facilities".[10]

The EU "calls on member states that have not yet done so to endorse the non-binding working definition of "antisemitism" used by the International Holocaust Remembrance Alliance (IHRA) as a useful guidance tool in education and training. It can also help law enforcement authorities to more efficiently and effectively identify and investigate antisemitic attacks and, in that sense, is included in the Resolution of the European Parliament on combating antisemitism.[11]

At the 1214th plenary meeting of its Permanent Council, the OSCE[12] made a similar statement in February 2019, on the occasion of International Holocaust Remembrance Day.

Similarly, in UN report A/74/358 of 23 September 2019 on the "Elimination of all forms of religious intolerance", the Special Rapporteur on freedom of religion or belief identified violence, discrimination and expressions of hostility motivated by antisemitism (including online expressions) as a serious obstacle to the enjoyment of the right to freedom of religion or belief. He also urged States to adopt a human-rights based approach in combating antisemitism and all other forms of religious intolerance. This document addresses antisemitism and characterises it from a religious perspective.[13]

---

9 Convenes governments and experts to reinforce, advance and promote education, remembrance and research around the world on the Holocaust, while upholding the commitments of the 2000 Stockholm Declaration.

10 In the following links you will find contemporary examples described in public life, in the media, in schools, in the workplace and in the religious sphere:
https://www.holocaustremembrance.com/es/resources/working-definitions-charters/definicion-del-antisemitismo-de-la-alianza-internacional

11 European Parliament Resolution of 1 June 2017 on combating antisemitism (2017/2692(RSP)) OJEU 30/8/2018, series C 307/183, p. 184. Available at:
https://www.europarl.europa.eu/doceo/document/TA-8-2017-0243_EN.html

12 OSCE Permanent Council No. 1214 Vienna, EU Declaration on the occasion of Holocaust Remembrance Day, 31 January 2019 p. 3 Available at
https://www.osce.org/files/PCOEJ1214.pdf

13 UN A/74/358, 23 September 2019 'Elimination of all forms of religious intolerance' Available at
https://undocs.org/pdf?symbol=en/A/74/358

Islamophobia

From the etymological and conceptual point of view, this term refers to aversion to Islam, Muslims or Islamic culture.

In the document entitled "Guidelines for educators on countering intolerance and discrimination against Muslims, addressing Islamophobia through education" the OSCE and UNESCO[14] use the general definition "intolerance and discrimination against Muslims" as this is the one most used by intergovernmental organisations, including the OSCE itself, UNESCO and the Council of Europe. There are other terms that refer to intolerance and discrimination of Muslims such as "Islamophobia" and "anti-Muslim racism". The term "Islamophobia", widely used by NGOs and frequently in the media, refers to fear, hatred or prejudice towards Islam and Muslims. The concept of "anti-Muslim racism" places intolerance towards Muslims in the broader framework of racism and implies a racial interpretation of a religious concept. The term emphasises the multidimensional aspect of intolerance towards Muslims, which may be based on factors other than religion. Although these terms are not synonymous and refer to different aspects of the problem, they are often used interchangeably".

The Council of Europe[15] defines Islamophobia as "(…) the fear of or prejudiced viewpoint towards Islam, Muslims and matters pertaining to them. Whether it takes the shape of daily forms of racism and discrimination or more violent forms, Islamophobia is a violation of human rights and a threat to social cohesion".

Islamophobia has also been defined as a feeling and attitude of rejection and hostility towards Islam and, by extension, towards Muslims. The term is recognized by the EU Fundamental Rights Agency (FRA) and was coined by the British organization Runnymede Trust that developed the concept and established eight characteristics that denote Islamophobia: Islam seen as a single monolithic bloc, static and unresponsive to new realities, radically different from other religions and cultures with which it does not share values or influences; inferior to 'Western culture' (barbaric, irrational, primitive, sexist); violent and threatening; political ideology and religion are intimately linked; criticisms made by Islam of the West rejected out of hand; and anti-Muslim hostility accepted as natural and normal.

---

14 OSCE-UNESCO "Guidelines for educators on countering intolerance and discrimination against Muslims, addressing Islamophobia through education". Available at **https://www.osce.org/es/odihr/91301?download=true**

15 Council of Europe, 'Islamophobia and its consequences on Young People, European Youth Centre'. Budapest, 1-6 June 2004, report by Ingrid Ramberg. Available at **https://rm.coe.int/16807037e1**

CONCEPTUAL FRAMEWORK AND GENDER APPROACH

The "Conceptual Definition Report on hate crime"[16], drafted within the framework of the Monitoring Committee of the aforementioned agreement signed by the General Council of the Judiciary, the Office of the Prosecutor General, the Ministry of Justice, the Ministry of the Interior, the Ministry of Education and Vocational Training, the Ministry of Labour, Migration and Social Security, the Ministry of the Presidency, Relations with the Courts and Equality, the Ministry of Culture and Sports and the Centre for Legal Studies, to cooperate at institutional level in the fight against racism, xenophobia, LGBTIphobia and other forms of intolerance, points out that: "In line with the definition under Recommendation R (97) 20 of the Committee of Ministers of the Council of Europe, and the subsequent ECRI definition, **Hate speech crime** is any criminally sanctioned act of speech (i.e. **expression-communication**) that can be considered a hate crime. Hate speech crimes typically provoke dangerous situations (leading to other hate crimes unrelated to speech). In these cases the punishable offence is incitement to hatred and violence, a crime classified under Article 510 of the Spanish Criminal Code (CP). However, other common behaviour related to expressive-communicative speech can also be considered hate speech crimes.

---

16 DÍAZ LÓPEZ, JA. "Conceptual Definition Report on hate crime", March 2018. Ed. Ministry of Employment and Social Security. Available at **https://www.inclusion.gob.es/oberaxe/es/publicaciones/documentos/documento_0125.htm**

This is the case of the crime of libel as an aggravating circumstance under Criminal Code Article 22. (4)ª".

Hate speech refers to public speech expressing hatred or prejudice towards a certain person because of a personal condition, or causing a discriminatory effect against a group characterised by a personal condition. Therefore, hate speech may or may not constitute a criminal offence. Hate speech does not necessarily have to be criminally punishable to be considered as such. In any case, its social impact is always negative insofar as it is directly related to prejudice, stereotypes and discrimination towards particular groups of people.

According to ECRI GPR No. 11, a **hate incident** is any incident which is perceived to be a hate crime by the victim or any other person. These are events that, while possibly qualifying as hate crimes or administrative offences related to a hate crime, are not themselves a crime: either because they do not constitute an offence, because they only constitute an administrative offence, or because a conviction has not yet been handed down for the hate crime in question. Hate speech that cannot be classified as a hate speech crime can be considered a hate incident.

For AL-RE-CO, whose objective is to avail itself of tools that help identify and monitor racist, xenophobic, Islamophobic, antisemitic and anti-Roma hate speech, it is also important to make some considerations about

the **Indicators** that can be used to identify hate speech and to establish criteria and identify the factors that can trigger social alerts with a view to preventing or counteracting hate incidents and crimes.

According to ECRI Recommendation No. 15, a characteristic feature of hate speech is that it may be used to incite others to commit acts of violence, intimidation, hostility or discrimination against those it targets or those on whom it is reasonably expected to have such an effect. **Incitement** means that either there is a clear intention to commit acts of violence, intimidation, hostility or discrimination, or there is an imminent risk that such acts will occur as a result of having used hate speech.

There is intention to incite when the person who uses hate speech unequivocally calls on others to commit acts of hatred, or when such intention can be inferred by the forcefulness of the language used and other noteworthy circumstances such as the speaker's prior behaviour. However, intentionality is not always easy to prove, especially when observations are ostensibly related to purported facts or when coded language is used.

It is also important to note that the Recommendation explicitly excludes from the definition of hate speech, forms of expression such as satire or reports or analyses carried out in an objective manner, which simply

offend, distress or annoy. In so doing, the Recommendation protects the definition adopted by the European Court of Human Rights in accordance with Article 10 of the Human Rights Convention[17]. Nonetheless, we would recall that the European Court also recognises that incitement to hatred can be the result of irresponsible insult, ridicule or defamation of certain population groups, the consequence of which may be unnecessary offence, defence of discrimination, the use of degrading or humiliating language, or it may include the inevitable exposure of the victim[18] to all these forms included in the Recommendation's definition.

Moreover, in determining whether there is indeed a risk of these acts coming to fruition, the **specific circumstances** in which hate speech is used must be taken into account.

---

[17] See, for example, Jersild v. Denmark [GC], No. 15890/89 of 23 September 1994; Sürek and Özdemir v. Turkey [GC], No. 23927/94, 8 July 1999, Giniewski v. France, No. 64016/00 of 31 January 2006; Alves da Silva v. Portugal, No. 41665/07 of 20 October 2009; and Fáber v. Hungary, No. 40721/06 of 24 July 2012.

[18] See, for example, Féret v. Belgium, No. 15615/07 of 16 July 2007 and Vejdeland and Others v. Sweden, No. 1813/07 of 09 February 2012.

The following should specifically be considered:

1) The context in which the hate speech concerned is being used (notably whether or not there are already serious tensions within society to which this hate speech is linked);

2) The ability of the person using the hate speech to exert influence over others (by virtue of being a political, religious or community leader, for instance);

3) The nature and strength of the language used (such as whether it is provocative and direct, involves the use of misinformation, negative stereotyping and stigmatisation or is otherwise capable of inciting acts of violence, intimidation, hostility or discrimination);

4) The context of the specific remarks (whether they are an isolated occurrence or are repeatedly reasserted and whether they can be regarded as being counter-balanced either by other remarks made by the same speaker or by someone else, especially in the course of a debate);

5) The medium used (whether it is liable to bring about an immediate response from the audience such as at a "live" event);

6) The nature of the audience (whether it has the means and/or is likely or apt to engage in acts of violence, intimidation, hostility or discrimination).

ECRI GPR No. 15

All indicators are important when it comes to identifying and monitoring hate speech. The organization "Movement Against Intolerance" has highlighted the importance of knowing the meaning and understanding the message of racist and neo-Nazi symbolism currently used by those who seek to spread hatred or violence and to that end search for references to honour, discipline, courage, spirituality, love of family, homage paid to leaders and glorification of war. Familiarity with symbols such as "88" (Heil Hitler), "14NS" (14 words of a National Socialist) and RA-HOWA (racial holy war), for example, is essential in identifying the nature of the crime and its authors.[19]

The Symbology Manual, compiled by the State Commission against violence, racism, xenophobia and intolerance in sports, is an open document that can be updated which includes a wide range of symbols, emblems and flags used by different groups in Spain and in other European countries, the display of which can incite violence, racism, xenophobia or intolerance. The Manual basically serves as a guide to police and security forces who patrol sports venues, and to the private security staff of sports clubs to identify fans who individually or as a group carry such emblems, banners or flags in stadium stands and in the vicinity of sports facilities.

19 Analysis Notebook No. 60 "Early warning of hate crimes". Movement against Intolerance. **http://www.educatolerancia.com/wp-content/uploads/2017/11/Cuaderno-de-analisis-60.pdf**

CONCEPTUAL FRAMEWORK AND GENDER APPROACH

3

The gender approach

PROTOCOL AND SYSTEM OF INDICATORS
to detect hate speech on social media

3. The gender approach

3.1. The gender perspective and intersectionality

Article 3 of the Council of Europe's Convention on preventing and combating violence against women and domestic violence[20] includes the following definitions as regards gender:

a) "Violence against women" is understood as a violation of human rights and a form of discrimination against women and shall mean all acts of gender-based violence that result in, or are likely to result in, physical, sexual, psychological or economic harm or suffering to women, including threats of such acts, coercion or arbitrary deprivation of liberty, whether occurring in public or in private life."

b) "Domestic violence" shall mean all acts of physical, sexual, psychological or economic violence that occur within the family or domestic unit or between former or current spouses or partners, whether or not the perpetrator shares or has shared the same residence with the victim."

c) "Gender" shall mean the socially constructed roles, behaviours, activities and attributes that a given society considers appropriate for women and men."

The mandate regarding gender equality and the empowerment of women has been agreed by the Member States and encompasses all areas of peace, development and human rights. Gender equality mandates are based on the United Nations Charter which unequivocally reaffirms equal rights between women and men.

---

20 The Council of Europe Convention on preventing and combating domestic violence, done at Istanbul on 11 January 2011, was ratified by Spain and published in the Official State Gazette (BOE) on 6 June 2014. Available at **https://www.boe.es/boe/dias/2014/06/06/pdfs/BOE-A-2014-5947.pdf**

The **Fourth World Conference on Women** held in 1995 defended the incorporation of a gender perspective as a fundamental and strategic approach to achieving commitments on gender equality and the resulting **Beijing Declaration and Platform for Action** call for action in this regard. Additional commitments are included in the final document of the Twenty-third Special Session of the General Assembly, the Millennium Declaration and various resolutions and decisions of the United Nations General Assembly, the Security Council, the Economic and Social Council and the Commission on the Status of Women.

The **ECOSOC conclusions of 1997** defined the incorporation of a gender perspective as: "…the process of assessing the implications for women and men of any planned action, including legislation, policies or programmes, in all areas and at all levels. It is a strategy for making women's as well as men's concerns and experiences an integral dimension of the design, implementation, monitoring and evaluation of policies and programmes in all political, economic and societal spheres so that women and men benefit equally and inequality is not perpetuated. The ultimate goal is to achieve [substantive] gender equality".

Gender is a central social operator in the configuration of social hierarchies, but it does not operate in a vacuum but rather in combination with other social operators, mainly social class and ethnicity, and intersectionality is the conceptual tool that enables us to visualise the interception of different social operators (gender, class, ethnicity, etc.). This approach interrelates different categories that contribute to forming socially, politically, economically, culturally and psychologically constructed identity, giving rise to differentiated positions between people in society (Helia del Rosario Rodríguez, Right to a life free from violence. Experiences and resistance of migrant women: case studies).

3.2. The gender approach in the ALRECO project protocol and indicators

The principle of equal treatment and non-discrimination between men and women is an essential factor in the ALRECO project as women constitute one of the most vulnerable groups and are therefore most at risk of being the target of racist or xenophobic incidents and speech.

Women facing discrimination based on racial or ethnic origin suffer dual discrimination based not only on gender but also on ethnic origin or other forms of intolerance. The disadvantages that minority women face in relation to the labour market, trafficking and racially based violence are of particular concern. Many women face this multiple inequality due to their gender and the fact that they belong to a certain ethnic or racial group, making it more difficult for them to fulfil their needs. According to the United Nations, for many women factors related to their social identity such as race, colour, ethnic origin and national origin, can create problems that affect particular groups of women, or that affect some women disproportionately compared to others, and to their ability to meet their needs.

In recent years, racist and xenophobic hate speech has often attempted to link immigration with gender-based violence. Associating the mistreatment of women, violence and sexism with immigrants aims to stigmatize certain groups. Recent reports warn of an increase in hatred with serious attacks on Muslim women resulting in a phenomenon that has been called Gender Islamophobia.[21]

Another example of hate speech and intolerance that disparages women

---

[21] Citizens' Platform against Islamophobia. Available at **http://www.observatorioislamofobia.org/que-es-la-islamofobia/**

is the term "feminazi" which seeks to justify discrimination and violence against women mainly through moral judgments of their lifestyle, dress, ideology, and so forth. It is a form of hate speech that reinforces misogyny and intolerance and justifies verbal and physical violence against women.

In this connection, to detect hate speech in social media, the ALRECO project's protocol and system of indicators includes those that take account of the gender factor as an identifying and/or aggravating element of hate speech, and mainly focuses alerts on the detection of:

- Racism/ Gender xenophobia.
- Linking migration to gender violence.
- Gender Islamophobia.

3.3. The human rights-based approach

According to the Universal Declaration of Human Rights, adopted by the General Assembly of the United Nations in 1948, the universality of the equal and inalienable rights of all human beings is the foundation of freedom, justice and peace in the world, based on personal dignity and freedom. The priority placed on human rights principles was the cornerstone of the United Nations reform initiatives that commenced in 1997.

The human rights-based approach focuses on vulnerable people and population groups that are subject to increased marginalization, exclusion and discrimination and who suffer from higher levels of intolerance, hate speech and hate crime. This approach often requires an analysis of gender norms and the different forms of discrimination and power imbalances in order to ensure that interventions reach the most oppressed and segregated segments of the population.

As previously mentioned, the legal classification of what has come to be referred to as hate crime, affects all criminal offences subject to the aggravating circumstance described in Article 22(4) of the Spanish Criminal Code, and together with so-called hate speech, affects the provisions of Chapter IV: **Crimes relating to the exercise of fundamental rights and public freedoms**, i.e. crimes against the exercise of fundamental rights and public freedoms guaranteed by the Constitution, and crimes against freedom of conscience, religious sentiments and respect for the deceased. This classification corresponds to the protection of the Fundamental Rights of Spaniards and foreigners and of their public freedoms under the Spanish Constitution, particularly under Article 10: 1. The dignity of the person, the person's inherent inviolable rights, the free development of the personality, the respect for the law and for the rights of others are the foundation of political order and social peace. 2. Provisions relating to the fundamental rights and liberties recognized by the Constitution shall be construed in conformity with the Universal Declaration of Human Rights and international treaties and agreements thereon ratified by Spain.

Hence the following, inter-alia, are considered best practices under a human rights-based approach:

- activities that see the full exercise of human rights as their ultimate goal;
- participation of people both as a means and an objective;
- strategies that empower and contribute to autonomy rather than deny it;
- situational analysis used to identify the immediate, underlying and fundamental causes of human rights violations, discrimination and hatred based on intolerance.

The analysis includes all stakeholders, including the State as the main guarantor of rights and other non-state agents such as NGOs and others. In this sense, a dual need should be highlighted: first, to inform individuals, communities and social groups of their rights, and secondly, to protect them from hate speech and hate crime.

METHODOLOGY

4

Methodology

PROTOCOL AND
SYSTEM OF INDICATORS
to detect hate speech on
social media

4. Methodology

The ALRECO project has developed indicators to detect online hate speech. It aims to develop an action protocol with a system of indicators and search criteria to identify speech that promotes online racism, xenophobia and hatred. The system also includes early warning indicators with which to assess the intensity, severity, dissemination and potential impact of hate speech, the objective being to devise initiatives to prevent potential discriminatory incidents and hate crimes.

Within the framework of this action, the following is considered a preliminary step: identify experiences and best practices in the European Union that serve as a basis for a discussion and classification of early warning indicators.

We will now summarize the methodology used to select best practices and to develop the protocol and indicator system.

4.1. Best practice selection methodology

It goes without saying that best practice is not only defined as good on its own merit, but rather a practice that has been shown to work well and produce good results and is therefore recommended as a model. It is a successful experience, which has been tested and validated, in a broad sense, which has been repeated and which deserves to be shared in order to be adopted by as many people as possible. It must meet at least 5 criteria:

Figure 1
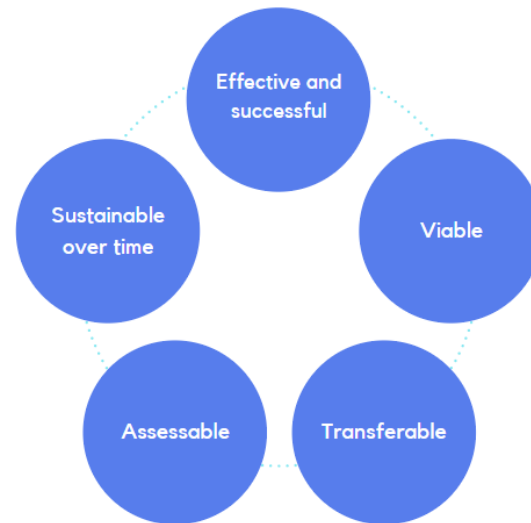
BEST PRACTICE

An action, methodology or tool developed in Europe in the field of online hate speech, which has proven itself able to effectuate change with positive results in the identification, analysis, monitoring and/or evaluation of online hate speech for racist, xenophobic, Islamophobic, antisemitic and anti-Gypsy purposes.

Specifically, value is placed on the impact of the experience, the fact that it has been implemented and has produced practical results and has evaluation mechanisms, that experiences are heterogeneous in terms of the agent promoting the practice and the beneficiaries of the experience (institutional, university, third sector-promoted, etc.) and that it is endowed with coordination mechanisms.

A definition of best practice in relation to tools and experiences in the field of online hate speech was agreed between the ALRECO project partners taking these considerations into account.

Second, different methodological tools were used to identify best practices:

Table 1

TOOLS	INSTRUMENTS
DOCUMENTARY ANALYSIS	Main documents: • Similar projects • Existing tools • Previous experiences • Academic articles and ongoing research • Platform analysis • Analysis of relevant press clippings
INTERVIEWS OF KEY PERSONS	• Information was requested from all project partners: OBERAXE, Ministry of the Interior, University of Barcelona (CREA research group), and the Asociación TRABE. Key entities and individuals from different European countries (Finland, the Netherlands, Austria, Italy, the United Kingdom and Greece) were contacted. • Experts were contacted: FRA Advisor • Social organisations were contacted: OXFAM
INFORMATION SHEET - QUESTIONNAIRE	• A sheet was drawn up to systematize the information used to assess best practices.

Third, different key people and entities were contacted to broaden the scope in the search for best practice:

Table 2

EXPERT / ENTITY CONTACTED	COUNTRY
European Training and Research Centre for Humans Rights and Democracy (ETC)	Austria
Centre for European Constitutional Law (CECL)	Greece
University of Milan	Italy
Ministry of Justice of Finland - Anti-discrimination and Fundamental Rights Team	Finland
Bradford Hate Crime Alliance	United Kingdom
Department of European and International Affairs / City of Utrecht	Netherlands
Rosa Bada, member FRA Advisory Board	European Expert
Jose Camacho-Collados, Cardiff University	United Kingdom
Juan Carlos Pereira Kohatsu	University Carlos III Spain Master's Thesis
Spanish Racism and Xenophobia Observatory (OBERAXE) (State Secretariat for Migration - Ministry of Labour, Migration and Social Security).	Spain
National Office for the Fight against Hate Crime (Secretariat of State for Security, Ministry of the Interior)	Spain
University of Barcelona (CREA research group)	Spain
Asociación TRABE	Spain

Thanks to the methodology used, information was obtained from a total of 53 experiences in the period between 9 and 30 January 2019. Of these, 18 met the established criteria and were selected. Regarding the areas of the selected experiences, it is interesting to note that most of them combine different aspects. In other words, there are hardly any "pure" tools for detecting hate speech but are rather part tool, part counter-narrative, part awareness-raising and part training. Of the 18 experiences selected, only two can be considered a "pure" tool.

Regarding the geographical areas to which the selected experiences refer, although the scope of the report is European, a tool from the United States was included because it was especially relevant for the purpose of the ALRECO project and because it is difficult to circumscribe geographical areas of impact. For example, another of the selected tools is limited to "Spanish-speaking" countries which also transcends Europe. The rest of the experiences are from different countries as they are promoted in consortia by several partners. Not only the country where the experience is promoted but also its language is an issue to take into account. For instance, we do not have any tools from Arabic speaking countries but we do work in that language. In short, we have gathered experiences from the following countries: Spain, France, Italy, United Kingdom, Greece, Germany and the United States.

Regarding the groups targeted by the selected experiences, most focus on migrants or ethnic minorities and different religions (mainly Muslim and Jewish). Some combine different criteria or are aimed at vulnerable groups (LGTBI, etc.).

There are basically three types of sponsors of the selected experiences: universities/academia, public institutions and NGOs. Most of the experiences were publicly financed although some were financed by companies such as Google or Facebook.

4.2. Methodology used to develop the protocol and system of indicators

A total of 16 tools were selected from the review included in the first best practices report and the review of specific articles taking an in-depth look at existing hate speech monitoring tools. This selection enabled us to thoroughly study the methodology. The following table shows the degree of relevance of each of the best practices, specifically its methodology, with the ultimate aim of creating a proprietary tool (Table 3).

Table 3

Tools to identify online hate speech

TOOL	RELEVANCE OF THE METHODOLOGY FOR **ALRECO**	OTHER COMMENTS
OBERAXE Classifier	High	Classification typology.
Somos más (we are more) (Google project)	Low	It lacks a monitoring tool. Project focused on awareness-raising, intervention and promoting cyber-activism to counter hate speech and radicalism. Especially focused on YouTube.
CiberHache	Average	It lacks a monitoring tool. Capture of tweets from hashtags- trending topics. The data mining analysis technique used is interesting. Analysis by expert peers and contrast using the Kappa test.

TOOL	RELEVANCE OF THE METHODOLOGY FOR **ALRECO**	OTHER COMMENTS
CibeRespect	Average	Project focused on intervention and promoting cyber-activism to combat hate speech. While hate speech is monitored, it lacks a monitoring tool.
PROXI Observatory	High	Monitoring of news on the migrant and Roma populations from three digital media outlets with a large audience in Spain. The analysis team is not currently operational. Software available in open access. Cyber activism proposal.
Be the key	Low	Project focused on intervention and promoting cyber-activism. Especially focused on Islamophobia. The main dissemination platform is Facebook.
Official rewind	Low	Project focused on intervention to promote counter-narratives through Twitter and Facebook.
Getthetrollsout	Low	Qualitative monitoring of social networks and media. Intervention-oriented project. Special attention to hate speech against religious minorities.
MANDOLA	High	Creation of an app to report online hate speech based on an algorithm to measure hate speech. It includes a comparative map showing hate speech per country.
Wordsarestones	Low	Project focused on intervention through training in detecting hate speech and promoting cyberactivism.

TOOL	RELEVANCE OF THE METHODOLOGY FOR **ALRECO**	OTHER COMMENTS
Save a hater	Low	Project focused on intervention to create counter-narratives through cyberactivism.
Hate meter	Average	Project in development that aims to build a tool focused on Islamophobia.
Silence hate	Low	Project focused on intervention
Contra l'odio	High	Includes a map with data on the type and frequency of hate speech in each area of Italy. It is based on a Twitter monitoring tool using an algorithm fostering machine learning and natural language processing for automatic identification in social media. It creates Twitter account "scores" for the purpose of identifying the tendency to use hate speech or follow people who do.
Hate base	High	Linguistic analysis of public conversations to calculate the likelihood they are hate-related. Based on natural language engine (Hate brain) algorithms. Uses a large vocabulary
Donate the hate	Low	Project focused on training, awareness-raising and intervention against hate speech.

TOOL	RELEVANCE OF THE METHODOLOGY FOR **ALRECO**	OTHER COMMENTS
Exploring Online Hate	Average	Analysis of specific users who produce hate speech. Identification of most frequently used hashtags and terms. Data mining and algorithm creation but not public information.
Saferlab		Uses qualitative and natural language methods to identify violence and hate speech.
Smartphone application (App) called "LIGHT ON RACISM"	Low	Promotes cyberactivism against hate speech. The tool encourages people to report offences and promote counter-narratives.

The tool website itself was carefully reviewed as were articles and documents about it. The following aspects were analysed for each tool:

- ◗ Tool methodology (how it works)
- ◗ Types of language specified in the tool (categories of hate speech, if any. Possible category indicators.)
- ◗ Types of objectives / purpose of the speech (categories established by the tool regarding the objectives or purposes of hate speech, if any. Possible category indicators.)
- ◗ Types of hate speech intensity (categories established by the tool regarding intensity, if any. Possible category indicators.)

To supplement this information, we contacted the groups responsible for some of the tools, specifically PROXI developed by the IDHC (Institut de Drets Humans de Catalunya) and Cyberhache.

We also consulted scientific articles through different scientific databases such as the ISI Web of Science search engine for the latest methodologies, developments and challenges addressed.

We used the following keywords in our search:

Web of Science Keywords (last 10 years)
Hate Speech and Social Media
Hate Speech and Social Networks
Hate Speech Detection
Cyberhate and Social Networks.

Then we used the following terms to narrow down the search: Hate Speech and Racism, Hate Speech and Xenophobia, Hate Speech and Immigration, Hate Speech and Islamophobia, Hate Speech and Muslims, Hate Speech and Antisemitism, Hate Speech and Antisionism.

We selected 18 useful articles from the analysis for the development of our tool and gleaned information from them on the following aspects:

- ❥ Types of hate speech analysis.
- ❥ Methodological aspects (description and classification of methodologies, review of different methodologies' pros and cons, etc.).
- ❥ Specific tools developed. (Here, objective and scope of the initiative, impact achieved and evaluation).
- ❥ Hate speech classification (semantics, content classification, intensity, others).
- ❥ Contributions related to the different areas of AL-RE-CO (racism, xenophobia, antisemitism, Islamophobia) and cross-cutting gender examination.
- ❥ Other aspects of interest.

The following conclusions can be drawn from this methodological process:

- ❥ Several different experiences, both tools and articles, proved to be very useful in developing the tool envisaged in the ALRECO project.
- ❥ It is important to be aware of the intrinsic limitations of algorithm-based tools. The experiences and articles analysed make special mention of these limitations.
- ❥ It is important to delimit the scope of different aspects of the tool as much as possible: search terms, language in which it will be developed (and geographical areas), specific areas (racism, xenophobia, antisemitism, Islamophobia and anti-Gypsyism) and a cross-cutting gender perspective).

BAG OF WORDS AND INDICATORS

BAG OF WORDS AND INDICATORS

5

Point of departure

PROTOCOL AND SYSTEM OF INDICATORS
**to detect hate speech on social media**

5. Point of departure

The protocol proposal is based on the experience gained by team members through the creation of social media speech analysis methodologies (Cabré–Olivé, Flecha–García, Ionescu, Pulido & Sordé–Martí, 2017; Pulido, Redondo–Sama, Sordé–Martí & Arrow, 2018) and on the theoretical framework of hate speech, specialized articles on automatic online detection of hate speech, and exploration of best practices.

The main theoretical contributions on discourse analysis and racism (Van Dijk, 2002, 2007; Wieviorka, 1992) and communication (Austin, 1961; Searle & Soler, 2004), highlight the complexity of racism and hate speech and how these are very closely linked to context and non-verbal elements. Online hate speech research is still scarce (Fortuna and Nunes, 2018). The articles consulted concur in pointing out the difficulties in identifying and classifying online hate speech (Hughey & Daniels, 2013), specifically: the importance of context and emitter is in determining whether a word or expression can be considered hate; non-explicit expressions of racism (subtle references, metaphors, etc.); discourse specificity according to the target group; intersectionality; volume of data.

Most of the experiences analysed use a set or "bag of words" to detect hate speech (Gitari, Zuping, Damien & Long, 2015; Greevy & Smeaton, 2004). As there are obvious limitations, often combinations of different words or parts of text are used. Many researchers are trying to improve automatic hate speech detection through machine learning, that is, through algorithms that improve over time or can be trained to detect hate speech. Step one of machine learning entails collecting tweets (using different criteria) by means of a list of pre-defined keywords (Burnap & Williams, 2015, 2016). Step two consists of training the algorithm, which is typically done with text coders.

The protocol that we are presenting is based on the following diagram:
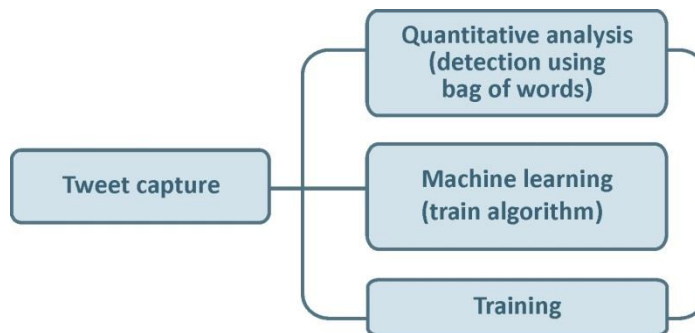
We now turn our attention to capture and content analysis strategies.

BAG OF WORDS AND INDICATORS

6

Tweet capture

6. Tweet capture

In order to attempt to limit hate speech monitoring to Spain, as there is no way to tell what country a tweet comes from, tweets in Spanish will be captured and the time difference between Spain and Latin America will help to differentiate between those originating in one or the other of these two geographical areas.

Given the volume of daily tweets, three tweet capture strategies were defined to obtain manageable samples with which to analyse hate speech.

6.1. Bottom Up: Tweet capture by daily trending topics

The Bottom Up strategy aims to help identify hate speech through trending topics on Twitter and their possible relationship to events or other factors that encourage it.

The Bottom Up capture process consists of:

1. Programming the automatic collection of the day's top 50 trending topics (Spain).
2. Once the list of 50 trending topics is obtained, all tweets published under each of them are collected, resulting in 50 complete tweet lists of each one of the topics.

This process of capturing and analysing the day's trending topics and obtaining the tweets corresponding to each of them will be performed every day at the same time and will be dated.

Example of the tweet capture on 17 April 2019:

#DiAlgoBonitoAUnaEscritora	Tour de la Manada	PSOE y Compromís	#STOPOKUPAS	#DebateTVE
#FelizMiércoles	España y Turquía	Feliz Miércoles Santo	#AR17A	#GabrielGarciaMarquez
#DebateESP	William Carvalho	El Consejo de Europa	#DiaMundialdelaHemofilia	#DíaMundialHemofilia
#MiércolesSanto	Arrimadas a Rufián	Premio Nobel de Literatura	#LaCafeteraJEcJEc	#17Abril
#ApodérateUnidasPodemos	Televisión Española	Open Arms	#RumboUrnasARV	#CasoGrúas
Atresmedia	Gabo	Damian Lillard	#STRP	#WednesdayMotivation
Gabriel García Márquez	cartaya	Junta Electoral a Vox	#Homecoming	#DíadelaInformaciónJuvenil
RTVE	El Hijo	McCollum y Lillard	#NoEsNo	#SuperSmashBrosUltimate
Manuel Alcántara	Mt 26	Pedro Saura	#BeyoncéHomecoming	#SoloSíEsSí
Chavela Vargas	Juanma Lillo	#YoigoTeDaUnPlus	#hemofilia	#HdadNervión19

In this case we capture *tweets* that explicitly target the communities included in the project, the aim being to determine whether they constitute hate speech and, if so, which type of hate speech. We start with keywords typically used to refer to these communities.

Table 4 shows the initial list of words used for tweet capture. All the tweets that contain the words listed in the table are captured. It is important to point out that the *tweets* containing these words cannot necessarily be considered hate speech. This is nothing more than a list designed to capture *tweets* related to the groups included in our analysis.

6.2. Top Down: Tweet capture by group

Table 4
Tweet-capturing keywords (provisional)

Group	Keyword
Migrants and foreigners	migracion inmigracion migrante migrantes inmigrante inmigrantes extranjero extranjera extranjeros extranjeras re- fugiado refugiada refugiados refugiadas marroqui marroquis magrebi magrebis africano africana africanos africanas latino latina latinos latinas rumano rumana rumanos rumanas subsahariano subsahariana subsaharianos subsaharia- nas sudamericano sudamericana sudamericanos sudamericanas sudaca sudacas chino china chinos chinas chinito chinita clandestino clandestinos mena menas
People from cultural, ethnic or religious minorities	negro negra negros negras* panchito panchita panchitos panchitas indio india indios indias conguito conguita conguitos conguitas machupichu machupichus
Roma	gitano gitana gitanos romani romanis patriarca
Muslims	musulman musulmana musulmanes musulmanas moro mora moros moras islamico islam arabe arabes sarraceno sarracenos yijadista yijadistas islamista islamistas fundamentalista fundamentalistas fundamentalismo hiyab burca
Jews	judio judia judios judias sion sionista sionistas hebreo israeli israelis holocuento holocausto hitler camara+gas nazi
Global	stopinvasion mantero manteros topmanta eurabia terrorismo atentado atentados guerra sumision boicot frontera fronteras expulsion

This is a provisional word list which will be improved over time. Tweet capture by means of predefined keywords will be done weekly. If the volume of data is excessive, the capture will be done on a daily basis.

6.3. Top Down: Tweet capture using specific criteria

Another strategy entails personalizing the capture by user, hashtag, topic or event. This strategy is used when hate speech is detected or is expected based on an event, for instance a terrorist attack, a public person, a news item affecting society. These capture routines are initiated manually based on need or opportunity depending on events.

BAG OF WORDS AND INDICATORS

44

BAG OF WORDS AND INDICATORS

7

Hate speech analysis

PROTOCOL AND
SYSTEM OF INDICATORS
to detect hate speech on social media

7. Hate speech analysis

Three types of hate speech analysis will be conducted for each tweet sample obtained.

7.1. Quantitative analysis by bag of words

This strategy enables the classification of the tweets into hate or non-hate speech depending on the content of words identified as hate speech (bag of words). We propose our own bag of words with two categories of terms (to which the Spanish language plural and feminine forms of some must be added).

This bag of words was the fruit of exploring five existing bags as well as a discussion group, a quantitative analysis of the lexicon of 20 racist profiles on Twitter and supplementary qualitative analyses of tweets (more than 1000) identified as hate speech. The methodology followed is detailed in Annex II.

The bag of words is necessarily limited as it is unable to capture the wide array of terms used in the myriad of hate expressions. While imperfect, it is a point of departure for designing and managing the hate speech monitoring tool. Hence, it must be a flexible bag of words open to modifications (additions, nuances, eliminations) during the pilot phase. Once in operation, we would like to provide open access to the bag of words allowing citizens to contribute new words and comments. This would give citizens the right to participate not only in identifying hate speech, but also in the tool itself.

7.2. Quantitative analysis and machine learning

This strategy aims to identify patterns and relationships between the words in the text in such a way that the tweet content can be classified according to hate intensity (ranging from extreme hatred to upstander speech).

Supervised machine learning techniques will be used for the algorithm implementation and training. This requires the following sequence events:

1. **Clean up data**: Symbols and characters that do not add meaning to tweets are eliminated, and words are separated and standardized to focus on content as opposed to form (convert lowercase characters, abbreviations, etc.).

2. **Tag tweets for training**: The content of the set of tweets will be analysed manually. Tweet reviewers are required to tag tweets according to an intensity scale:

Extreme hatred	Speech that incites violence.
Offensive hatred	Speech that is individually or collectively offensive that incites discrimination and propagates clichés and falsehoods.
Neutral speech	Descriptive speech void of hatred.
Upstander speech	Alternative discourse which contributes to a counter-narrative, defying clichés or defending groups subjected to hatred.

Created in-house from external sources: Watanabe, Bouazizi & Ohtsuki, 2018; Gitari, Zuping, Damien & Long, 2015; Hate Base

Reviewers receive a codebook (with definitions) in advance. In the training process, we will look into the possibility of examining the degree of agreement when assigning categories (for example, the Kappa coefficient of agreement).

3. **Obtain the model**: The manually tagged tweets will be used to create a model enabling the algorithm to classify tweets according to the defined hate speech intensity scale. This process requires adjusting different models and selecting the one that responds best to the desired classification.

4. **Classify or predict**: Lastly, the predictive model will be applied to unprocessed tweets to classify them automatically.

The different phases of the analysis algorithm implementation process entail repeated adjustments based on analysis of the results obtained and the fine-tuning of selection parameters.

The resulting model will be updated and adjusted on a regular basis by incorporating new sets of pre-classified hate speech tweets.

7.3. Manual qualitative analysis

In addition to the aforementioned quantitative analyses, the tool allows for a manual (not automated) qualitative tweet analysis. This is a voluntary option for users who want more in-depth analysis that algorithms and automation do not provide.

This is explained in a guideline document with examples focusing on different characteristics: types of speech, purpose of speech and linguistic registers.

Types of language used in hate speech

- ◗ Insulting and humiliating language that incites or advocates violence
- ◗ Justification, jokes, trivialization of violence towards 'others'
- ◗ Divisive and otherness (them versus us) language
- ◗ Stereotype-Prejudice
- ◗ Rumours
- ◗ False Facts
- ◗ Flawed argumentation
- ◗ Metaphor; dehumanizing comments; irony

References: Noriega & Iribarren, 2012; Observatorio Proxi, 2015; Van Dijk, 2002, Cyberhache, Control'Odio

Types of 'otherness'

As scientific literature points out, hate speech is characterized by marked polarization between 'Us' and 'Them', where negative qualities are attributed to 'them' as opposed to positive qualities attributed to 'us'. The message may be addressed directly to 'them' or could refer to 'them' in the third person.

◗ Language addressing 'them'
◗ Language about 'them'

Different objectives of hate speech

◗   To offend
◗   To offend an individual person
◗   To offend a group
◗   To incite violence
◗   To incite discrimination
◗   To incite segregation
References: Miró Llinares, 2016; Cyberhache, Proxi, Miró.

Types of linguistic registers

◗   Derogatory non-verbal structures. e.g. emojis, punctuation, capital letters, etc.
◗   Syntax. e.g. to emphasize or de-emphasize an act using active as opposed to passive voice
◗   Lexicon. e.g. terrorist versus freedom fighter
◗   Local meaning of an expression
◗   Vague or indirect meaning of 'our' action versus detailed meaning about 'their' inappropriate behaviour
◗   Global meaning of speech. e.g. positive topics addressed for us (solidarity, tolerance, etc.) versus negative topics for them (crime, violence, etc.)
◗   Diagrams. e.g. simplification
◗   Rhetoric, metaphor, metonymy, hyperbole, euphemism, etc.
◗   Interaction. Interrupt, finish earlier, aggressively disagree, failure to respond
References: (Van Dijk, 2002, 2007)

BAG OF WORDS AND INDICATORS

Types of Upstander speech

Upstander speech is characterized by proactive condemnation and/or positive speech in reference to the group targeted by hate speech. Contents can be classified as follows:

- ◗ Publication of one's own comments: argumentation with evidence that negates prejudices; drawing attention to positive experiences.
- ◗ Anti-rumour interventions: counter-rumour data and explanations.
- ◗ Pedagogical interventions: facts and data to shed light on the debate.
- ◗ Awareness-raising interventions: emotional appeal.
- ◗ Qualification of comments from other users: support or disqualify positive / negative comments.
- ◗ Reporting hateful comments: through social media tools and apps created specifically for that purpose.

References: Observatorio Proxi, 2015

These definitions and categories are a starting point and will be modified over the tool's development and piloting phases according to the results obtained.

In addition to being used for qualitative analysis, the tool should provide indications for actively intervening in "counter-narrative".

BAG OF WORDS AND INDICATORS

8

Indicator systems

to detect hate speech on social media

8. Indicator system

We draw a distinction between the indicators that enable us to identify a *tweet* as hate speech (and its different intensities) and the hate speech indicators that will be able to be obtained from implementing the hate speech monitoring tool.

Hate speech indicators in social media

The following are hate speech indicators that will be used in the development of the tool for algorithm's piloting and training phase (to ensure tweets are correctly identified as hate speech and properly labelled according to their intensity).

Types of hate speech

- Presence of certain words (bag of words)
- Language that incites or advocates violence
- Justification, jokes, trivialization of violence towards 'others'
- Divisive and otherness (them versus us) language
- Reproduction of stereotypes or prejudices
- Spread of rumours
- Spread of false facts
- Flawed Argumentation
- Metaphors, dehumanizing comments, ironies
- Derogatory non-verbal structures e.g. emojis, punctuation, capital letters, etc.

These characteristics will then be used to draw a distinction between extreme and offensive hate speech.

- **Extreme**: incitement to violence.
- **Offensive**: Speech that is individually or collectively offensive, incites discrimination and propagates clichés and falsehoods.

Indicators to monitor hate speech

This tool will make it possible to regularly analyse hate speech on social media. A set of indicators will be generated to monitor the evolution of hate speech over time, its intensity and other characteristics.

Table 6

Hate speech indicators

Indicator	Description and variations
No. of tweets expressing hate	Number of tweets with hateful words (bag of words). Applicable to: • Each trending topic in a day • Set of trending topics in a day • Each of the groups analysed during a certain period • Set of tweets related to a specific event / topic
Percentage of tweets containing hate speech	Percentage of tweets with hate words compared to tweets without these words. Applicable to: • Each trending topic in one day • Set of trending topics in one day • Each of the groups analysed during a certain period • Set of tweets related to a specific event / topic
Percentage of hate speech in the day's topics	Percentage of the day's topics featuring tweets with hateful words
Growth of hate speech in tweets targeting groups	Hate speech growth (in percentage terms) between two specific dates. Applicable to: • Total number of tweets featuring hate speech • Percentage of tweets featuring hate speech

Observations: The best possible bag of words is needed to obtain these indicators. As already mentioned, these indicators will not collect all hate speech but only that which uses clearly identifiable keywords expressing hate. While they will not be able to collect hate speech that does not use the words from the bag of words, they do provide useful information on how an explicit type of hate speech develops.

Table 7

Hate speech intensity indicators

Indicator	Description
No. of tweets classified as extreme hate	Number of tweets that incite violence. Applicable to: • Each trending topic in a day • Set of trending topics in a day • Each of the groups analysed during a certain period • Set of tweets related to a specific event / topic
No. of tweets classified as offensive hate	Number of tweets that are individually or collectively offensive, that incite discrimination and propagate clichés and falsehoods. Applicable to: • Each trending topic in a day • Set of trending topics in a day • Each of the groups analysed during a certain period • Set of tweets related to a specific event / topic
No. of tweets classified as neutral	Number of tweets that do not contain hate speech or a counter-narrative. Applicable to: • Each trending topic in a day • Set of trending topics in a day • Each of the groups analysed during a certain period • Set of tweets related to a specific event / topic
No. of tweets classified as upstander	Number of tweets contributing to a counter-narrative, defying clichés or defending groups subjected to hatred. Applicable to: • Each trending topic in a day • Set of trending topics in a day • Each of the groups analysed during a certain period • Set of tweets related to a specific event / topic

Observations:

These indicators can only be obtained by training an algorithm, which entails people tagging tweets. Again, the great complexity of racist discourse makes it difficult to build an effective model to identify varying degrees of hate speech intensity. If the model enables us to obtain some or all of these categories, other indicators such as percentages and trends can be developed over time.

9

Technical specifications

**PROTOCOL AND
SYSTEM OF INDICATORS**
to detect hate speech on
social media

9. Technical specifications

The three data collection strategies will be implemented with an application developed using the Python programming language to perform search and capture tasks.

Users, roles and limitations

A large volume of information needs to be processed before hate speech can be monitored and the results displayed. Computing resources such as data storage, memory and computing power are critical and therefore resource consumption must be controlled at all times. That is why it is important to limit the execution of algorithms in specific time slots and also limit the number of users and the functionalities available to them.

Following is the initial user and role proposal:

Administrator	Configure tweet capture parameters. Configure the hate speech detection algorithm. Configure hate speech indicators and statistics. Configure data display and filter parameters.
Reviewers	Introduce the tweets classified according to the intensity scale into the system. A codebook will be provided for this.
Users	View hate speech according to filters. Perform qualitative analysis. Propose addition of words / provide other information or comments
Technician	Initially to verify that the capture and analysis processes are working properly. This will be activated manually to check that there are no errors in terms of execution or analysis. Once it is determined that they are working correctly, they will run automatically in certain time slots.

BAG OF WORDS AND INDICATORS

Data processing

A lot of calculation time will be needed to view hate speech statistics from a large volume of tweets in real time, which means that the application will be very slow. This problem can be mitigated by pre-processing results such that a file is created for each day with the results of the statistics and indicators of the tweets corresponding to that day. Historical hate speech data will come from these pre-processed files and not from all tweets. This will greatly enhance calculation and data viewing efficiency.

We propose the guidelines for conducting qualitative hate speech analysis and related tools be made available to stakeholders with no prior processing of the results generated.

User interface (viewing data)

A graphic interface will be designed to view the hate speech results enabling users to select and specify the desired results. Basically, the parameters that can be selected will correspond to the identification parameters designed in the project and described in the previous sections.

During the pilot phase, the requirements for displaying hate speech analysis results will be redefined, as will the design of the graphic interface, the aim being to adapt it to the needs of the different application users. The application will display:

1. Hate speech indicators and main statistics (according to the filters established which must be feasible).
2. Table and graph of how hate speech is evolving (number of tweets) over a defined period of time (day/ week/month/year) and selected filters.
3. Table and point cloud graph of frequent words according to the selected filters.
4. An export the list of tweets (.xls) according to selected filters.
5. Automatically prepared predefined reports.

10

References

10. References

Austin, J. (1961). Cómo Hacer cosas con palabras. Palabras y acciones (Paidós). Barcelona.

Burnap, P., & Williams, M. L. (2015). Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeli ng for Policy and Decision Making. Policy and Intenet, 7(2), 223–242. **https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0072-6**

Burnap, P.,& Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data Science, 5(1). **https://doi.org/10.1140/ epjds/s13688-016- 0072-6**

Cabré-Olivé, J., Flecha-García, R., Ionescu, V., Pulido, C., & Sordé-Martí, T. (2017). Identifying the Relevance of Research Goals through Collecting Citizens' Voices on Social Media. International and Multidisciplinary Journal of Social Sciences, 6(1), 70. **http://diposit.ub.edu/dspace/bitstream/2445/126740/1/678905.pdf**

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering, 10(4), 215–230. **https://preventviolentextremism.info/sites/default/files/A%20Lexicon-Based%20Approach%20for%20Hate%20Speech%20Detection.pdf**

Greevy, E., & Smeaton, A. F. (2004). Classifying racist texts using a support vector machine (p. 2).

Association for Computing Machinery (ACM). **https://doi.org/10.1145/1008992.1009074** Hughey, M. W., & Daniels, J. (2013).

Racistcommentsatonlinenewssites: Amethodologicaldilemmafordiscourseanalysis. Media, Cultureand Society,35(3), 332–347. **https://journals.sagepub.com/doi/10.1177/0163443712472089**

Miró Llinares, F.(2016). Taxonomía de la comunicación violenta yeldiscurso delodio en Internet.IDP. Revistade Internet, Derecho y Política, 22(22). **https://dialnet.unirioja.es/servlet/articulo?codigo=5849356**

Noriega, C. A., & Iribarren, F. J. (2012). Social Networks for Hate Speech: Commercial Talk Radio and New Media. UCLA Chicano Studies, 2. Retrieved from **https://www.chicano.ucla.edu/publications/report-brief/social-networks-hate-speech**

Observatorio Proxi. (2015). Informe del Observatorio Proxi. Barcelona. Retrieved from **https://www.observatorioproxi.org/images/pdfs/INFORME-proxi-2015.pdf**

Pulido, C. M., Redondo-Sama, G., Sordé-Martí, T., & Flecha, R. (2018). Social impact in social media: A new method to evaluate the social impact of research. PLOS ONE, 13(8), e0203117. **https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0203117**

Searle, J., & Soler, M. (2004). Lenguaje y ciencias sociales. Barcelona: El Roure.

Van Dijk, T. A. (2002). Discourse and racism. In D. T. Goldberg & J. Solomos (Eds.), A Companion to Racial and Ethnic Studies. Oxford: Blackwell.

Van Dijk, T. A. (2007). Discurso racista. In Igartua, J.; Muñiz, C. (2007) Medios de comunicación, Inmigración y Sociedad. (pp. 9–16). Salamanca: Universidad de Salamanca. Retrieved from **http://www.discursos.org/oldarticles/Discurso%20racista.pdf**

Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speec h Detection. IEEE Access, 6, 13825–13835. **https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8292838**

Wieviorka, M. (1992). El espacio del Racismo. Barcelona: Paidós.

Creation of the bag of words

ANNEX 1

ANNEX 1

In most cases, hate speech analysis uses lists of words that denote or tend to denote hatred, but this is not enough. That is why our analysis of articles and tools seeks to identify a bag of words that can be used in the hate speech monitoring and detection tool. The methodology developed to create a specific bag of words for the project is described below.

1. Review of existing bags of words and unified wordlist

We gleaned words from our analysis of scientific articles and tools in an attempt to come up with wordlists to identify hate speech. However, most of the bags of words from the tools we explored and the articles we consulted are not available, a fact that according to Fortuna and Nunes (2018) slows down progress in this field. While in some cases, articles or documentation related to existing tools describe methods for building a bag of words, the actual list compiled is not provided. Many of the articles analysed give examples of words or expressions, but do not offer a complete list of the vocabulary used.

Language is an added difficulty since our objective is to obtain or create a bag of words in Spanish. Even if we were able to get a bag of words in English, they would not be directly translatable due to the very specific way this sort of language (often slang) is used.

Of the articles reviewed where the authors used word lists, we checked which lists are available (like Mathew et.al. 2018) and which are not (Gitari et. al. 2015). Some recent studies started out with tweets and manual annotations, for example Crowdflower and Github open source repository **(https: // github. com / ZeerakW / hatespeech**), which are also in English (the lexicons from these resources are practically worthless in Spanish).

We compiled lexicons (bags of words) from the tools available and translated them into Spanish. Specifically, we obtained the following lists:

a. Bag of words from Hate meter
b. Bag of words from Hate base
c. Bag of words from Hate meter (hashtags)
d. Bag of words from PROXI

ANNEX 1

a. Bag of words from Hate meter

Website: **http://hatemeter.eu**

Collection date: April 2019

Collection methodology: consultation of the proposed list and verification it has been used in Spanish.

Words: Isis; Terrorismo; Fundamentalismo; Jihad; Bastardo; Guerra; Islam; Mohammed; Halal; Islam; Ramadan; Stop Invasión; Invasión; Invasores; Violencia; Sangre; Tolerancia cero; Sumisión; Guerra Civil; Ciudadanía; Integración; No integración; jóvenes musulmanes; Cultura; Migrantes; Expulsión; Acogida; Clandestinos; Africanos; Stop Inmigración; Recursos; Asilo; Refugiados; No Eurabia; Civilización; Patria; EU; Europa; Fronteras; Primero Nuestra Gente; Occidente; comunistas; No liberales; No democracia; Partidos; Idiotas; Mujeres; Izquierda; Burca; Niqab; Violación; Papa; ONG;
Gypsies; Blacks; Nigerians; Hebrews; Jews; Latinos; China; Filipinos; Bangladesh

b. Bag of words from Hate base

Website: http://hatemeter.eu
Collection date: April 2019

Collection methodology: consultation of the proposed list and verification that it has been used in Spanish. We later asked three people from different Autonomous Communities (Madrid, Andalusia, Catalonia) and one person from a Muslim association to review the list obtained and delete the words they were unfamiliar with. All the words included by at least one of the participants were kept on the list.

Words: amariconada; amariconadas; amariconado; **amariconados**; amariconar; **amariconé**; **amaricono**; **blanco**; **bollera**; **bolleras**; **cabezas cuadradas**; cabecita negra; **catalufo**; **catalufos**; **champinon**; **charnega**; **charnego**; **chele**; **chelo**; conguito; **conguitos**; **culandrón**; **culandrones**; **disminuido**; **disminuidos**

ANNEX 1

c. Bag of words - Hatemeter (hashtags)

Website: http://hatemeter.eu/
Collection date: April 2019

Methodology: a list of hashtags in English related to terrorism, religion, invasion, national/European identity, multiple target groups was obtained. Each hashtag was translated and tested on twitter to see whether it is used in Spanish. Hashtags used in Spanish were kept on the list.

Terrorism: #Isis; #Terrorismo (#Terrorism); #Terrorista (#Terrorists); #Fundamentalista (#Fundamentalists); #Jihadista (#Jihadists); #Jihad; #Seguridad (#Security); #AllahAkbar; #Atentado #AtaqueTerrorista; #Bastardo (#Bastards); #Guerra (#War); #HermanosMusulmanes (#Brotherhood, one of the most important international Islamist organizations); mention of cities / countries where terrorist attacks have occurred such as: #Paris (#Paris), #Alemania (#Germany); #Münster; #Francia #France); #Belgica (#Belgium).

Religion: #Islam; #Sharia; #Coran (#Quran); #Mahoma (#Mohammed, often associated with #hashtags #Pedofilo – #Pedofilia, (for having allegedly taken a child bride); #Halal; #Islam; #Mezquita (#Mosques); #AsiaBibi (Christian woman convicted for blasphemy in Pakistan in 2010); #Ramadan; #Halal. Often in juxtaposition to #Cristianos (#Christians); #Catolicos (#Catholics); #Cristiandad (#Christianity).

Invasion: #Stopinvasion (#Stopinvasion), #Invasion (#Invasion); #Invasores, (#Invaders); Violencia (#Violence); #Sangre (#Blood); #masacre (#Bloodbath); #ToleranciaCero (#ZeroTolerance); #Submisión (#Submission); #Guerracivil (#Civicwar).

Social Integration: #Ciudadania (#Citizenship); #Integración (#Integration); (#Sons, referring to the "second generation); #NoIntegración (#NoIntegration); #Jovenimusulman (#YoungMuslims); #Cultura (#Culture).

Immigration: #Migrant (#Migrants); # Expulsion (#Expelled); #Hospitality (#Hospitality); #Clandestino (#IllegalImigrants); #A-Africans (#Africans); #Africa; #TodosenCasa; #TodosVueltalPaisdeOrigen #EverybodyGoBacktotheirCountriesofOrigin);

ANNEX 1

#StopInmigracion (#StopImmigration); #Recursos (#Resources, in reference to migrants; an expression coined by the centre-right); #CerrarPuertos (#Closetheports); #DemandanteAsilo(#AsylumSeekers); #Aeropuertos (#Landings); #Refugiados (#Refugees).

National/European Identity: #DefensaEuropa; #NoEurabia; #Civilización (#Civilization); #España; #Españoles; #Pa- tria (#Homeland); #EU; #Europa (#Europe); #Frontera (#Borders); #PatriotaEspañol (#ItalianPatriots); #PrimerolosdeCasa (#FirstOurPeople); #Occidente (#West).

Multiple victims target group: centre-left coalition (#Comunista – #Communists; #Buonismo – #BleedingHearts; #Izquierdabuonista– #BleedingHeartLiberals; #Buenismooccidental –#NOPD – #NODemocraticParty; Conditions affecting women: #Mujer (#Women); #Burqa; #Burka; #Niqab; #Violación – #Rape; Non- governmental organisations (#ONG). Social minorities: (#Rom – #Gitano; #Negro – #Niggers; #Ebreos – #Judios; #Latinos; #Bangla; #China; #Filipinos).

d. Bag of words from PROXI

Website: http://www.observatorioproxi.org/
Collection date: April 2019

Methodology: We collected the word clouds they have published that are related to the project's objective.

Expressions:
IMMIGRANTS: "Distribution of refugees", instead of reception; "Immigrants intercepted, detained and held", as if they were criminals, "second generation immigrants", "foreigners with dual nationality", referring to Spanish citizens, "illegal", "undocumented", and even "Human burden", "avalanche", "wave" and "tide" of immigrants, "Not enough work for everyone here", "Host them at your house", "They come looking for handouts", "They are criminals, parasites, beggars", "They all engage

ANNEX 1

in illegal street sales, gorillas...", "pull factor", "Muslims are potential terrorists", "Muslims do not want to integrate" "They are invaders", "There are not enough jobs for everyone here", "Host them at your house",

ROMA: "They are parasites", "They do not want to integrate", "They steal for drugs", "They live off aid, abusing the system", "They do not respect the rules of coexistence", "They are parasites".

e. Bag of words from Wiktionary - ES Category - Derogatory terms

Website: **https://es.wiktionary.org/wiki/Apéndice:1000_palabras_básicas_en_español**

Collection date: April 2019

Methodology: On the link you will find a list of derogatory terms, including racist terms. However, it is not classified by target group.

A unified list of potential hate words was compiled, alphabetized, and repeated words were deleted.

ANNEX 1

2. Word selection focus group

A focus group was formed to select the relevant entries for the bag of words. The group was comprised of 10 male and female researchers of different ages and from different parts of Spain. All the words on the list were discussed and two different word categories were created:

I. Words that by themselves denote hatred.
II. Words that can be used in hate speech but whose meaning depends on the context or their relationship with adjacent words.

Word selection and classification into one of the two groups was done by consensus on a term-by-term basis and with the understanding that this was a starting point.

ANNEX 1

3. Quantitative analysis of racist profiles and incorporation of words

Another strategy consisted of obtaining a list of words from an openly racist set of Twitter user profiles.

First, a group of 20 profiles was identified based on obviously racist, Islamophobic or antisemitic hashtags or expressions. Only Spanish profiles written in Spanish were selected. A few highly politicized profiles were excluded. The proposal was validated by three team members. The following profiles were identified:

Profile account	Type of hate	Number of tweets timeline	Number of followers	Likes
@ account1	Racist - Islamophobia	986	206	1327
@ account2	Racist - Islamophobia	4777	152	1402
@ account3	Racist - Islamophobia	43.2K	2646	50.1K
@ account4	Racist - Islamophobia	2781	321	3856
@ account5	Racist - Islamophobia	48K	3399	34.8K
@ account6	Racist - Islamophobia	134K	2973	87.1K
@ account7	Racist - Islamophobia	6700	128	10.6K
@ account8	Racist - Islamophobia	21.9K	26.1K	25K
@ account9	Antisemitic	14.7K	4452	26.8K
@ account10	Antisemitic	4394	485	10.9K
@ account11	Antisemitic	2609	199	2197

ANNEX 1

Profile account	Type of hate	Number of tweets timeline	Number of followers	Likes
@ account12	Antisemitic	2673	507	364
@ account13	Antisemitic	7281	4004	44.7K
@ account14	Antisemitic	1196	35	287
@ account15	Antisemitic	105K	3775	27.3K
@ account16	Antisemitic	5537	126	2626
@ account17	Antisemitic	11.5K	5411	1880
@ account18	Antisemitic	16.4K	7171	1758

An application was developed using the Python programming language to capture and analyse tweets from these profiles. Tweets were captured using Twitter APIs: (**https://developer.twitter.com/en/docs**)**.** All tweets from the selected profiles were captured and their content analysed in two ways: First, a hashtag frequency analysis was conducted on all the tweets from the selected profiles enabling us to create a list of the most frequently used hashtags. Then, a frequency analysis of the words used in the tweets was conducted. For this analysis, standard Spanish libraries and sets of stopwords were used; words expressed in symbols or abbreviations were deleted manually. Following are the hashtags and words identified:

Frequent hashtags employed by racist users		
bds: 841	defensemddhh: 121	paro: 59
stopinvasion: 497	barcelona: 107	atletijuegolimpio: 57
israel: 461	últimahora: 93	stopbulos: 54

ANNEX 1

Frequent hashtags employed by racist users		
palestina: 350	eleccionesya: 92	corrupcion: 54
boycotteurovision2019: 319	artistasqueyadijeronno: 89	freepalestine: 49
stopislam: 306	ddhh: 87	pleasedontgo: 49
closeborders: 288	spexit: 81	nakba: 45
elecciones2019votoderechas: 272	openarms: 76	womentogaza: 45
gaza: 210	maccabiesapartheid: 72	matisyahu: 45
apartheid: 181	ivreich: 68	aquarius: 43
vox: 176	forolocalbds: 68	templarios: 41
openborders: 169	españa: 67	politicalcorrectness: 39
ue: 166	venezuela: 66	defiendeespaña: 37
españaviva: 150	merkel: 65	concertperpalestina: 37
eurabia: 135	nakba70: 64	ot18galafinal: 37
yonocomproapartheid: 130	palestine: 64	elai: 37
ongs: 126	bdsesddhh: 60	

ANNEX 1

Frequent words employed by racist users				
españa: 2028	tener: 391	venezuela: 262	sionistas: 201	votar: 165
ser: 1704	mal: 391	nacional: 262	ningún: 201	@dioshorus796: 165
israel: 1389	todas: 389	violencia: 260	persona: 200	religión: 164
: 1266	tal: 388	sistema: 259	hora: 200	grande: 164
@vox_es: 1139	@bdspaisvalencia: 387	andalucía: 259	mundial: 200	ó: 164
: 1132	alguien: 385	@orbitaeduardo: 255	madre: 200	des: 164
gracias: 1123	@jguaido: 385	haber: 254	@larecolectiva: 200	liberal: 164
años: 1104	https: 381	frente: 254	veces: 199	buen: 163
@marubimo: 1077	quiere: 381	musulmanes: 253	podría: 199	ataque: 163
así: 1069	tras: 379	@elhadadevox: 253	cristianos: 198	ue: 163
ahora: 1021	año: 378	tres: 252	ilegales: 198	: 163
vox: 1011	historia: 377	dar: 252	mierda: 197	régimen: 162
hoy: 1004	sino: 377	cosa: 251	quiero: 196	subvenciones: 162
the: 978	según: 376	derecho: 250	@ortega_smith: 196	mira: 162
va: 950	dicen: 370	cierto: 249	@anebald: 195	manera: 162
bien: 918	presidente: 370	cualquier: 249	@mimariban: 195	actual: 162
españoles: 915	dios: 369	siendo: 248	@el_pais: 194	ciudad: 162
hace: 911	barcelona: 366	for: 247	saber: 194	horas: 161
per: 859	usted: 366	ilegal: 247	artículo: 194	meses: 161
#bds: 847	millones: 365	gaza: 246	@elnahu_atr: 194	partidos: 161

ANNEX 1

Frequent words employed by racist users				
@rubnpulido: 834	psoe: 362	veo: 246	tema: 193	pronto: 161
ver: 788	quieren: 361	forma: 244	queda: 193	solidaritat: 161
puede: 779	els: 361	mayoría: 242	@irrintzialaves: 193	poner: 160
día: 775	@bdsmadrid: 358	trump: 242	general: 192	: 160
gente: 774	inmigración: 358	boicot: 242	seguro: 191	@sandonaegui: 160
hacer: 749	cosas: 357	solidaridad: 238	foto: 190	normal: 159
país: 731	: 356	policía: 237	control: 189	calle: 159
@santi_abascal: 722	política: 352	espero: 236	dia: 189	onu: 159
aquí: 715	judío: 352	@mariagtriana: 236	alguna: 189	libro: 159
solo: 709	#palestina: 352	medio: 235	terroristas: 189	ong: 159
dice: 707	apoyo: 350	vergüenza: 234	deja: 189	demonios: 159
palestina: 707	m: 346	casi: 233	@mnopasana: 189	acuerdo: 158
amb: 707	niños: 345	sabe: 232	saben: 188	última: 158
vez: 706	caso: 344	pasado: 229	demás: 188	mes: 158
sólo: 706	bueno: 342	paz: 228	humanos: 187	puedes: 158
mismo: 702	nunca: 341	semana: 228	buena: 187	encima: 157
vía: 694	és: 341	on: 227	gustó: 187	nombre: 157
gobierno: 689	libertad: 339	dijo: 227	peor: 186	puesto: 157
inmigrantes: 654	sociedad: 338	dels: 227	imagen: 186	siguen: 156
verdad: 638	casa: 331	justicia: 226	habla: 185	bandera: 155
cada: 637	ejemplo: 330	defensa: 226	paso: 185	voto: 155

ANNEX 1

Frequent words employed by racist users				
europa: 634	is: 329	canal: 226	político: 185	poble: 155
personas: 625	igual: 328	francia: 225	libre: 184	noticia: 154
menos: 622	sigue: 328	familia: 225	pide: 183	vosotros: 154
pues: 613	países: 327	soros: 225	@tiradorfranco13: 183	razón: 154
pueblo: 604	izquierda: 325	haciendo: 224	#apartheid: 182	rey: 154
dos: 603	vamos: 325	seguridad: 224	tierra: 182	sale: 154
vídeo: 601	internacional: 324	población: 223	cambio: 182	fuerza: 154
@idealismonazi: 596	#boycotteurovision2019: 322	iglesias: 223	primero: 181	entidad: 153
siempre: 595	: 321	sionismo: 222	mano: 181	ejército: 153
parte: 595	quién: 320	primera: 222	judía: 181	fronteras: 153
gran: 591	pp: 318	bajo: 222	feminismo: 181	@carrascomarimar:153
mundo: 581	debe: 317	@e_cycni: 222	palestinos: 181	través: 152
mujer: 570	campaña: 310	ciudadanos: 221	defender: 181	are: 151
mejor: 567	pedro: 309	@psoe: 221	total: 181	idea: 151
@bdscatalunya: 558	favor: 308	programa: 221	islam: 181	género: 151
decir: 543	#stopislam: 306	twitter: 220	avui: 181	situación: 151
da: 535	hacen: 304	visto: 220	palabras: 180	lucha: 151
cómo: 533	ir: 304	: 220	datos: 180	posible: 150
podemos: 532	ley: 302	derecha: 219	movimiento: 180	ninguna: 150
van: 524	medios: 302	entonces: 219	israelià: 179	amigos: 150

ANNEX 1

Frequent words employed by racist users				
mujeres: 516	@jadouken10: 302	alemania: 219	lleva: 178	centro: 150
vida: 502	además: 300	final: 217	#vox: 177	primer: 150
mientras: 500	nueva: 298	ello: 217	ayuda: 177	#españaviva: 150
#stopinvasion: 497	hombres: 298	saludos: 216	@elmundoes: 177	sabes: 149
@sanchezcastejon:493	pasa: 297	muerte: 215	padre: 177	htt: 149
sánchez: 493	@salvameoficial: 296	hijos: 214	cultura: 176	euros: 148
parece: 492	hablar: 292	lugar: 214	unas: 175	not: 148
judíos: 491	cataluña: 291	@hermanntertsch: 214	suport: 175	you: 148
@rescop1: 482	aunque: 290	l'apartheid: 214	algún: 174	militar: 148
hombre: 478	madrid: 289	incluso: 213	existe: 174	vídeos: 148
guerra: 477	elecciones: 288	nación: 213	@cama1610: 174	respecto: 148
claro: 474	ayer: 288	@edmondd09082129: 213	with: 173	liberales: 148
mas: 473	#closeborders: 288	social: 212	dan: 173	dejar: 148
tan: 473	ahí: 287	@casoaislado_es: 212	marruecos: 173	acto: 148
después: 470	fin: 286	#gaza: 212	miedo: 173	llegar: 148
#israel: 462	hilo: 285	momento: 211	@cristinasegui_: 173	@historiaespanna: 148
israelí: 452	mañana: 284	com: 211	palestí: 173	: 148
español: 451	vaya: 283	palestino: 210	invasión: 172	drets: 148
tiempo: 451	muchas: 282	hacia: 210	nivel: 171	discurso: 147
partido: 450	: 282	real: 209	maduro: 171	blanco: 147

ANNEX 1

Frequent words employed by racist users				
hecho: 442	problema: 277	dentro: 209	gusta: 170	señor: 147
luego: 436	falta: 277	grupo: 209	viene: 169	@varyingweion: 147
toda: 436	dicho: 275	único: 207	público: 169	sociales: 146
nadie: 435	ve: 275	http: 206	:/: 169	varios: 146
@youtube: 432	seguir: 274	tipo: 206	#openborders: 169	luz: 146
apartheid: 432	misma: 273	: 206	grandes: 168	feminista: 146
dinero: 422	odio: 271	voy: 205	eeuu: 168	refugiados: 146
sido: 419	#elecciones2019votoderechas: 271	ddhh: 205	información: 168	sr: 146
cuenta: 418	pueden: 270	lado: 204	pedir: 168	chile: 146
sionista: 418	trabajo: 269	hola: 204	@danaeon_: 168	@solof1sincirco: 146
poder: 413	@voxnoticias_es: 269	més: 204	origen: 167	feministas: 145
@rodrickgamer: 413	derechos: 268	digo: 203	tampoco: 167	zona: 145
realidad: 407	políticos: 268	hijo: 203	jajaja: 167	iglesia: 145
video: 405	cara: 265	debería: 203	@ino_forever: 167	aquest: 145
creo: 403	aún: 265	@a3noticias: 203	comunidad: 166	número: 144
@agnosis9: 399	@proucomplicitat: 265	civil: 202	plan: 166	cabeza: 144
días: 398	mayor: 264	mismos: 202	bds: 166	amigo: 144
nuevo: 395	sé: 264	pablo: 202	#ue: 166	dado: 143
española: 393	etc: 263	llama: 201	democracia: 165	puedo: 143

ANNEX 1

4. Qualitative analysis of racist tweets and incorporation of words

Lastly, the research team conducted a qualitative analysis of tweets containing hate speech to identify other frequent words to include in the initial list. More than 1500 tweets obtained using the above-described top-down capture strategy by target group were reviewed.

These strategies were used to generate our bag of words. This bag has two categories of words (words that denote hate on their own and words that can be used in hate speech).

Examples of tweets

ANNEX 2

ANNEX 2

Tweets containing words from the bag of words

Tweets containing words from the bag of words	Sudaca vuelvete al cono sur, eres un puto charnego
	Por cierto, los gilipollas son los progres de Podemos, PSOEy demás basura y toooooooooda la banda maricomplejines dePPyC's,puessus"líderes"están encantadosconestodelainvasiónprogramadaylaEurabiaquesenosviene encima.
	Detenidos dos machupichus por estafar 28.000 euros a un discpacitado

ANNEX 2

Hate speech tweets classified according to intensity

EXAMPLES OF TWEETS

Extreme hatred	Speech that incites violence.	Hay que matarlos a todos jatejode, un negro de mierda mas un negro menos, no se pierde nada.
		que negro re de mierda wacho. Después los hijos de mil putas se quejan cuando dicen de matarlos a todos a estos hijos de puta
		Que asco de gente!! A estos hijos de perra habría que colgarlos de los huevos hasta que mueran. Así de claro. Como hacen ellos. Ni expulsión ni hostias. Pena de muerte.
		Palestina resiste! Viva el pueblo palestino, Viva su heroica resistencia! Muerte a los sionistas!
		COTO UNA AYUDITA PARA ELIMINAR A LOS JUDIOS. UN ABRAZO Y GRACIAS
Offensive hate	Speech that is individually or collectively offensive that incites discrimination and propagates clichés and falsehoods.	Hay que joderse, otro extranjero que quiere que España se vaya a la mierda, que asco de gentuza, seguro que algo les debe a estos golpistas catalufos
		Que dices del "Pueblo Elegido" invasor y asesino del pueblo Palestino? Que dices de Angela Merkel que se ha reelegido en 3 ocasiones y va por la 4a? Que dices de los Jeques árabes que nadie eligió y gobiernan desde hace décadas...
		Gracias al #RamadanMubarak los délitos de #menas se concentrarán fuera del horario de ayuno
Neutral speech	Descriptive speech void of hatred.	¿Sabes cómo diagnosticar el #eccema en personas de piel negra? ¿ #SabíasQue las personas negras son más propensas a desarrollar formas más severas de eczema que las personas de otras etnias? #Dermatología https://www.medicalnewstoday.com/articles/325066.php …
Upstander speech	Alternative discourse which contributes to a counter-narrative, defying clichés or defending groups subjected to hatred.	Yo soy gitano y no soy delincuente
		¿Cómo te sonaría "el clan mafioso de argentinos", "el clan mafioso de judíos", "el clan mafioso de porteños"? Estigmatizar por religión, identidad, procedencia social o territorial se llama Xenofobia. Por cierto, parece que t... https://twitter.com/gabicerru/status/1127206684655529985 , https://twitter.com/gabicerru/status/1127206684655529985
		Bulo sobre actos delictivos de un grupo de menores en Calella de Mar. Hemos contactado por teléfono con Policía Local de Calella de Mar y Mossos d'Esquadra de Pineda de Mar y confirman que ha habido hurtos pero el resto de la información es FALSA.

ANNEX 2

Tweets classified by type of language

Language that incites or advocates violence	Yo diferencio legalidad de JUSTICIA, estoy absorto de la capacidad de aceptación de la población, por ejemplo cuando un moro viola a una hija o sobrina. Yo sin duda haría justicia y luego entraría en Soto, pero feliz. No se si me explico Lynn
	La hermana de un amigo sale de rendir de recibirse y viene un negro y le roba el celular... Ah pero uno dice que a los negros hay que matarlos apenas nacen y te miran mal.
	No seais fachas. Es solo un pobre inmigrante que ha dejado su pais de origen y su familia en busca del sueño español. MECAGOENSUPUTAMADRE. Lo que le hace falta a este mierda es una buena ensalada de ostias, que por lo que se ve, todavia no se las han dado.Y despues a su puto pais.
Justification, jokes, trivialization of violence towards 'others'	El problema reside en que si ahora uno le parte la cara al moro, saldría en todos los telediarios con el titular de ataque ra cista de un blanco a un pobre y desvalido musulman, recordemos lo que pasó en Barcelona con el vigilante de seguridad del tren,que fue ha sta sancionado.
	Si todo el mundo que dice ser superviviente del Holocausto lo es de verdad, entonces a quien mato Hitler? **#hitler #holocuento**
	Ojalá el Karma tape la boca a ésa sinvergüenza, ósea q está diciendo q x ser extranjero y negro tienen derecho a matar,violar y apalear pués para q su familia o ella sea la próxima en recibir tan maravilloso trato.
Divisive and otherness (them versus us) language	Seguro k era un moro??que casualidad si es que Barcelona ya casi parece Casablanca.
	Yo os traduzco: los menas tendrán en Cataluña una paga de 665 euros mensuales hasta los 23 años. El efecto llamada será brutal. Cuando os roben, cuando os atraquen, cuando os agredan o violen, volved la cabeza hacia el Parlamento de Cataluña y en las próximas elecciones haced algo.
	Por culpa de los negros que violan y matan mujeres nos tratan a todos los hombres como el enemigo. Pues NO, el enemigo está e n el gobierno y es el que deja impune estos actos y permite que cualquier inmigrante se quede en nuestro país y pueda hacer lo que le de la gana.

ANNEX 2

Propagation of stereotypes and prejudices	Peleas tribales, comidos y pagados por los españoles, se dedican más cómodamente a lo que han estado haciendo durante más de 2000 años a peleas tribales por el control. ....Claro que de un puntapié los problemas va a parar al otro lado de la verja..y todos tranquilos #Stopinvasion [Link a noticia sobre pelea entre marroquís en Ceuta]
	Por esos gitanos que van a la autoescuela en su propio coche!!
	Hemos votado que queremos seguir trayendo moros para que sigan haciendo esto y viviendo de nuestras ayudas. Es lo que hay, ahora toca joderse.
Spread of rumours	Este año desgraciadamente ha habido 20 manadas, solo se habla de esta, las otras víctimas no tienen derechos pues han sido violadas por moros sudacas y demás escoria.
	Nooooo, así no es la noticia, la noticia es QUE OTRO EXTRANJERO, MORO, la más señas, a si el causante de esta muerte, y ya va n unas cuantas muertes, violaciones y maltrato a mujeres por parte de esta gentuza que vosotros seguís protegiendo y ayudando,sois una banda SINVERGÜENZAS.
	Y siguen viniendo este finde vinieron tres cientos como no cierren las fronteras vamos de culo y a la ruina se necesita control ayudas a la gusrdia civil para proteger las fronteras.
Dissemination of false facts	Este año desgraciadamente ha habido 20 manadas, solo se habla de esta, las otras víctimas no tienen derechos pues han sido violadas por moros sudacas y demás escoria.
	En Móstoles una manada de moros, para variar, trataron de violar a una chica y al salir sus amigos en su defensa estos animal es les atacaron con navajas. No habrá ni manifestaciones feministas ni a ministra dirá de cambiar el código penal. Más asco no podéi s dar.
	Sin querer hacerme el experto, Europa Occidental está más cerca de Europistán que otra cosa. Lo que no se dice en los medios es terrible. Alemania tiene colegios donde el alumnado nativo ya no resiste porque es atormentado por los alumnos musulmanes.

ANNEX 2

Flawed argumentation	La sociedad española es una de las más seguras. El asesino de Parla es un moro magrebí con antecedentes, uno de ellos por tentativa de homicidio, que habría sido deportado si gobernara #VOX. La gran mayoría de los violadores y asesinos son de origen inmigrante. Sois cómplices. pic.twitter.com/6Y2cmMxD74
	Acabo de oir en la radio, que la pelea entre marroquíes y sudamericanos se debió a que los putos sarracenos intentaron violar a una chica dominicana. Está claro que el problema, no es la inmigración, el problema son los moros y su cultura
	Esa etnia por casualidad no será la misma que lleva decenios, sino siglos, saltan2las leyes de los payos, robando, engañando y mintiendo a los payos, pero al mismo tiempo queriendo dar pena para vivir del cuento. Hay señol guardia sivil que me han roba do la fregoneta. Haaay i
Metaphor-dehumanizing comments-irony	Menas alojados en hotel de Calella, se espera un plácido verano allí
	No es lo mismo que un judío se bañe a que me bañe con un judío
	Ya mismo soy un musulmán me tocas y exploto culiaaaa
Derogatory non-verbal structures e.g. emojis, punctuation, capital letters, etc.	JARED LETO PARECE UN MUSULMAN LPM EN CUALQUIER MOMENTO GRITA ALLAHU AKBAR, DE CARA PARECE JESUS Y NI HABLAR DE QUE ANDA A SABER XQ LLEVO UNA CABEZA FALSA #MetGala #MetGala2019 pic.twitter.com/UTUaTyh6RH
	Continúa en tu HOLOCUENTO, asno repugnante JUDIO, descendiente de SIÓN. Sigue viendo y creyendo en la "la vida es bella" o "el niño con la pijama de rayas". Ponte a leer el verdadero holocausto: El holocausto PALESTINO, por parte de los de tu especie. "Fiesta del Purim"

ANNEX 2

Tweets featuring different types of upstander speech

Publication of one's own comments: argumentation with evidence denying prejudices; drawing attention to positive experiences.	Yo soy gitano y no soy delincuente
Anti-rumour interventions: counter-rumour data and explanations.	Bulo sobre actos delictivos de un grupo de menores en Calella de Mar. Hemos contactado por teléfono con Policía Local de Calella de Mar y Mossos d'Esquadra de Pineda de Mar y confirman que ha habido hurtos pero el resto de la información es FALSA.
Pedagogical interventions: facts and data to shed light on the debate.	Más de 68'5 millones de personas viven fuera de sus hogares por la guerra, la violencia y graves violaciones de sus derechos fundamentales. Esto supone el número más alto jamás registrado **#DíaMundialDelRefugiado #RefugeeDay**
Awareness-raising interventions: emotional appeal.	¿Cómo te sonaría "el clan mafioso de argentinos", "el clan mafioso de judíos", "el clan mafioso de porteños"? Estigmatizar por religión, identidad, procedencia social o territorial se llama Xenofobia.
Qualification of comments by other users: support or disqualify positive / negative comments.	"Dos clases de menas". Qué penita joder. Qué penita. Hay dos clases de adultos: quienes creen en los chavales y les dan apoyo bajo las circunstancias que sean y quienes no. Os cuento una historia en este hilo
Reporting hateful comments: through social media tools and apps created specifically for that purpose.	Todavía quedan otras elecciones. Queremos cubrirlas y estar preparad@s para los bulos que van a intentar colarnos, como que "el peligro es el inmigrante". ¿Nos ayudas a seguir?