METHODS & CONCEPTS

DOI: 10.51480/1899-5101.19.1(42).970

Detecting and Categorizing Islamophobic Hate Speech on Spanish Twitter/X: Discursive Patterns, Toxicity Scores, and BERT-Based Classification

William González-Baquero
0000-0001-5760-9771
University of Salamanca, Spain

Carlos Arcila Calderón
0000-0002-2636-2849
University of Salamanca, Spain

Javier J. Amores.
0000-0001-7856-5392
University of Salamanca, Spain

Rocío Zamora
0000-0002-0541-2456
University of Murcia, Spain

Pilar Garrido
0000-0002-3305-5257
University of Murcia, Spain

Maria del Mar Grandio
0000-0002-2577-4059
University of Murcia, Spain

Pedro Rojo
Al Fanar Foundation, Spain

Abstract: Islamophobic hate speech on social media includes overt toxicity and contextual frames that associate Muslims with invasion, criminality, cultural threat, or racialized exclusion. This study examines 21,326 Spanish-language tweets about Islam and Muslims posted in Spain between 2010 and 2022. We combine human annotation, exploratory discursive coding, Perspective API scores, and a multilingual BERT classifier. Human coders classified 28.2% of tweets as Islamophobic hate speech. Explicit hostility was the most frequent pattern, followed by invasion, cultural threat, criminality, race/ethnicity, and socioeconomic burden. Although hate-labeled tweets received significantly higher toxicity-related scores than non-hate tweets, many remained below a 0.70 high-probability threshold. The BERT classifier achieved 0.89 accuracy, 0.97 precision, 0.85 recall, and 0.91 F1 for the hate class. The findings show that generic toxicity measures capture overt aggression more effectively than contextual hostility, supporting target-specific, context-sensitive detection.

Keywords: Religion; Islamophobia; Hate speech; Toxicity; Stereotypes.

INTRODUCTION

Islamophobic hate speech on social media is not limited to insults, profanity, or threats. It can also be expressed through frames that present Muslims as invaders, criminals, cultural threats, or racialized outsiders (Awan, 2014; Evolvi, 2018; Vidgen & Yasseri, 2020). This distinction matters for automated detection because general toxicity systems are designed to recognize overt verbal aggression and may miss hostility that depends on context or group-based framing (Lees et al., 2022; Wulczyn et al., 2017). Social media facilitates the rapid circulation of discriminatory discourse and connects anti-Muslim hostility with wider debates about migration, security, national identity, and European values. In Spain, online conversations about Islam frequently associate religious identity with migration, terrorism, cultural difference, and political conflict (Alcántara-Plá & Ruiz-Sánchez, 2017; Fuentes-Lara & Arcila-Calderón, 2023; González-Baquero et al., 2023; Zamora Medina et al., 2021). Twitter/X is not representative of Spanish public opinion, but its open structure and historical availability through the API make it a useful setting for examining public platform discourse at scale.

This study analyzes 21,326 Spanish-language tweets about Islam and Muslims published in Spain between 2010 and 2022. It makes three contributions. First, it estimates the prevalence and temporal distribution of Islamophobic hate speech. Second, it describes recurrent and overlapping discursive patterns within the hateful subset. Third, it compares human annotation with a general toxicity-scoring system and a target-specific multilingual BERT classifier. The central premise is that computational architecture should follow the construct being measured: if Islamophobic hostility can be explicit or frame-based, its detection requires more than generic toxicity scores.

THEORETICAL FRAMEWORK

ISLAM AND MUSLIM COMMUNITIES IN SPAIN: DEMOGRAPHIC AND SOCIO-POLITICAL CONTEXT

Islam is one of Spain's main minority religions. Muslims represent around 4% of the national population, slightly over two million people, and vary considerably in nationality, citizenship, migration history, ethnicity, religiosity, and generation. The largest Muslim populations are concentrated in Catalonia, Andalusia, Madrid, and the Valencian Community, while Ceuta and Melilla have the highest relative proportions (Comisión Islámica de España, 2022; Observatorio Andalusí, 2023). Despite this diversity, public and online discourse often reduces Muslims

to recurring associations with migration, terrorism, cultural difference, gender oppression, and incompatibility with Western or European values (Alcántara-Plá & Ruiz-Sánchez, 2017; Fuentes-Lara & Arcila-Calderón, 2023; González-Baquero et al., 2023; Zamora Medina et al., 2021). Spain's historical relationship with Islam, including memories of Al-Andalus and later national narratives, together with its proximity to North Africa and position as a southern European border country, also shapes contemporary debate (García Sanjuán, 2016, 2020; Ríos Saloma, 2008). These factors make narratives of invasion, cultural threat, security, and national identity especially salient.

ISLAMOPHOBIA, ANTI-MUSLIM RACISM, AND DISCURSIVE FRAMES OF HOSTILITY

Islamophobia is a contested but widely used concept for describing hostility, prejudice, discrimination, exclusion, or violence directed at Islam, Muslims, or people and practices perceived as Muslim. Although the term has earlier historical uses, its consolidation in public and academic debate is commonly associated with the Runnymede Trust's report Islamophobia: A Challenge for Us All, which helped frame anti-Muslim hostility as a specific form of discrimination requiring social and political attention (Elahi & Khan, 2017; Gómez, 2021). In this article, Islamophobia is understood as a multidimensional form of anti-Muslim hostility that overlaps with racism, xenophobia, religious intolerance, securitization and aporophobia, while retaining specific features linked to the stigmatization of Islam and "Muslimness".

Following Bleich's formulation, Islamophobia may be defined as indiscriminate negative attitudes or emotions directed at Islam or Muslims (Bleich, 2011, cited in Vidgen & Yasseri, 2020). However, for the purposes of this study, it is important to move beyond purely attitudinal definitions and consider how Islamophobia is expressed discursively. Islamophobic discourse does not only communicate negative evaluations of Islam or Muslims; it also constructs "Muslimness" through recurring associations with threat, foreignness, inferiority, violence, cultural incompatibility, and exclusion (Awan, 2014; Evolvi, 2018; Meer, 2013; Taras, 2013; Zempi & Awan, 2019). Corpus-assisted discourse studies have also shown that media representations of Islam and Muslims are often organized around recurrent associations with conflict, threat, extremism, cultural difference, and national belonging. This literature is relevant because it treats anti-Muslim discourse not only as a set of isolated hostile expressions, but as a patterned discursive phenomenon shaped by repeated lexical, semantic, and narrative associations (Baker et al., 2013).

A central feature of contemporary Islamophobia is the racialization of religion: although Islam is a religion and not a race, Muslims are often treated in public

discourse as if they constituted a homogeneous, visible, inherited and culturally fixed group (Meer, 2013; Meer & Modood, 2009; Taras, 2013). In this process, "Muslimness" becomes a social marker that can be attributed through names, appearance, clothing, language, nationality, ethnicity, migration background or perceived cultural practices, so that people may become targets of Islamophobic hostility not only because of their religious beliefs or practices, but also because they are perceived as Muslim. This logic also helps explain the different levels of explicitness through which Islamophobic discourse operates: some messages articulate anti-Muslim hostility through direct verbal aggression, including insults, threats, profanity, dehumanizing language or open calls for exclusion, while others do so through indirect, coded or frame-based forms that construct Muslims as a collective threat without necessarily relying on insults or profanity (Awan, 2014; Evolvi, 2018; Vidgen & Yasseri, 2020). For this reason, the distinction between explicit hostility and frame-based hostility is central to this study: the former refers to messages in which the hateful orientation is linguistically direct, for example through derogation, threats, dehumanization or the rejection of Muslims as legitimate members of society; the latter refers to messages that articulate anti-Muslim meanings through broader interpretive frames, presenting Muslims as invaders, criminals, terrorists, welfare burdens, threats to Western identity or culturally incompatible subjects, even when they avoid overtly abusive vocabulary.

Several frames are especially relevant for this study. The first is invasion, in which Muslim presence is represented as a demographic, territorial, or cultural takeover. The language of invasion transforms migration and religious diversity into a threat to national continuity, European identity, or social cohesion, condensing demographic fear, cultural replacement, territorial anxiety, and political emergency (Gualda & Rebollo, 2020; Gorman & Culcasi, 2020). The second is securitization, in which Islam and Muslims are associated with terrorism, radicalization, violence or threats to national security. This discourse shifts the discussion from rights, citizenship, and coexistence to risk, surveillance, control, and emergency (Saeed, 2016; Awan & Zempi, 2016). A third frame is criminalization, which links Muslims or Muslim-coded migrants with delinquency, disorder, sexual violence, theft or social danger. Although it may overlap with securitization, criminalization refers more specifically to everyday insecurity and moral disorder rather than terrorism. This frame is supported by research showing how anti-Muslim discourse often connects Muslim identity with crime, deviance, sexual threat, and public disorder, especially when religion, migration, and racialized suspicion are combined in media or political narratives (Cockbain & Tufail, 2020; Awan, 2014).

A fourth frame is a cultural threat, in which Islam is represented as incompatible with Spanish, European, Western, secular, liberal, or democratic values.

This frame may appear in apparently civil or value-based language while still legitimizing exclusionary attitudes toward Muslims (Meer, 2013; Taras, 2013). A fifth frame concerns race and ethnicity, as Islamophobic discourse frequently racializes religion by linking Muslim identity to Arabness, Moroccanness, foreignness, migration status, skin color, or ethnic origin (Meer, 2013; Meer & Modood, 2009; Taras, 2013). A sixth frame involves gender, in which women's rights, gender equality, the veil, sexuality or family norms are mobilized to portray Islam as inherently patriarchal, oppressive or incompatible with liberal values. Research on gendered Islamophobia has shown that visibly Muslim women are often positioned both as oppressed subjects and as symbolic markers of cultural threat (Perry, 2014; Hopkins, 2016; Zempi & Chakraborti, 2012, 2015). Finally, the socio-economic burden frame represents Muslims or Muslim-coded migrants as a cost to public services, welfare systems, employment, housing, or social order, connecting economic anxiety with cultural and racialized exclusion. This frame overlaps with broader research on anti-migrant hostility, welfare chauvinism, classed exclusion, and aporophobia, where racialized or migrant populations are represented as undeserving recipients of public resources or as economic threats to the majority group (Awan, 2014; Gorman & Culcasi, 2020).

These frames are analytically separable, but they often co-occur in real discourse. A single message may simultaneously represent Muslims as invaders, criminals, welfare burdens, and threats to Western values. For this reason, the present study treats the subcategories of Islamophobic discourse as non-mutually exclusive and exploratory. They are not intended as a definitive taxonomy of Islamophobia, but as an empirical tool for identifying recurrent discursive patterns within messages classified as Islamophobic hate speech. This distinction is important because the main binary classification, Islamophobic hate speech versus no Islamophobic hate speech, provides the strongest basis for quantitative inference, whereas the subcategories are used to describe internal variation within the hateful subset.

This conceptualization also has methodological implications. If Islamophobic discourse includes both explicit hostility and frame-based hostility, then generic toxicity detection may capture only part of the phenomenon. Messages containing insults, profanity, or direct threats may receive high toxicity scores, whereas messages organized around invasion, cultural threat, racialization, or securitization may be less likely to reach conventional toxicity thresholds. For this reason, the article examines not only whether messages are classified as Islamophobic hate speech, but also how toxicity scores vary across different exploratory discursive patterns.

HATE SPEECH, OFFENSIVE LANGUAGE, AND TOXICITY

Research on online hostility must distinguish hate speech from offensive language and toxicity. Hate speech targets individuals or groups based on protected or vulnerable characteristics and stigmatizes, threatens, devalues, or legitimizes their exclusion (Fortuna & Nunes, 2018; Schmidt & Wiegand, 2017). Offensive language is broader and may include insults or profanity without targeting a protected group. In computational moderation, toxicity refers to language likely to disrupt participation in an online discussion (Wulczyn et al., 2017). Perspective API estimates probabilities for toxicity, severe toxicity, identity attack, insult, profanity, and threat (Jigsaw, n.d.; Lees et al., 2022). These scores indicate similarity to previously annotated toxic language rather than directly measuring hate speech or Islamophobia. Here, Islamophobic hate speech comprises messages that stigmatize, threaten, or legitimize exclusion against Islam, Muslims, or people perceived as Muslim. Toxicity is treated as a separate linguistic attribute. Perspective API, therefore, assesses overlap with explicit toxic language, allowing us to test whether generic measures capture indirect, coded, or frame-based Islamophobic hostility.

AUTOMATED DETECTION OF ISLAMOPHOBIC HATE SPEECH

Automatic hate-speech detection has progressed from keyword lists and rule-based systems to supervised machine learning and Transformer models. Early methods identified explicit slurs but struggled with context, irony, coded language, negation, quotation, and counter-speech (Schmidt & Wiegand, 2017). Classifiers such as support vector machines and logistic regression improved detection using annotated examples yet remained dependent on surface lexical patterns and often missed implicit or target-specific hostility (Burnap & Williams, 2015; Fortuna & Nunes, 2018). Transformer-based models such as BERT address some of these limitations by capturing contextual relationships and adapting pretrained representations to particular tasks (Devlin et al., 2019; Mozafari et al., 2020). Monolingual and multilingual variants are especially useful for Spanish corpora grounded in specific cultural and political contexts (Cañete et al., 2023; Conneau et al., 2020).

Islamophobic hate speech remains challenging because its meaning can depend on local debates, migration narratives, or coded associations among Islam, race, security, and national identity. Anti-Muslim hostility is not always expressed through direct abuse (Awan, 2014; Evolvi, 2018; Vidgen & Yasseri, 2020). General-purpose tools such as Perspective API measure toxicity, identity attacks, insults, profanity, and threats, but may overlook frame-based hostility involving invasion, cultural incompatibility, racialized suspicion, or securitization. Comparing these scores with human annotation and a target-specific

BERT classifier therefore tests whether generic toxicity measures are sufficient or specialized models provide more robust detection.

In this study, automated detection is approached in two complementary ways. First, Perspective API is used as an external scoring tool to assess the degree to which tweets classified by human coders as Islamophobic hate speech overlap with toxicity-related attributes. Second, a multilingual BERT-based classifier is trained and evaluated against the human-coded binary classification of Islamophobic hate versus non-hate messages, excluding discarded messages from training and evaluation. Because the subcategories of Islamophobic discourse showed lower intercoder reliability than the binary classification, the model is not trained to predict those exploratory subcategories. Instead, the classification task focuses on the more reliable main distinction between Islamophobic hate speech and non-Islamophobic content.

Based on this theoretical framework, the study addresses the following research questions:

- RQ1: What is the prevalence and temporal evolution of Islamophobic hate speech in Spanish-language tweets about Islam and Muslims between 2010 and 2022?
- RQ2: What exploratory discursive patterns characterize tweets classified as Islamophobic hate speech, and how do these patterns co-occur?
- RQ3: To what extent do tweets classified as Islamophobic hate speech receive higher Perspective API toxicity-related scores than tweets classified as non-hate, and which exploratory subcategories are associated with higher toxicity scores?
- RQ4: How accurately can a multilingual BERT-based classifier detect Islamophobic hate speech compared with human annotation?

METHODOLOGY

DATA COLLECTION

The corpus comprises 21,326 Spanish-language tweets about Islam and Muslims in Spain, posted between January 1, 2010, and December 31, 2022. Data were collected through Twitter API v2 using Python and Tweepy, with Boolean combinations of relevant keywords. The search was limited to tweets associated with Spain when location metadata were available. Retweets, replies, duplicates, link-only posts, and texts without analyzable content were excluded, while engagement metrics were retained. Because Twitter/X users are not representative of Spain's population, the corpus reflects platform discourse rather than public opinion.

HUMAN ANNOTATION, CATEGORIES, AND RELIABILITY

A manual content analysis classified each tweet as Islamophobic hate, non-hate, or discarded. Hate included messages attacking, stigmatizing, threatening, or legitimizing exclusion against Islam, Muslims, or perceived Muslims. Non-hate covered informational, neutral, supportive, or otherwise non-hateful content; discarded tweets were irrelevant or targeted other groups. Hate-labeled tweets were also assigned non-mutually exclusive exploratory categories: socioeconomic burden, criminality, terrorism, invasion, illegality, cultural threat, gender, race/ethnicity, classism/aporophobia, and explicit hostility. Seven trained external coders each annotated approximately 3,000 tweets using a detailed codebook and calibration exercises. An eighth coder independently reannotated 15% of every batch, creating a stratified validation subset. Krippendorff's α for the binary hate/non-hate distinction was .712, indicating acceptable reliability (Krippendorff, 2018). Reliability for the ten subcategories was lower, ranging from .50 to .66. Consequently, the binary classification supports quantitative inference and model evaluation, while subcategory findings are interpreted as exploratory.

PERSPECTIVE API TOXICITY SCORING

Perspective API measured overlap between human-coded Islamophobic hate and explicit toxic language; it was not used to detect Islamophobia directly. It returned probability scores from 0 to 1 for toxicity, severe toxicity,identity attack, insult, profanity, and threat. Discarded messages were excluded, and independent-samples *t* tests compared hate and non-hate tweets. Descriptive statistics for the hateful subset assessed scores at or above a conservative .70 threshold. This was an operational choice, not an API recommendation or universal boundary: Islamophobic messages may lack overt toxicity, while toxic messages are not necessarily Islamophobic.

MULTILINGUAL BERT-BASED CLASSIFICATION

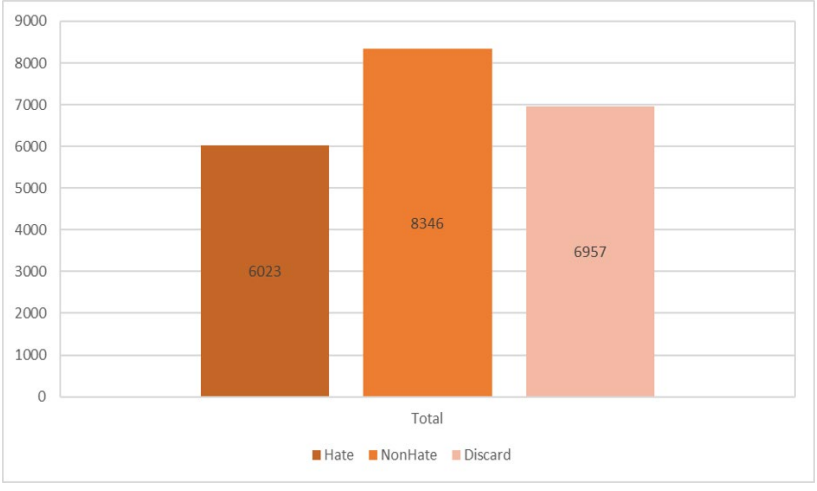
A multilingual BERT classifier (bert-base-multilingual-uncased) was fine-tuned on Spanish-language tweets using Hugging Face's TFBertForSequenceClassification, BertTokenizer, and BertConfig. The binary labels were 0 = Islamophobic hate and 1 = non-hate; discarded messages were excluded from training and evaluation. Performance was assessed against human annotations using accuracy, precision, recall, F1, macro and weighted averages, and class-level support. Because the exploratory subcategories had lower intercoder reliability, the model predicted only the more reliable binary distinction rather than individual discursive patterns.

RESULTS

FREQUENCY OF CLASSIFIED HATE SPEECH

The manual content analysis classified the corpus into three initial categories. Of the 21,326 tweets analyzed, 6,957 messages were discarded because they were not relevant to the object of study or did not specifically target Islam or Muslims, representing 32.6% of the corpus. A further 8,346 messages were classified as non-hate, representing 39.1% of the corpus. Finally, 6,023 messages were classified as Islamophobic hate speech, representing 28.2% of the total sample category (see Figure 1). Regarding the temporal evolution of Islamophobic hate speech, the highest number of hateful messages was identified in 2021, with 985 tweets, representing 16.35% of all messages classified as Islamophobic hate speech. This was followed by 2020, with 746 messages, and 2022, with 716 messages. These three years therefore concentrate a substantial proportion of the Islamophobic hate messages identified in the corpus. By contrast, the highest number of non-hate messages appeared in 2016, with 2,209 messages, representing 26.47% of the non-hate.

Figure 1. Frequency of tweets



Source: Own elaboration

Public engagement metrics also varied across the period studied. Tweets classified as Islamophobic hate speech reached their highest average number of likes in 2021 (M = 11.23), while the highest average number of retweets was observed in 2017 (M = 8.52). However, these averages should be interpreted with caution because engagement metrics were highly skewed. In 2021, likes showed a standard deviation of 88.11 and a maximum value of 1,387, while in 2017 retweets

showed a standard deviation of 46.3 and a maximum value of 576. These high standard deviations indicate that the averages were strongly influenced by a small number of highly visible tweets.

EXPLORATORY DISCURSIVE PATTERNS AND CATEGORY CO-OCCURRENCE

Within the subset of messages classified as Islamophobic hate speech, the study identified ten non-mutually exclusive exploratory subcategories. These categories should be interpreted as descriptive discursive patterns rather than as a definitive taxonomy of Islamophobia. The most frequent category was explicit hostility, with 2,448 messages, representing 40.64% of the Islamophobic hate subset. This was followed by invasion, with 1,165 messages (19.34%), cultural threat, with 979 messages (16.25%), criminality, with 961 messages (15.96%), race/ethnicity, with 926 messages (15.37%), and socio-economic burden, with 814 messages (13.52%). Other categories appeared less frequently but remained analytically relevant: terrorism, with 615 messages (10.21%); gender, with 409 messages (6.79%); classism/aporophobia, with 315 messages (5.23%); and illegality, with 267 messages (4.43%) (see Table 1).

Table 1. Frequency of messages classified in the exploratory categories of Islamophobic discourse.

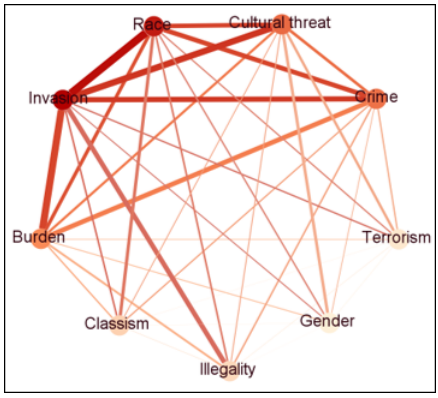
Years	Burden	Crime	Terrorism	Invasion	Illegality	Cultural threat	Gender	Race	Classism	Irrational	Total Hate
2010	0	0	0	0	0	0	0	0	0	4	4
2011	1	17	8	5	1	16	15	4	10	83	114
2012	10	38	15	11	3	24	10	10	30	208	241
2013	11	55	26	10	6	58	21	25	42	390	442
2014	20	51	56	32	10	72	44	27	43	352	522
2015	16	26	112	38	2	33	14	9	3	199	414
2016	14	7	73	73	1	52	32	22	11	172	349
2017	40	28	110	28	1	189	28	97	6	64	474
2018	51	47	41	39	10	76	30	60	0	91	342
2019	135	193	30	86	29	139	62	2	0	177	674
2020	166	125	64	339	47	121	63	197	34	221	746
2021	256	203	62	413	116	141	58	377	102	216	985
2022	94	171	18	91	41	58	32	96	34	271	716
Total	814	961	615	1165	267	979	409	926	315	2448	6023

Source: Own elaboration

Because the subcategories were coded as non-mutually exclusive, a single message could contain more than one discursive pattern. For this reason, we examined descriptive co-occurrence patterns among categories. The most frequent co-occurrence was invasion and race/ethnicity, with 254 messages coded in both categories. This was followed by socio-economic burden and invasion, with 234 shared messages, and cultural threat and invasion, with 209 shared messages. Other frequent overlaps included criminality and invasion, invasion and illegality, socio-economic burden and criminality, and criminality and race/ethnicity (see Figure 2).

These co-occurrences suggest that Islamophobic discourse often combines multiple frames within the same message. Invasion appears as a prominent frame in this corpus, frequently overlapping with race/ethnicity, socio-economic burden, cultural threat, criminality, and illegality. However, these results should be interpreted descriptively. They show patterns of overlap among exploratory categories, but they do not establish causal relationships or confirm a definitive structure of Islamophobic discourse.

Figure 2. *Co*-**occurrence patterns among exploratory categories of Islamophobic discourse.**



Source: Own elaboration

PERSPECTIVE API TOXICITY SCORES

To evaluate the relationship between human-coded Islamophobic hate speech and automated toxicity-related attributes, we compared Perspective API scores between messages classified as Islamophobic hate and messages classified as non-hate. Discarded messages were excluded from this analysis. Independent samples t-tests showed statistically significant differences across all six Perspective API attributes (see Table 2).

Table 2. Descriptive API Perspective attributes in the hate/non-hate categories

Attribute	Hate	N	M	D.st	D Cohen
TOXICITY	0	8178	.24	.214	1.269
	1	5928	.53	.242	
SEVERE_TOXICITY	0	8178	.07	.139	1.042
	1	5928	.27	.233	
IDENTITY_ATTACK	0	8178	.13	.151	1.025
	1	5928	.31	.197	
INSULT	0	8178	.20	.221	1.262
	1	5928	.50	.253	
PROFANITY	0	8178	.17	.269	0.808
	1	5928	.42	.345	
THREAT	0	8178	.03	.078	0.434
	1	5928	.08	.143	

Source: Own elaboration

Messages classified as Islamophobic hate scored higher than non-hate messages in toxicity, severe toxicity, identity attack, insult, profanity, and threat. For toxicity, hate messages obtained a mean score of 0.53 (SD = 0.242), compared with 0.24 (SD = 0.214) among non-hate messages, $t(14104) = -75.123$, $p < .001$, $d = 1.269$. For severe toxicity, hate messages also scored higher ($M = 0.27$, $SD = 0.233$) than non-hate messages ($M = 0.07$, $SD = 0.139$), $t(14104) = -65.114$, $p < .001$, $d = 1.042$. Similar differences were observed for identity attack, insult, profanity, and threat. Effect sizes ranged from moderate to large, with the smallest effect observed for threat ($d = 0.434$) and the largest effects observed for toxicity and insult.

However, the descriptive analysis of the hateful subset shows that most human-coded Islamophobic hate messages did not reach high-probability toxicity thresholds. Among messages classified as Islamophobic hate speech, toxicity had the highest mean score ($M = 0.526$, $SD = 0.242$), followed by insult ($M = 0.499$, $SD = 0.252$), profanity ($M = 0.421$, $SD = 0.344$), identity attack ($M = 0.308$, $SD = 0.197$), severe toxicity ($M = 0.274$, $SD = 0.233$), and threat ($M = 0.084$, $SD = 0.142$) (see Table 3).

The third quartile reached or approached the operational threshold of 0.70 only for toxicity ($Q3 = 0.736$), profanity ($Q3 = 0.784$), and insult ($Q3 = 0.699$). This indicates that only the upper segment of Islamophobic hate messages displayed high levels of explicit toxicity. Perspective API scores also varied across the exploratory subcategories of Islamophobic discourse. The highest toxicity-related scores were observed in explicit hostility, race/ethnicity, and classism/aporophobia. Explicit hostility obtained the highest mean scores for toxicity ($M = 0.666$), insult ($M =$

0.637), and profanity (M = 0.648). Race/ethnicity also showed relatively high scores for toxicity (M = 0.555) and insult (M = 0.536), while classism/aporophobia showed elevated scores for toxicity (M = 0.539), insult (M = 0.521), and profanity (M = 0.470). These results suggest that Perspective API is more sensitive to categories involving overt verbal aggression, insult, and profanity than to frame-based categories such as invasion, socio-economic burden, or illegality.

Table 3. API Perspective Attribute Descriptions in the Hate messages

		TOXICITY	S_TOXICITY	ID_ATTACK	INSULT	PROFANITY	THREAT
N	Valid	5919	5919	5919	5919	5919	5919
	Missing	104	104	104	104	104	104
M		.526	.274	.308	.499	.421	.084
SD		.242	.233	.197	.252	.344	.142
Mín		.002	.0001	.001	.006	.008	.005
Máx		.982	.993	.952	1	.985	.929
Perc.	25	.338	.023	.113	.307	.079	.010
	50	.507	.232	.288	.513	.321	.020
	75	.736	.458	.460	.699	.784	.073

Source: Own elaboration

Taken together, these findings answer RQ3 by showing that messages classified as Islamophobic hate speech receive significantly higher Perspective API scores than non-hate messages. At the same time, the fact that most Islamophobic hate messages do not reach high-probability toxicity thresholds supports the argument that generic toxicity scores capture explicit verbal aggression more effectively than indirect, coded, or frame-based Islamophobic hostility.

RESULTS OF THE BERT CLASSIFIER

To evaluate automatic detection of Islamophobic hate speech, we fine-tuned a multilingual BERT-based classifier on the human-coded binary labels. The test set included 2,873 messages, with moderate class imbalance: 1,844 messages classified as Islamophobic hate and 1,029 messages classified as non-hate. Discarded messages were excluded from training and evaluation. Under a 0.5 decision threshold, the model achieved an overall accuracy of 0.89. For the Islamophobic hate class, precision was 0.97, recall was 0.85, and F1-score was 0.91. For the non-hate class, precision was 0.78, recall was 0.96, and F1-score was 0.86. Macro-averaged precision, recall, and F1-score were 0.88, 0.90, and 0.88, respectively. Weighted averages were 0.90 for precision, 0.89 for recall, and 0.89 for F1-score (see Table 4).

Table 4. Model evaluation metrics for hate vs. non-hate.

Class	Precision	Recall	F1-score	Support
0 (hate)	0.97	0.85	0.91	1844
1 (non_hate)	0.78	0.96	0.86	1029
Accuracy			0.89	2873
Macro avg	0.88	0.90	0.88	2873
Weighted avg	0.90	0.89	0.89	2873

Source: Own elaboration

These results indicate that the classifier performed strongly in detecting Islamophobic hate speech compared with human annotation. The model was especially conservative when assigning the hate label: it achieved very high precision for Islamophobic hate, indicating a low false-positive rate, but this came at the cost of some false negatives, as reflected in a recall of 0.85. Conversely, the non-hate class showed very high recall but lower precision. Accordingly, RQ4 is answered as follows: the multilingual BERT-based classifier achieved strong out-of-sample performance in detecting Islamophobic hate speech, with high overall accuracy and particularly high precision for the hate class. This profile may be useful for monitoring tasks in which avoiding false accusations of hate speech is especially important. However, if the goal is to reduce missed hate cases, future implementations could calibrate the decision threshold using validation data and precision-recall curves or explore class weighting and additional balancing strategies.

DISCUSSION

This study examined Islamophobic hate speech in Spanish-language tweets about Islam and Muslims in Spain between 2010 and 2022. By combining human annotation, exploratory discursive categorization, Perspective API toxicity scores, and a multilingual BERT-based classifier, the article contributes to the analysis of Islamophobic hate speech in this corpus and to the discussion of methodological challenges involved in detecting it automatically. The results are consistent with the argument that Islamophobic hate speech cannot be reduced to overt toxicity. Some messages contain explicit insults, profanity, threats, or direct derogation, but others operate through frame-based forms of hostility that associate Islam and Muslims with invasion, cultural threat, criminality, racialization, gendered oppression, or socio-economic burden. This pattern aligns with previous research showing that anti-Muslim hostility may be expressed through both direct abuse and more indirect or contextual forms of Islamophobic

discourse (Awan, 2014; Evolvi, 2018; Vidgen & Yasseri, 2020). The descriptive results show that 28.2% of the corpus was classified as Islamophobic hate speech. This proportion indicates that, within the keyword-based Twitter/X corpus analyzed, anti-Muslim hostility represented a substantial part of the conversation captured by the data collection strategy. Longitudinally, the highest number of Islamophobic hate messages appeared in 2021, followed by 2020 and 2022. This concentration suggests that the final years of the period studied were especially visible in the corpus, although the study does not claim that Twitter/X activity directly reflects public opinion in Spain. Rather, the corpus should be understood as a large-scale sample of public platform discourse about Islam and Muslims.

The analysis of exploratory subcategories provides further insight into the internal variation of Islamophobic hate speech. The most frequent pattern was explicit hostility, followed by invasion, cultural threat, criminality, race/ethnicity, and socio-economic burden. These findings are consistent with studies showing that Islamophobic discourse often combines direct verbal aggression with broader frames that construct Muslims as threatening, foreign, inferior, incompatible, or undeserving of equal belonging (Meer, 2013; Taras, 2013; Zempi & Awan, 2019). Importantly, the subcategories were not mutually exclusive. A single message could simultaneously represent Muslims as invaders, criminals, welfare burdens, and threats to Western values. The co-occurrence analysis therefore suggests that, in this corpus, Islamophobic discourse often operates through combinations of frames rather than through isolated themes.

Among these frames, invasion appears as a prominent pattern. It was one of the most frequent categories and one of the most common in co-occurrence with race/ethnicity, socio-economic burden, cultural threat, criminality, and illegality. This suggests that invasion may function, in this corpus, as a flexible interpretive frame capable of linking demographic fear, territorial anxiety, cultural replacement, and security concerns. This interpretation is consistent with research on invasion metaphors and anti-refugee or anti-Muslim discourse, where migration and religious diversity are represented as threats to national continuity or social cohesion (Gualda & Rebollo, 2020; Gorman & Culcasi, 2020). Race/ethnicity also plays an important role, as it points to the racialization of religion: Islam is treated not only as a religious identity, but as a visible, inherited, and culturally fixed marker associated with Arabness, Moroccanness, foreignness, migration status, skin colour, or ethnic origin. In this sense, the results are consistent with approaches that conceptualize Islamophobia as a form of anti-Muslim racism rather than as religious intolerance alone (Meer, 2013; Meer & Modood, 2009; Taras, 2013).

The findings also clarify the relationship between Islamophobic hate speech and toxicity. Perspective API scores were significantly higher among messages classified as Islamophobic hate speech than among non-hate messages across all

measured attributes. This indicates that human-coded Islamophobic hate speech often overlaps with toxicity-related language. However, the descriptive distribution of scores shows that most Islamophobic hate messages did not reach high-probability toxicity thresholds. Only the upper quartile reached or approached the operational threshold of 0.70 for toxicity, insult, and profanity. In this corpus, therefore, Perspective API scores appear more aligned with explicit verbal aggression than with coded, indirect, or frame-based hostility. This interpretation is consistent with the distinction between toxicity as an operational moderation category and hate speech as a target-specific form of discriminatory discourse (Wulczyn et al., 2017; Fortuna & Nunes, 2018; Schmidt & Wiegand, 2017).

This does not mean that Perspective API is ineffective. Rather, it suggests that generic toxicity scores should not be treated as stand-alone detectors of Islamophobic hate speech. Perspective API is useful for measuring the degree to which Islamophobic messages resemble language annotated as toxic, insulting, threatening, or identity-attacking. However, it may be less sensitive to hostile meanings that depend on context, racialization, securitization, cultural threat, or coded associations. The analysis by subcategory reinforces this interpretation. Explicit hostility obtained the highest mean scores for toxicity, insult, and profanity, followed by categories such as race/ethnicity and classism/aporophobia. By contrast, frame-based categories such as invasion, socio-economic burden, illegality, or cultural threat tended to produce lower toxicity scores, even though they were part of the human-coded Islamophobic hate subset. This pattern suggests that toxicity detection is more sensitive to surface-level verbal aggression than to discursive frames that legitimize exclusion without necessarily using abusive vocabulary.

The multilingual BERT-based classifier suggests a useful complementary approach for automatic detection. The model achieved an overall accuracy of 0.89 on the test set, with very high precision for the hate class (0.97) and strong F1 performance (0.91). This indicates that, in this test set, when the model classified a message as Islamophobic hate, it was rarely wrong. However, recall for the hate class was 0.85, meaning that some hateful messages were still missed. The model therefore behaves conservatively when assigning the hate label. This profile may be useful for monitoring contexts where false accusations of hate speech should be minimized, although future applications could calibrate the decision threshold to increase recall if the objective is to detect a larger proportion of hateful messages. These results are consistent with research showing the value of Transformer-based models for hate speech detection, while also confirming that model performance depends on the annotated corpus, the target category, and the evaluation setting (Devlin et al., 2019; Mozafari et al., 2020).

Taken together, the results support the value of combining human annotation, discursive categorization, toxicity scoring, and supervised classification. Human

annotation provides the reference classification for identifying Islamophobic hate speech. The exploratory subcategories describe internal variation within this discourse. Perspective API helps assess the overlap between Islamophobic hate and explicit toxic language. The multilingual BERT-based classifier shows that a target-specific model trained on human-coded data can perform well within this corpus. Each layer captures a different aspect of the phenomenon, and none should be treated as sufficient on its own.

CONCLUSION

The study shows that Islamophobic hate speech on Spanish Twitter/X is recurrent and diverse, combining explicit abuse with frames of invasion, cultural threat, criminality, racialization, gendered oppression, and socioeconomic burden. Generic toxicity measures capture only part of this phenomenon: Perspective API scores were higher for human-coded hate, yet many messages remained below high-probability thresholds. A multilingual BERT classifier performed strongly but assigned the hate label conservatively. These findings support context-sensitive definitions, reliable human annotation, and target-specific models. Future research should test this approach across platforms, multimodal content, and offline events.

Several limitations apply to this study. Twitter/X users are not representative of Spanish society, and platform features may shape the discourse observed. Keyword-based collection can miss-coded, visual, or unlisted content and retrieve irrelevant messages. Although binary annotation showed acceptable reliability, the exploratory subcategories were less reliable and should be treated descriptively rather than as a definitive taxonomy. Finally, Perspective API measures general toxicity rather than Islamophobic hate and may overlook contextual or frame-based hostility. The BERT classifier was trained and evaluated on this corpus, so its generalizability to other platforms, periods, countries, and varieties of Spanish remains to be tested.

FUNDING

This research was developed within the framework of the project "Evaluando Campañas contra el Odio" (ECO), funded by the European Union through the Citizens, Equality, Rights and Values Programme (CERV-2022-EQUAL).

REFERENCES

Alcántara-Plá, M., & Ruiz-Sánchez, A. (2017). The framing of Muslims on the Spanish Internet. *Lodz Papers in Pragmatics, 13*(2), 261–283. https://doi.org/10.1515/lpp-2017-0013

Awan, I. (2014). Islamophobia and Twitter: A typology of online hate against Muslims on social media. *Policy & Internet, 6*(2), 133–150. https://doi.org/10.1002/1944-2866.POI364

Awan, I., & Zempi, I. (2016). The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts. *Aggression and Violent Behavior, 27*, 1–8. https://doi.org/10.1016/j.avb.2016.02.001

Baker, P., Gabrielatos, C., & McEnery, T. (2013). *Discourse analysis and media attitudes: The representation of Islam in the British press*. Cambridge University Press.

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet, 7*(2), 223–242. https://doi.org/10.1002/poi3.85

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2023). *Spanish pre-trained BERT model and evaluation data* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2308.02976

Cockbain, E., & Tufail, W. (2020). Failing victims, fuelling hate: Challenging the harms of the "Muslim grooming gangs" narrative. *Race & Class, 61*(3), 3–32. https://doi.org/10.1177/0306396819895727

Comisión Islámica de España. (2022, March 14). *Estudio demográfico de la población musulmana* [Demographic study of the Muslim population]. https://comisionislamica.org/2022/03/14/estudio-demografico-de-la-poblacion-musulmana/

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Elahi, F., & Khan, O. (Eds.). (2017). *Islamophobia: Still a challenge for us all*. Runnymede Trust.

Evolvi, G. (2018). Hate in a tweet: Exploring internet-based Islamophobic discourses. *Religions, 9*(10), Article 307. https://doi.org/10.3390/rel9100307

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys, 51*(4), Article 85. https://doi.org/10.1145/3232676

Fuentes-Lara, C., & Arcila-Calderón, C. A. (2023). El discurso de odio islamófobo en las redes sociales: Un análisis de las actitudes ante la islamofobia en Twitter [Islamophobic hate speech on social media: An analysis of attitudes toward Islamophobia on Twitter]. *Revista Mediterránea de Comunicación, 14*(1), 225–240. https://doi.org/10.14198/MEDCOM.23044

García Sanjuán, A. (2017). La persistencia del discurso nacionalcatólico sobre el medievo peninsular en la historiografía española actual [The persistence of National-Catholic discourse on medieval Iberia in current Spanish historiography]. *Historiografías*, (12), 132–153. https://doi.org/10.26754/ojs_historiografias/hrht.2016122367

García Sanjuán, A. (2020). Weaponizing historical knowledge: The notion of Reconquista in Spanish nationalism. *Imago Temporis: Medium Aevum, 14*, 133–162. https://doi.org/10.21001/itma.2020.14.04

Gómez Godino, D. (2021). Islamofobia: La construcción social de un prejuicio y su abordaje educativo. Revisión teórico-crítica y estado de la cuestión [Islamophobia: The social construction of prejudice and its educational approach: A theoretical-critical review and state of affairs]. *RES: Revista de Educación Social*, (32), 307–327. https://eduso.net/res/revista/32/miscelanea/islamofobia-la-construccion-social-de-un-prejuicio-y-su-abordaje-educativo-revision-teorico-critica-y-estado-de-la-cuestion

González-Baquero, W., Amores, J. J., & Arcila-Calderón, C. A. (2023). The conversation around Islam on Twitter: Topic modeling and sentiment analysis of tweets about the Muslim community in Spain since 2015. *Religions, 14*(6), Article 724. https://doi.org/10.3390/rel14060724

Gorman, C. S., & Culcasi, K. (2021). Invasion and colonization: Islamophobia and anti-refugee sentiment in West Virginia. *Environment and Planning C: Politics and Space, 39*(1), 168–183. https://doi.org/10.1177/2399654420943897

Gualda, E., & Rebollo, C. (2020). Big data y Twitter para el estudio de procesos migratorios: Métodos, técnicas de investigación y software [Big data and Twitter for the study of migration processes: Methods, research techniques, and software]. *EMPIRIA: Revista de Metodología de Ciencias Sociales*, (46), 147–177. https://doi.org/10.5944/empiria.46.2020.26970

Hopkins, P. (2016). Gendering Islamophobia, racism and White supremacy: Gendered violence against those who look Muslim. *Dialogues in Human Geography, 6*(2), 186–189. https://doi.org/10.1177/2043820616655010

Jigsaw. (n.d.). *Perspective API*. https://perspectiveapi.com/

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). SAGE.

Lees, A., Tran, V. Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., & Vasserman, L. (2022). *A new generation of Perspective API: Efficient multilingual character-level transformers* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2202.11176

Meer, N. (2013). Racialization and religion: Race, culture and difference in the study of antisemitism and Islamophobia. *Ethnic and Racial Studies, 36*(3), 385–398. https://doi.org/10.1080/01419870.2013.734392

Meer, N., & Modood, T. (2009). Refutations of racism in the "Muslim question." *Patterns of Prejudice, 43*(3–4), 335–354. https://doi.org/10.1080/00313220903109250

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on BERT model. *PLOS ONE, 15*(8), Article e0237861. https://doi.org/10.1371/journal.pone.0237861

Observatorio Andalusí. (2023). *Estudio demográfico de la población musulmana: Explotación estadística del censo de ciudadanos musulmanes en España referido a fecha 31/12/2022* [Demographic study of the Muslim population: Statistical analysis of the census of Muslim citizens in Spain as of December 31, 2022]. Unión de Comunidades Islámicas de España. https://ucide.org/wp-content/uploads/2023/03/estademograf22.pdf

Perry, B. (2014). Gendered Islamophobia: Hate crime against Muslim women. *Social Identities, 20*(1), 74–89. https://doi.org/10.1080/13504630.2013.864467

Ríos Saloma, M. F. (2008). La Reconquista: Génesis de un mito historiográfico [The Reconquista: Genesis of a historiographical myth]. *Historia y Grafía*, (30), 191–216.

Saeed, T. (2016). *Islamophobia and securitization: Religion, ethnicity and the female voice*. Palgrave Macmillan. https://doi.org/10.1007/978-3-319-32664-6

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In L.-W. Ku & C.-T. Li (Eds.), *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1–10). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1101

Taras, R. (2013). "Islamophobia never stands still": Race, religion, and culture. *Ethnic and Racial Studies, 36*(3), 417–433. https://doi.org/10.1080/01419870.2013.734388

Vidgen, B., & Yasseri, T. (2020). Detecting weak and strong Islamophobic hate speech on social media. *Journal of Information Technology & Politics, 17*(1), 66–78. https://doi.org/10.1080/19331681.2019.1702607

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In R. Barrett, R. Cummings, E. Agichtein, & E. Gabrilovich (Eds.), *Proceedings of the 26th International Conference on World Wide Web* (pp. 1391–1399). Association for Computing Machinery. https://doi.org/10.1145/3038912.3052591

Zamora Medina, R., Garrido Clemente, P., & Sánchez Martínez, J. (2021). Análisis del discurso de odio sobre la islamofobia en Twitter y su repercusión social en el caso de la campaña "Quítale las etiquetas al velo" [Analysis of hate speech involving Islamophobia on Twitter and its social repercussion in the case of the campaign "Remove the labels from the veil"]. *Anàlisi: Quaderns de Comunicació i Cultura*, (65), 1–19. https://doi.org/10.5565/rev/analisi.3383

Zempi, I., & Awan, I. (Eds.). (2019). *The Routledge international handbook of Islamophobia*. Routledge.

Zempi, I., & Chakraborti, N. (2012). The veil under attack: Gendered dimensions of Islamophobic victimization. *International Review of Victimology, 18*(3), 269–284. https://doi.org/10.1177/0269758012446983

Zempi, I., & Chakraborti, N. (2015). *Islamophobia, victimisation and the veil*. Palgrave Macmillan. https://doi.org/10.1057/9781137356154

APPENDIX

OPERATIONAL DEFINITIONS OF EXPLORATORY SUBCATEGORIES OF ISLAMOPHOBIC HATE SPEECH

This appendix provides the operational definitions used to identify the exploratory subcategories within messages previously classified as Islamophobic hate speech. These categories were coded as non-mutually exclusive labels. Therefore, a single message could be assigned to more than one subcategory when it articulated multiple forms of anti-Muslim hostility. The examples below are paraphrased and do not reproduce original user-generated content.

Exploratory subcategory	Operational definition	Main coding indicators	Paraphrased example
Socio-economic burden	Messages that represent Muslims or Muslim-coded migrants as an unfair economic cost, a burden on public resources, or undeserving beneficiaries of welfare, housing, employment, education, subsidies, or public assistance.	References to welfare abuse, public spending, jobs supposedly taken from natives, housing, or aid allegedly given preferentially to Muslims or Muslim-coded migrants, or claims that Muslims should not receive rights or resources.	The message claims that Muslim migrants receive public benefits while native citizens are neglected.
Criminality	Messages that associate Muslims or Muslim-coded people with crime, delinquency, theft, sexual violence, assault, disorder, drug trafficking, or everyday insecurity.	Generalization from alleged individual crimes to Muslims as a group, the association between Muslim identity and criminal behavior, calls for expulsion or punishment based on presumed criminality.	The message presents Muslims as naturally linked to theft, violence, or public disorder.
Terrorism	Messages that associate Islam, Muslims, or Muslim-coded people with terrorism, jihadism, radicalization, extremist violence, ISIS/Daesh, Al-Qaeda, bombings, attacks, or generalized security threats.	Use of terms such as jihad, jihadist, terrorist, ISIS/Daesh, Al-Qaeda, radical Islam, or claims that Muslims as a group support or embody terrorism.	The message generalizes terrorist violence to Muslims as a whole or treats Islam as inherently terrorist.
Invasion	Messages that represent Muslim presence, migration, religious visibility, or demographic change as a territorial, cultural, political, or civilizational takeover.	Language of invasion, replacement, occupation, Islamization, mass arrival, expulsion, border control, "go back" claims, or calls to remove Muslims from Spain or Europe.	The message frames Muslim migration as an invasion that threatens Spain or Europe.

Exploratory subcategory	Operational definition	Main coding indicators	Paraphrased example
Illegality	Messages that associate Muslims or Muslim-coded migrants with irregular migration, lack of documents, unlawful presence, or illegitimate residence.	References to undocumented status, illegal entry, border crossing, deportation, repatriation, papers, or the presumption that Muslim-coded migrants are in the country illegally.	The message assumes that a Muslim-coded person is undocumented and should be deported.
Cultural threat	Messages that represent Islam or Muslims as incompatible with Spanish, European, Western, secular, liberal, democratic, Christian, or national values.	Claims that Islam cannot coexist with democracy, freedom, secularism, gender equality, national identity, Christian tradition, or "Western" ways of life.	The message claims that Islam is incompatible with European values and should not be allowed in public life.
Gender	Messages that mobilize women's rights, the veil, burka, hijab, sexuality, family norms, feminism, or gender equality to portray Islam or Muslims as inherently oppressive, patriarchal, violent, or incompatible with liberal values.	References to Muslim women as passive victims, the veil as proof of oppression, Muslim men as inherently sexist, or accusations that feminists ignore Muslim misogyny.	The message presents Muslim women only as oppressed subjects and uses that claim to attack Islam or Muslims as a group.
Race/ethnicity	Messages that racialize Muslimness by linking Islam or Muslim identity to Arabness, Moroccanness, foreignness, skin color, ethnicity, ancestry, nationality, or inherited group traits.	Use of ethnicized or racialized labels, claims that Muslims are a fixed racial or ethnic group, references to blood, origin, skin color, Arabness, Moroccanness, or foreignness as markers of Muslim identity.	The message treats Muslims as a fixed foreign or ethnic group that can never belong to Spain.
Classism/ aporophobia	Messages that associate Muslims or Muslim-coded people with poverty, dirtiness, social inferiority, marginality, ignorance, lack of education, precarious work, or stigmatized lower-class status.	References to Muslims as poor, dirty, backward, uncivilized, socially inferior, undesirable, or associated with stigmatized occupations or spaces.	The message depicts Muslim-coded people as dirty, poor, socially inferior, or unworthy of equal treatment.
Explicit hostility	Messages that express direct verbal aggression against Islam, Muslims, or Muslim-coded people through insults, profanity, dehumanization, threats, violent wishes, or open calls for exclusion.	Direct insults, abusive language, threats, dehumanizing labels, calls to expel, harm, eliminate, silence, or exclude Muslims or Islam.	The message directly insults Muslims as a group and expresses a desire for their exclusion or harm.

Coding note: The subcategories are exploratory and descriptive. They should not be interpreted as a definitive taxonomy of Islamophobia. They were used to identify recurring discursive patterns within the subset of messages already classified as Islamophobic hate speech. Because the categories are non-mutually exclusive, coders could assign multiple labels to the same message when more than one frame or form of hostility was present.