Hate Speech in Times of Crises: a Cross-Disciplinary Analysis of Online Xenophobia in Greece

Maria Pontiki[1,2], **Vasiliki Georgiadou**[2,3], **Lamprini Rori**[4], **Maria Gavriilidou**[1]

[1]Athena Research Center, [2]Panteion University of Social and Political Sciences, [3]National Center for Social Research, [4]National and Kapodistrian University of Athens

mpontiki@athenarc.gr, vgeorg@panteion.gr, lrori@pspa.uoa.gr, maria@athenarc.gr

OFFENSIVE CONTENT WARNING: This report contains examples of hateful content. This is strictly for the purposes of enabling this research, and we have sought to minimize the number of examples where possible. Please be aware that this content could be offensive and cause you distress.

Abstract

Bridging NLP with political science, this paper examines both the potential and the limitations of a computational hate speech detection method in addressing real-world questions. Using Greece as a case study, we analyze over 4 million tweets from 2015 to 2022—a period marked by economic, refugee, foreign policy, and pandemic crises. The analysis of false positives highlights the challenges of accurately detecting different types of verbal attacks across various targets and timeframes. In addition, the analysis of true positives reveals distinct linguistic patterns that reinforce populist narratives, polarization and hostility. By situating these findings within their socio-political context, we provide insights into how hate speech manifests online in response to real-world crises.

1 Introduction

Although hate speech predates the digital era—having historically served as a means of reinforcing stereotypes and dehumanizing individuals or groups, often leading to discrimination, marginalization, and, in extreme cases, genocide (Graham and Smith, 2024)—its manifestation in online spaces has significantly amplified both its reach and impact, fueling polarization and contributing to the erosion of democratic discourse (Sunstein, 2018). Social media platforms, by affording users a degree of anonymity, often reduce accountability, thereby enabling the open expression of prejudiced views (Mondal et al., 2017) and hostile discourse, which in turn entrenches social and ideological divisions. The normalization of hate speech by influential figures (e.g., political leaders) has further legitimized hateful rhetoric, embedding it within mainstream discourse and leading to far-reaching societal consequences, particularly in polarized or crisis-driven contexts; online platforms can serve as catalysts for offline violence, as exemplified by the January 6th attack on the U.S. Capitol (Lupu et al., 2023). Similarly, the surge in online hate speech in Europe has been linked to the refugee crisis (Ross et al., 2016) and has coincided with a documented rise in anti-refugee hate crimes (Müller and Schwarz, 2020). During periods of crisis or perceived threat, there is a common tendency to scapegoat outgroups blaming them for societal problems and uncertainties which serve to activate and amplify stereotypes and prejudices (Kim et al., 2016; Wodak, 2015). This pattern was evident during the COVID-19 pandemic, which saw a surge in anti-Asian sentiment on social media platforms (Ghenai et al., 2025). More recently, the Israel–Hamas conflict has triggered a global rise in both Islamophobic and anti-Semitic narratives, reflecting the reactivation of deep-seated prejudices (Rose and Matlach, 2024).

NLP research has made significant progress in detecting various aspects of hateful content (e.g., Jurgens et al., 2019; Sap et al., 2020; Caselli et al., 2021; ElSherief et al., 2021;

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 241–253

Yoder et al., 2022) laying the groundwork for targeted interventions such as moderation, debiasing, and counter-speech (Hee et al., 2024). Recent advances in LLMs can improve performance and interpretability, enabling more nuanced hate speech analysis (e.g., Yang et al., 2023; Wang et al., 2023). However, real-world application of hate speech detection remains challenging. Unlike many other NLP tasks, it is culturally sensitive (Schmidt and Wiegand, 2017), as hate speech is deeply embedded in the sociocultural contexts in which it emerges (Warner and Hirschberg, 2012; Kennedy et al., 2022). Ethically, it requires careful consideration of the risks involved in labeling communicative practices as hate speech, particularly for the communities implicated in such research (Gagliardone et al., 2022). Therefore, models must be rigorously validated to ensure they accurately capture complex social issues—especially since false-positive errors can inadvertently censor online speech and further marginalize specific groups (Yang et al., 2023). A further challenge arises from the differing goals of NLP and social sciences (McGillivray et al., 2020). NLP focuses on developing new computational systems or improving existing ones, so it is important that these are evaluated on standard datasets using reproducible methods which, however, are optimized for restricted datasets and languages –most of them in English (Arango et al., 2022). Social scientists seek insights stemming from research questions that are formulated using constructs relevant to their fields and not in terms of NLP. This divergence highlights the need for interdisciplinary approaches that tailor computational tools to domain-specific questions and real-world complexities (McGillivray et al., 2020).

In this paper, we apply a rule-based NLP method on real-word questions in the context of political science research. Taking Greece as a case study, we present a large-scale yet fine-grained analysis of online verbal aggression (VA) targeting key groups: Albanians and Pakistanis (the largest migrant communities), Muslims and Jews (significant religious and ethnic minorities), and migrants and refugees (both statuses of foreign population). Using a publicly available VA analysis tool for the Greek language (Pontiki et al., 2018; Pontiki et al., 2020) we analyze over 4 million tweets posted on Greek Twitter/X between 2015 and 2022. Greece is one of the few countries which experienced the concomitant turmoil of four different crises (Rori, 2021): a financial, a refugee, a foreign policy crisis (e.g., tensions with neighboring Turkey), and the COVID-19 pandemic crisis. In this context of polycrisis, beyond assessing the tool's performance, our study addresses two key research questions (RQs):

RQ1: Which groups, situated within specific socio-political contexts, were the primary targets of hate speech on Greek Twitter/X during the examined period(s)?

RQ2: Are there target-specific linguistic patterns, prejudices, or stereotypes?

The contribution of our work is two-fold: first, we provide both quantitative and qualitative analyses of false positives, highlighting the challenges of accurately detecting different types of verbal attacks across various targets and timeframes with regard to domain-specific RQs. Second, our analysis of the fluctuation and content of detected verbal attacks uncovers key linguistic patterns that reinforce populist narratives, polarization and hostility in Greek online discourse. By contextualizing hate speech within real-world grievances and socio-political tensions, our findings illuminate how hate discursive patterns manifest, evolve, and interact with broader crises. Furthermore, our findings can provide a framework for informed countermeasures and deeper exploration of the link between online aggression and offline political violence, particularly in times of crises.

2 Background

Most NLP approaches treat hate speech detection as a binary (e.g., Djuric et al., 2015) or multiclass classification task (e.g., Waseem and Hovy, 2016), typically relying on explicit linguistic cues. Recent research has shifted toward addressing implicit hate speech (Kennedy et al., 2018; Sap et al., 2020), identifying different types of verbal attacks (Pontiki et al., 2018; ElSherief et al., 2021), analyzing group-specific targeting (Kennedy et al., 2018; Pontiki et al., 2018; Yoder et al.,

2022), and using free-text annotations to better capture the pragmatic implications of hateful messages (Sap et al., 2020; ElSherief et al., 2021). Computational approaches focusing on the Greek language include the development and evaluation of classifiers for tasks such as offensive tweet detection (Pitenis et al., 2020) and abusive content moderation in user comments (Pavlopoulos et al., 2017). Perifanos and Goutsos (2021) proposed a multimodal approach that combines Computer Vision and NLP to detect abusive contexts in tweets targeting refugees and migrants. Pontiki et al. (2018; 2020) employed a linguistically-informed rule-based framework to identify and categorize specific forms of VA—such as criticism, swearing, and calls for ousting—against predefined minority and migrant groups on Twitter. Arcila-Calderón et al. (2022) developed both shallow and deep learning models for detecting online anti-immigration hate speech in Spanish, Greek and Italian. Their models are incorporated within the PHARM project interface (Vrysis et al. 2021; Kotsakis et al., 2023) developed with the goal to monitor and model hate speech against refugees and migrants in Greece, Italy, and Spain.

Political science literature on xenophobia examines primarily fears and hostility towards 'foreigners', focusing on the motives and mechanisms of their mobilization mostly by populist-radical and far-right parties (Mudde, 2007; Georgiadou et al., 2018). It also explores the institutionalization of xenophobia through public policies on immigration and securitization (Lahav and Messina, 2024). A notable gap persists in the academic discourse: while much attention has been paid to persistent forms of "non-violent discrimination" (Del Fabbro, 1995), less attention has been devoted to xenophobia "as a violent practice" (Galariotis et al., 2017).

Xenophobic attitudes in Greece are fluid and context-dependent, historically targeting socio-economically marginalized–during the period first entered the country–migrants like Albanians and Pakistanis, while directing hostility at perceived "dominant" outgroups such as Jews (Galariotis et al., 2017). Although political motivated violence is grounded in the history and culture of Greece, in constant presence since the transition to democracy in 1974 (Rori and Georgiadou, 2023), anti-immigrant and xenophobic violence marks unprecedented levels during the financial and the refugee crises. The rise of the neo-Nazi party Golden Dawn (GD) in the context of the economic crisis normalized anti-Semitic and xenophobic discourse in mainstream politics (Georgiadou, 2020). GD managed to emerge as the third largest party in the national elections of 2015, despite an ongoing judicial investigation into its involvement in violent attacks mostly against migrants and refugees. The rise of GD, which has not entered the national parliament since 2019, played a central role in mobilizing hate narratives and coordinating street-level violent attacks against 'foreigners' and left-wing activists, particularly during the 2015 refugee influx in the midst of economic crisis (Dinas et al., 2016). During this period, alarmist coverage in traditional and social media amplified economic and cultural fears, reinforcing exclusionary attitudes toward refugees and other minorities. An analysis of 504 incidents of far right violence registered from 2008 to 2019 revealed high-escalation attacks primarily targeting humans, highly correlated with the fear of economic losses, sensitive to increases in immigration flows and fuelled by the representation of extremist parties in parliament (Rori et al., 2022).

3 Data Collection and Processing

3.1 Data Collection

For each Target Group (TG) relevant tweets were retrieved using related keywords (i.e. Albanian, Pakistani, Muslim, Islam, Jew, immigrant, refugee). 4,386,501 tweets were retrieved through the queries, covering the period from 2015 to 2022. As illustrated in Figure 1, the total volume of tweets is highest for refugees (1,568,308) and migrants (1,359,610). Migration-related discourse dominates Twitter discussions in our datasets, likely influenced by political and social events. The spike in tweet volume for these groups in 2019 and 2020 aligns with increased tensions in Europe regarding migration policies and border conflicts (e.g., Greece-Turkey border crisis), while the decline in 2021 and 2022 may indicate shifting focus towards other crises (e.g., COVID-19, the Ukraine war). Muslims

(493,013 tweets) also feature prominently, suggesting significant online discourse related to Islam and related socio-political issues. The lowest volume is observed for Jews (170,928) and Pakistanis (313,021), indicating comparatively lower levels of public discussion about these groups. However, as demonstrated by our findings in Section 5—in line with previous research (Pontiki et al., 2018)—lower mention volumes do not necessarily equate to reduced levels of hate speech.

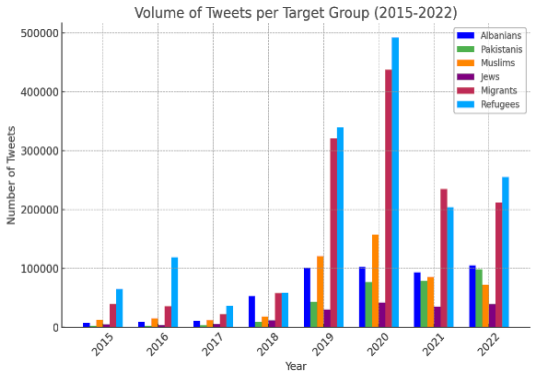

Figure 1: Total amount of tweets retrieved per TG and year.

3.2 Data Processing

The collections of tweets were processed using the GR_VA_Analyzer[1] web service that is freely accessible through the CLARIN:EL infrastructure (Gavriilidou et al., 2024). The workflow consists of the following processing steps: sentence splitting, tokenization, POS tagging, lemmatization, and VA detection and classification. The preprocessing is performed using the ILSP suite of NLP tools for the Greek language (Prokopidis et al., 2011). The VA analysis tool is a rule-based method that comprises a variety of lexical resources and linguistic patterns for the detection of explicit verbal attacks against a variety of targets related to xenophobia (Pontiki et al., 2018; Pontiki et al., 2020) and political violence (Pontiki et al., 2022). In particular, the method is designed to capture the following types of verbal attacks (Pontiki, 2019): **Criticism** (disapproval or negative evaluations of specific attributes of the target), **Swearing** (taboo or profane language to degrade or insult the target), **Irony** (sarcastic, humoristic, or satirical messages), **Ousting** (intentions or calls for ouster), and **Physical Abuse** (intentions or calls for physical violence/harm or physical extinction).

The VA analyzer is implemented as a cascade of Finite State Transducers using JAPE grammars (Cunningham et al., 2000) within the GATE framework (Cunningham et al., 2002). In the initial phase, the analyzer identifies candidate verbal attacks and potential targets based on predefined lexical resources. Subsequently, the grammars assess these candidates to determine which ones constitute valid verbal attacks and targets. The grammar system follows a multi-phase algorithmic structure, where each phase consists of several modules containing contextual lexico-syntactic patterns. These patterns act as templates for rule generation, analyzing the local context around each candidate using primarily shallow syntactic relations. For each identified attack, the tool returns a structured tuple containing specific annotation types[2]. For example, the output for the Tweet "*Rub out some migrants until they stop coming*" is: `[TG_id= "TG5", TG_ evidence= "migrants", VA_type= "VAM2B", VA_evidence= "Rub out"]`.

The tool detected a total of 365,669 verbal attacks against the six TGs under examination in our datasets. The output was recorded in a database along with other Twitter metadata for each tweet (e.g. timestamp). To check the validity of the detected verbal attacks for each TG, we filtered the attacks by year and by their type. Given that our goal is to address specific RQs and also given the size of our datasets, we did not go through an exhaustive inspection of all the results. Instead, we explored thousands of randomly selected samples with the aim to check the reliability (in terms of precision) of the extracted hate speech trends for our targets in the period under examination. Based on this qualitative evaluation, the final database, having been revised for the removal of false positives, contains a total of 310,587 verbal attacks.

[1] https://inventory.clarin.gr/tool-service/1241

[2] https://inventory.clarin.gr/storage/media/1bb0b8da4ce1421ab228a60f86fecff6_u37_GR_VA_Analyzer_AnnotationTypes.txt

4 Analysis of False Positives

Table 1 presents the approximate precision of the tool, though the actual precision is probably lower since not all results were manually inspected. Despite the limitations (further discussed in the respective section), the results provide insights into the tool's effectiveness in identifying different types of verbal attacks across various targets and time periods. The high precision observed for Muslims, Albanians, Pakistanis, and migrants may be due to the more explicit nature of attacks targeting these groups in our datasets—often featuring overtly negative portrayals or derogatory language that the specific tool is better equipped to identify. In contrast, tweets targeting Jews and refugees required extensive manual review due to a high number of false positives in our samples. This discrepancy aligns with prior research indicating that **hate speech varies significantly by the identities it targets** (e.g., Yoder et al., 2022).

	verbal attacks	false positives	approx. precision
Albanians	35.813	2669	92.55%
Pakistanis	30.692	1650	94.62%
Muslims	50.124	1105	97.80%
Jews	17.669	8860	49.86%
Migrants	178.962	16.360	90.86%
Refugees	52.271	23.200	55.61%

Table 1: Approx. precision per TG.

We also calculated approx. precision per year for the three TGs with the lowest overall precision (Fig. 2). The precision of the tool in detecting verbal attacks targeting Jews fluctuates significantly, peaking at 79% in 2016 and dropping to its lowest point (41.8%) in 2020. The highest precision (2015–2017) coincides with the period when GD was a major source of explicit antisemitic discourse contributing to the activation of deep-seated prejudices and dormant biases (Antoniou et al., 2020). During this time, GD openly promoted Holocaust denial, conspiracy theories rooted in historical and contemporary antisemitism, and incited violence, including vandalism of Jewish cemeteries and synagogues (Galariotis et al., 2017). The decline in precision from 2018 onward can be partly attributed to the prevalence of ironic tweets that mimic antisemitic rhetoric to criticize antisemitism.

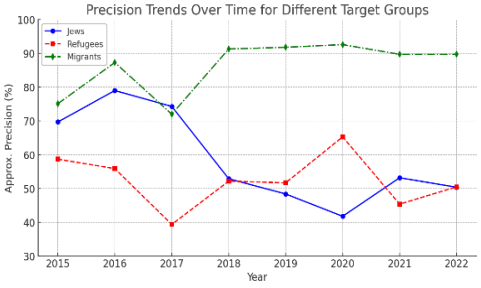

Figure 2: Approx. precision trends over time.

For example, messages blaming Jews for creating or spreading COVID-19 are often satirical, aiming to expose and condemn antisemitic conspiracy theories. The tool misclassified these tweets as genuine attacks, overlooking their underlying sarcastic intent. **Irony is frequently used to mock far-right ideologies by echoing their rhetoric without endorsing these views**. For example, in the tweet "*Abortion is murder!!!! Well calm down, we'll only kill Jews, blacks and commies*" the tool detected the explicit call for physical abuse targeting Jews. However, this tweet aims to criticize perceived hypocrisy in moral or political arguments, particularly within far-right ideologies. We removed such tweets from our database since the detected calls for physical abuse against our TGs are not true in terms of the users' intent.

However, even when framed ironically, references to "Jews, blacks, and commies" remain inflammatory, evoking historical atrocities such as the Holocaust, racial violence, and Cold War hostilities. The invocation of violence against marginalized groups in such messages perpetuates harmful stereotypes and may be interpreted as hate speech, depending on context and perspective. A significant portion of false positives in our datasets stems from such ambiguous or contextually complex content, raising the question: how should an NLP method handle such cases? The distinction between abusive and offensive language remains a topic of debate within the NLP community (e.g., Founta et al., 2018; Zampieri et al., 2019; Vidgen et al., 2019; Caselli et al., 2020). Definitions of offensive content are often shaped by the sensitivities of online audiences, which, in turn, influence annotation practices and dataset construction (Vidgen et al., 2019). These definitions tend to emphasize surface- level features—such as the presence of

profanity—or the emotional response of the reader, rather than the producer's intent (Caselli et al., 2020). Detecting and analyzing online hate speech poses further complex conceptual, ethical, and methodological challenges (Gagliardone et al., 2022), that become even more pronounced when aggression and hate speech appear within discourse that ostensibly aims to combat hate, as illustrated in the example above. Recent literature in Critical Discourse Studies explores the fuzzy boundaries between racist and antiracist discourse, "*which originate in the hegemony of racist discourse and further normalize it*" (Archakis and Tsakona, 2024).

For refugees, precision ranges from 39.4% (2017) to 65.3% (2020). The low precision in 2017 coincides with Greece's ongoing struggle to manage the refugee crisis, exacerbated by deaths in overcrowded camps such as Moria on Lesvos. Fatalities due to hypothermia and inadequate living conditions in the harsh winter of early 2017 sparked widespread criticism of the Greek government. Similarly, in 2021 (45.4%), Greece faced scrutiny over reports of pushbacks, violence, and abuse by authorities, alongside multiple refugee fatalities from shipwrecks. The tool struggled to classify tweets reporting on these events. It often misinterpreted descriptions of violence against refugees as verbal attacks against them or incorrectly assigned attacks targeting other groups (e.g., coast guard personnel) to refugees. Hate speech detection for migrants consistently outperforms that for Jews and refugees, with precision ranging from 72% (2017) to 92.6% (2020). The drop in 2017 may be explained by the overlap in discourse surrounding migrants and refugees (Gabrielatos and Baker, 2008), highlighting the broader challenge of detecting nuanced variations in language use, particularly in highly politicized or emotionally charged discussions.

Another key limitation of the tool is that it does not account for **opinion holder identification**, a crucial component in hate speech detection (Chetty and Alathur, 2018). As a result, it captures attacks expressed by any actor during the period under examination e.g., in "*Czech President: "Send refugees to uninhabited Greek islands*"" the tool correctly identified the explicit ousting message targeting refugees. Such cases were considered "***out-of-scope true positives***" and were removed from the database, since our goal is to capture and monitor hate speech expressed in X by Greek users. This limitation produces **another significant set of false positives in tweets that contain explicit hate speech but quote historical figures or past events to highlight and condemn discrimination**. For example, the tweet: *"When interviewed, Brunner said, 'The Jews deserved to die. I have no regrets. If I had the chance, I would do it again.'"* directly quotes Alois Brunner, a Nazi official. While the tweet reproduces hate speech, its intent is to expose antisemitism rather than promote it. The tool classifies such content as aggression targeting Jews due to its explicit language. Similarly, historical references are used in tweets opposing discrimination against refugees e.g., tweets referencing the mass displacement of ethnic Greeks from Asia Minor after the Greco-Turkish War (1919–1922), illustrating how past refugees faced xenophobia when entering Greece, despite being Greek Orthodox.

In addition, **retweets and quoted tweets** complicate the analysis because they may include context or commentary that alters the original meaning. In general, the tool lacks the cultural and contextual understanding required to differentiate between hate speech expression and hate speech critique. To sum up, our analysis shows that **the distinction between false and true positives is not that straightforward, when computational online hate speech detection is examined in the context of domain-specific real-world questions as opposed to specific/restricted test datasets**. Furthermore, the yearly fluctuations of the precision scores suggest that **the performance is influenced by changes in the volume and nature of hate speech on Greek X, as well as by shifts in public discourse and events affecting specific TGs**.

Focusing on the types of verbal attacks, the evaluation of the results suggests that **swearing**, due to its explicit and overt nature, enhances the tool's ability to reliably detect offensive language patterns across all TGs. **Criticism** detection shows moderate precision for Jews and refugees, indicating that while it often follows direct linguistic structures, its

accurate identification is complicated by nuances in tone and intent that vary across TGs. **Irony** detection also demonstrates moderate precision, as expected with a rule-based approach. The most challenging categories were **calls for physical abuse** and **ousting messages**, as their linguistic patterns frequently overlap with neutral reports on displacement and refugee fatalities—issues that remain highly visible in Greece. This overlap reinforces keyword bias (De la Peña Sarracén and Rosso, 2023), leading to misclassifications when the tool fails to distinguish between objective reporting of violent incidents and actual verbal attacks.

5 Analysis of True Positives

The analysis of true positives includes the calculation of the VA rate (i.e., number of verbal attacks per total tweets) for each TG, enabling the identification of the primary targets of Twitter-based verbal attacks over the full period examined. We also perform a year-by-year analysis of VA rates to track their evolution over time and identify potential peak periods. Additionally, we investigate the fluctuation of individual VA types per TG to gain further insights into the variation of verbal attacks both temporally and within each group. Below, we present our findings in relation to the RQs; due to space limitations, most visualizations are included in Appendix A.

5.1 RQ1: Which groups, situated within specific socio-political contexts, were the primary targets of hate speech on Greek Twitter/X during the examined period(s)?

As illustrated in Figure 3, Migrants, Muslims and Pakistanis are the main targets of Twitter verbal attacks for the whole period under examination.

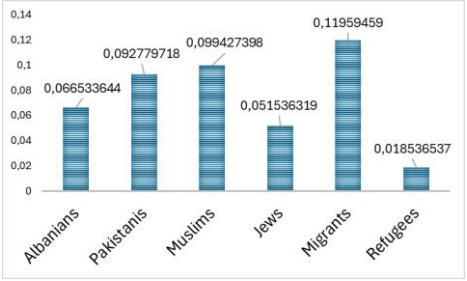

Figure 3: Overall VA rate per TG for the period 2015-2022.

In line with previous research (Pontiki et al., 2018; Pontiki et al., 2020), our results suggest that groups framed as *migrants* are more likely to be verbally attacked than those framed as *refugees*, likely due to the differing connotations and sociopolitical implications associated with these two lexicalizations.

The per year VA rates (Fig. 4) indicate a global increase of attacks from 2018 onwards. The increased rate of the attacks against migrants can be possibly attributed to the ongoing refugee crisis and mainly to the fact that the effect of this crisis has started to be tangible in Greek society, especially at the severely overcrowded camps on the islands. There might also be a noticeable time lag between the actual processes of events and the discursive articulation of them suggesting a delay in verbalization of them in public and social media discourses (Van Dijk, 1998; Wodak, 2015).

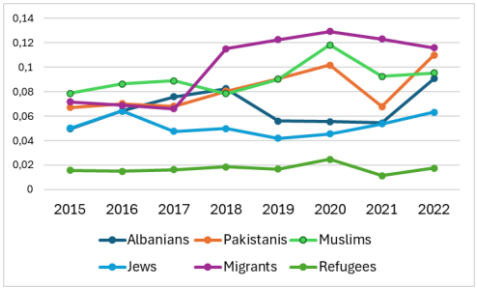

Figure 4: VA rate per year and TG.

As Greece officially exits the bailouts in the summer of 2018, increases in VA against migrants over the same period might as well reflect a shift in the prioritization of perceived scapegoats for grievances and backlashes. The highest peaks for most TGs are nonetheless observed in 2020, an evolution certainly fueled by a series of migration-related critical incidents which occurred in Evros in February and in Lesvos in March and September, all of which increased and sustained the salience of immigration in the public sphere, prompting the government to adopt a tougher stance than its predecessor (Rori, 2021). Greece was further placed in strict lockdown in March 2020. Due to the restrictions on physical contact during the COVID-19 pandemic, the internet came to function as the principal means of expression and communication.

5.2 RQ2: Are there target-specific linguistic patterns, prejudices, or stereotypes?

Verbal attacks expressing criticism constitute the main type of VA detected in our datasets, and are mostly directed against Muslims, Jews, Albanians and refugees.

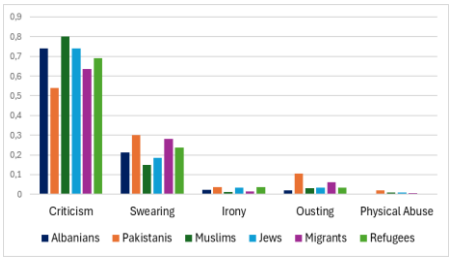

Figure 5: Overall VA type rates per TG for the period 2015-2022.

Pakistanis and migrants receive the most obscene messages. Pakistanis are mostly evaluated as inferior beings with derogatory morphological variations of their nationality name (e.g., *Pakistania*). The main recipients of ironic tweets are Pakistanis, Jews and refugees. Calls for ousting constitute the third most frequent type of VA targeting mostly Pakistanis and migrants. Pakistanis are also the main targets of messages calling for physical violence (Fig. 10), a finding which highlights the association of the previous dehumanizing discourse towards this specific TG with the calls for real-world violence against it.

Focusing on criticism (Fig. 6), the yearly distribution for each TG indicates a relatively stable frequency, highlighting it as the most consistently used linguistic strategy to target them throughout our datasets. Significant peaks and drops are observed only for Pakistanis and refugees. However, the decrease of criticism rates does not indicate a decrease of VA against them in general, but rather a shift in the VA type, with aggression moving from criticism towards swearing (Fig. 7) and irony (Fig. 8). Fluctuation of swearing rates is not only observed across time, but also within each TG. Compared to criticism, swearing is more emotional, driven by spontaneous reactions and strong feelings such as anger or contempt, e.g., as reactions/responses to news reporting crimes committed by Pakistanis. Ousting messages (Fig. 9) display two significant peaks for Pakistanis—in 2016 and 2020. The populist radical right party named Greek Solution, founded in 2016, launched a hardline anti-immigrant narrative, in which Greece shall be a "fortress" without migrants. Interestingly, the increase in 2020 relates to official online party discourse of the Greek Solution in 2020, whereas it is also associated with an ongoing crisis with Turkey during the same period (Rori, 2021). Another interesting peak is the one for Albanians in 2018, which according to the qualitative analysis is mostly related to the murder of Konstantinos Katsifas, a member of the Greek minority of Northern Epirus at the southern part of Albania, who was killed by the Albanian police. GD members were asking for mass deportations of Albanians. Finally, we can see a significant increase of ousting messages targeting refugees in 2021, triggered by online debates on toughening policies of other EU countries towards Syrian refugees.

The qualitative analysis of the attacks reveals that Albanians are mainly framed as *murderers* indicating a continuity of the so-called stereotype of the Balkanian criminal. Another frequent set of attacks entails the perception of Albanian nationalism and a particular enmity towards the Greek nation. Pakistanis are also frequently associated with crime (particularly sexual violence). The most frequent term in the construction of the image of Muslims is the word *fanatic*; the attacks against them are often lexicalized through evaluative and dysphemistic terms of insult or abuse to debase core Islamic values, practices, etc. indicating irrationalism, sexist behavior and fanaticism, and framing them as terrorists. Jews are mainly framed as *Greek haters* with the attacks against them entailing the perception of enmity towards the Greek nation and Christianity and blame attribution patterns (e.g., for the Greek crisis). This rhetoric and mobilization drew on a symbolic competition between perceived victimised ingroups and outgroups, wherein the targeting of Jews served to fuel antisemitic stances (Antoniou et al., 2020). The most frequent verbal attacks against migrants and refugees challenge their identity, framing them as *illegal*. This rhetoric aims to undermine their legitimacy, humanity, and right to asylum by questioning their motives, authenticity, and cultural compatibility. In this context, they are

frequently framed as economic, cultural and national security threats.

6 Conclusions

Analyzing over 4 million Greek tweets from 2015 to 2022—a period marked by overlapping crises—we assessed the effectiveness of a publicly available rule-based system in detecting different types of verbal attacks against various TGs. Our findings indicate that while the tool performs well in terms of precision for explicitly targeted groups in our datasets, it struggles with more contextually complex content, particularly tweets involving Jews and refugees. The high rate of false positives in these cases often results mainly from satire, irony, or quoted hate speech that aims to critique rather than endorse discriminatory views. Moreover, instances of aggression and hate speech embedded in discourse that ostensibly aims to combat hate further complicate the analysis.

Despite its limitations, the method offers valuable insights into the ways hate speech manifests online in Greece in response to real-world grievances and crises. Yearly fluctuations in precision and verbal attack rates reflect the impact of external socio-political developments and shifts in public discourse. Our research among others has shown the frequent association of the VA against migrants with criminality; that dehumanizing specific ethnic groups also renders those groups as targets for physical harm; that the bundled crises in the period under study scapegoated migrants and refugees for grievances and insecurities triggered by crises unrelated to them. Ultimately, our analysis underscores the need for interdisciplinary approaches that adapt computational tools to sociopolitical contexts, and incorporate human oversight—crucial for capturing nuance, intent, domain-specific and cultural specificities in language use, as well as the ethical ambiguities of hate speech detection.

Limitations

As discussed in previous sections, the actual precision of the VA analysis tool is likely lower—particularly for tweets targeting Albanians, Pakistanis, Muslims, and migrants—since not all retrieved results were manually reviewed, unlike those for Jews and refugees. We manually examined over 100,000 detected verbal attacks and removed 55,082 false positives, the majority of which pertained to the latter two groups. While qualitative analysis of true positives served as an additional validation step, some false positives may remain in the database. Moreover, due to the inclusion of retweets and quoted tweets in our datasets, many instances reflect repeated occurrences of the same verbal attack. A key limitation of our study is the lack of recall evaluation, which is expected to be moderate to low for certain TGs and time periods. Given our research focus and the size of our datasets (over 4 million tweets), our priority was on results reliability rather than exhaustiveness. We also acknowledge that, as a rule-based method, the tool may fail to detect implicit or ironic verbal attacks, potentially omitting a significant portion of hateful content. Furthermore, our keyword-based data collection method may have excluded tweets using alternative terms or emerging slurs. Finally, we recognize the possibility that some of the detected content originates from bots or fake accounts.

Acknowledgments

The work presented in this paper is supported by DeMoLiSH research project, implemented in the framework of H.F.R.I call “Basic research Financing (Horizontal support of all Sciences)” under the National Recovery and Resilience Plan “Greece 2.0” funded by the European Union – NextGenerationEU (H.F.R.I. Project Number: 15576).

Ethics Statement

In accordance with both the GDPR[3] and the Developer Policy of X[4], we have anonymized all personal and sensitive data included in the datasets under research. User identification information, such as username/handle and post ID have been deleted from the dataset. Verbatim expressions have been reproduced in this publication solely to support our claims.

[3] https://gdpr-info.eu/

[4] https://developer.x.com/en/developer-terms/policy#4-e

References

Georgios Antoniou, Elias Dinas, and Spyros Kosmidis. 2020. Collective victimhood and social prejudice: A post-Holocaust theory of anti-semitism. *Political Psychology* 41(5), 861-886.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2022. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, 105, 101584.

Argiris Archakis and Villy Tsakona. 2024. Antiracist and Racist Discourse as Antagonistic and Overlapping. In *Exploring the Ambivalence of Liquid Racism: In between Antiracist and Racist Discourse*, edited by Argiris Archakis and Villy Tsakona, 1–40. Amsterdam/Philadelphia: John Benjamins.

Carlos Arcila-Calderón, Javier J. Amores, Patricia Sánchez-Holgado, Lazaros Vrysis, Nikolaos Vryzas, and Martín Oller Alonso. 2022. How to detect online hate towards migrants and refugees? Developing and Evaluating a Classifier of Racist and Xenophobic Hate Speech Using Shallow and Deep Learning. *Sustainability*, *14*(20), 13094.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for Abusive Language Detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Association for Computational Linguistics.

Naganna Chetty and Sreejith Alathur. 2018. Hate speech review in the context of online social networks. *Aggression and Violent Behavior,* 40, 108–118.

Hamish Cunningham, Diana Maynard, and Valentin Tablan. 2000. JAPE: A Java annotation patterns engine. Technical report, University of Sheffield, Department of Computer Science.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: an Architecture for Development of Robust HLT applications. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 168–175, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM 2017)*, Montreal, Canada. arXiv:1703.04009.

René Del Fabbro. 1995. Germany: A victor of the street. In B. Baumgartl, and A. Favell (eds.). *New xenophobia in Europe*. London, The Hague, Boston, pp. 132-147.

Gretel Liz De la Peña Sarracén and Paolo Rosso. 2023. Systematic keyword and bias analyses in hate speech detection. *Information Processing and Management 60* (5).

Elias Dinas, Vasiliki Georgiadou, Ioannis Konstantinidis, and Lamprini Rori. 2016. From Dusk to Dawn. Local party organization and party success of right-wing extremism. *Party Politics* 22(1), 80-92.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web (WWW 2015)*, Florence, Italy.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1).

Costas Gabrielatos and Paul Baker. 2008. Fleeing, Sneaking, Flooding: A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the UK Press, 1996-2005. *Journal of English Linguistics*, *36*(1), 5-38.

Ioannis Galariotis, Vasiliki Georgiadou, Anastasia Kafe, and Zinovia Lialiouti. 2017. Xenophobic manifestations, Otherness, and violence in Greece: Evidence from an event analysis of

Media collections. EUI Working Paper MWP 2017/08.

Iginio Gagliardone, and Matti Pohjonen, Kate Orton-Johnson (Ed.). 2022. How to Analyze Online Hate Speech and Toxic Communication [How-to Guide]. Sage Research Methods: Doing Research Online.

Maria Gavriilidou, Stelios Piperidis, Dimitrios Galanis, Kanella Pouli, Penny Labropoulou, Juli Bakagianni, Iro Tsiouli, Miltos Deligiannis, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, and Katerina Gkirtzou. 2024. The CLARIN:EL infrastructure: Platform, Portal, K-Centre. Selected papers from the CLARIN Annual Conference 2023.

Vasiliki Georgiadou, Lamprini Rori, and Costas Roumanias. 2018. Mapping the European far right in the 21st century: A meso-level analysis. *Electoral Studies* 54, 103-115.

Vasiliki Georgiadou. 2020. The Far Right. In K. Featherstone, and D. A. Sotiropoulos (eds.) The Oxford Handbook of Modern Greek Politics. Oxford: Oxford University Press, 2020, pp. 242-255.

Amira Ghenai, Zeinab Noorian, Hadiseh Moradisani, Parya Abadeh, Caroline Erentzen, and Fattane Zarrinkalam. 2025. Exploring hate speech dynamics: The emotional, linguistic, and thematic impact on social media users. *Information Processing and Management*, *62*(3), 104079, ISSN 0306-4573. Roderick S. Graham and Shawn K. Smith. 2024. Cybercrime and Digital Deviance. 2nd Edition, Routledge.

Ming Shan Hee, Shivam Sharma, Rui Cao, Palash Nandi, Preslav Nakov, Tanmoy Chakraborty, and Roy Ka-Wei Lee. 2024. Recent Advances in Online Hate Speech Moderation: Multimodality and the Role of Large Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4407–4419, Miami, Florida, USA. Association for Computational Linguistics.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using nlp to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666.

Brendan Kennedy, Mohammad Atari, Aida M. Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs Jr, Shreya Havaldar, Gwenyth PortilloWightman, Elaine Gonzalez, et al.. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech. PsyArXiv.

Brendan Kennedy, Mohammad Atari, Aida M. Davani,. et al.. 2022. Introducing the Gab Hate Corpus: defining and applying hate-based rhetoric to social media posts at scale. *Lang Resources and Evaluation* 56, 79–108.

Heejung S. Kim, David K. Sherman, and John A. Updegraff. 2016. Fear of Ebola: The Influence of Collectivism on Xenophobic Threat Responses. *Psychological Science*, *27*(7), 935-944.

Gallya Lahav, and Anthony M. Messina, "Securitizing and Politicizing Immigration: Political Party Competition in Spain, UK, and the US," in *Immigration, Security, and the Liberal State: The Politics of Migration Regulation in Europe and the United States*, Cambridge: Cambridge University Press, 2024, pp. 262–319.

Rigas Kotsakis, Lazaros Vrysis, Nikolaos Vryzas, Theodora Saridou, Maria Matsiola, Andreas Veglis, and Charalampos Dimoulas. 2023. A web framework for information aggregation and management of multilingual hate speech. *Heliyon, 9*(5): e16084.

Yonatan Lupu, Richard Sear, Nicolas Velásquez, Rhys Leahy, Nicholas Johnson Restrepo, Beth Goldberg, and Neil F Johnson. 2023. Offline events and online hate. PLoS ONE 18(1): e0278511.

Barbara McGillivray, Thierry Poibeau, and Pablo Ruiz. 2020. Digital Humanities and Natural Language Processing: "Je t'aime... Moi non plus". *Digital Humanities Quarterly, 14* (2).

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 85–94. Association for Computing Machinery, New York, NY, USA.

Cas Mudde. 2007. Populist radical-right parties in Europe. Cambridge: Cambridge University Press.

Karsten Müller and Carlo Schwarz. 2020. Fanning the Flames of Hate: Social Media and Hate Crime. *Journal of the European Economic Association*, *19*(4), 2131–2167.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper Attention to Abusive User Content Moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark. Association for Computational Linguistics.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive Language Identification in Greek. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.

Konstantinos Perifanos and Dionysis Goutsos. 2021. Multimodal Hate Speech Detection in Greek Social Media. *Multimodal Technologies and Interaction*, *5*(7), 34.

Maria Pontiki. 2019. Fine-grained Sentiment Analysis. PhD Thesis. University of Crete.

Maria Pontiki, Konstantina Papanikolaou, and Haris Papageorgiou. 2018. Exploring the predominant targets of xenophobia-motivated behavior: A longitudinal study for Greece. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Natural Language Meets Journalism Workshop III*, pages 11 –15, Miyazaki, Japan. European Language Resources Association.

Maria Pontiki, Maria Gavriilidou, Dimitris Gkoumas, and Stelios Piperidis. 2020. Verbal Aggression as an Indicator of Xenophobic Attitudes in Greek Twitter during and after the Financial Crisis. In *Proceedings of the Workshop about Language Resources for the SSH Cloud*, pages 19–26, Marseille, France. European Language Resources Association.

Maria Pontiki, Nikos Saridakis, Dimitris Gkoumas, and Maria Gavriilidou. 2022. #le_petit_koulis and #tsipras_the_traitor: Verbal Aggression as an Aspect of Political Violence on Greek Twitter. *Journal of Modern Greek Studies*, 40(1): 63-93.

Prokopis Prokopidis, Byron Georgantopoulos, and Harris Papageorgiou. 2011. A suite of NLP tools for Greek. In *Proceedings of the 10th International Conference of Greek Linguistics*, pages 373–383, Komotini, Greece.

Lamprini Rori. 2021. From 'black sheep of the eurozone' to 'European shield': Ten years of crisis politics in Greece. In C. Spanou (ed.) *Crisis, reform and the way forward in Greece. A turbulent decade. London & New York:* Routledge, 2021, pp. 64-82.

Lamprini Rori, and Vasiliki Georgiadou. 2023. Far Left Organised Violence in Greece. The Second Generation. In *The Palgrave Handbook of Left-Wing Extremism, Volume 1* (pp. 223-246). Springer International Publishing, Cham.

Lamprini Rori, Vasiliki Georgiadou, and Costas Roumanias. 2022. Political violence in crisis-ridden Greece: Evidence from the far right and the far left. *Journal of Modern Greek Studies* 40(1): 1-37.

Hannah Rose, and Paula-Charlotte Matlach. 2024. Narratives of Hate. Post-7 October Antisemitism and Anti-Muslim Hate on Social Media. Institute for Strategic Dialogue (ISD), Amman, Berlin, London, Paris, Washington DS.

Bjorn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of the 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pages 6-9, Bochum, Germany, 22 September, 2016.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Cass R. Sunstein. 2018. *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Lazaros Vrysis, Nikolaos Vryzas, Rigas Kotsakis, Theodora Saridou, Maria Matsiola, Andreas Veglis, Carlos Arcila-Calderón, and Charalampos Dimoulas. 2021. A Web Interface for Analyzing Hate Speech. *Future Internet* 13 (3): 80.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2023. Large language models are latent variable models: explaining and finding good demonstrations for in-context learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 687, 15614–15638.

William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Ruth Wodak. 2015. The Politics of Fear: What Right-Wing Discourses Mean. Sage Publications Ltd.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. HARE: Explainable hate speech detection with step-by-step reasoning. *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505, Association for Computational Linguistics.

Michael Yoder, Lynnette Ng, David West Brown, and Kathleen Carley. 2022. How Hate Speech Varies by Target Identity: A Computational Analysis. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 27–39, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Appendix A. Fluctuation of the individual VA types per TG.

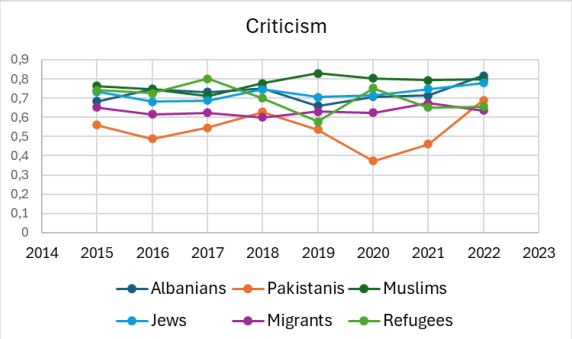

Figure 6: Criticism rates per year and TG.

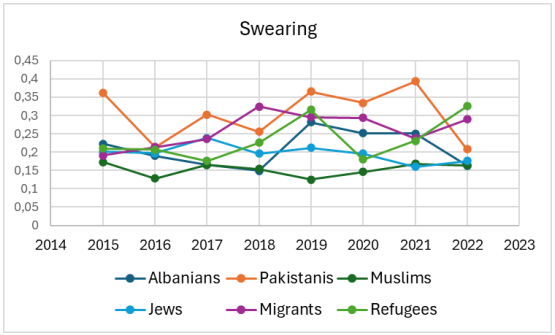

Figure 7: Swearing rates per year and TG.

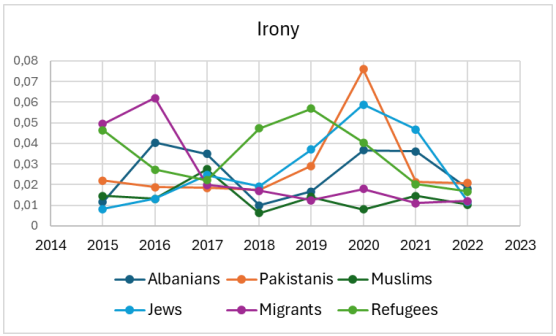

Figure 8: Irony rates per year and TG.

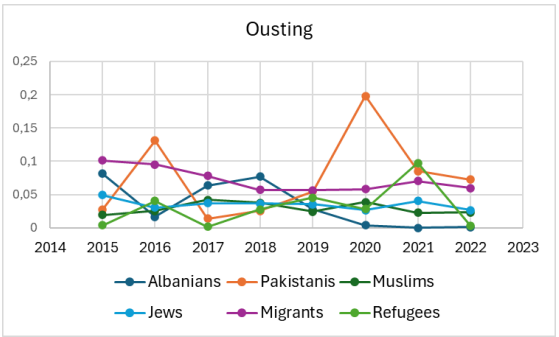

Figure 9: Ousting rates per year and TG.

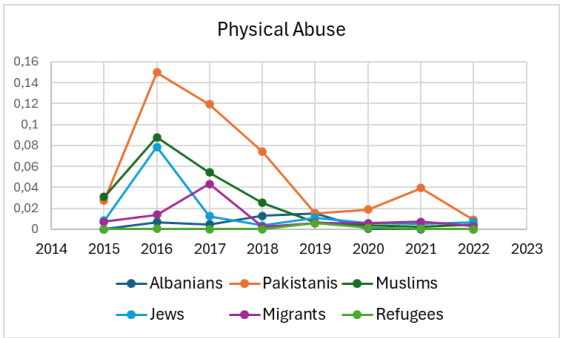

Figure 10: Physical Abuse rates per year and TG.