Hate Speech and Hate Crime: a Cross-Disciplinary Analysis of Xenophobia in Greece

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

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

Abstract

This paper investigates the correlation of hate speech and hate crime, in a inter-disciplinary approach, using a computational hate speech and hate crime detection method in a socio-political science framework, coupling Natural Language Processing with Political Sciences. The study focuses on Greece in the turbulent period from 2015 to 2022 (a period marked by economic, refugee, foreign policy, and pandemic crises); it analyzes tweets to discern linguistic patterns used to verbally attack predefined target groups consisting of ethnic and religious minorities in the country. Furthermore, it investigates hate crimes reported in the press, against the same target groups and during the same period and proceeds to examine correlations between xenophobic attitudes expressed verbally through social media, and those manifested as physical attacks in real life.

Keywords: computational political science, NLP, hate speech, hate crimes, xenophobia

1. Introduction

In this study, we examine xenophobia in Greece, as expressed online via social media, and offline as physical attacks recorded in the press. We conceptualize xenophobia as a two-dimensional violent practice and analyze attacks occurring between 2015 and 2022 - a period spanning the economic, refugee, and pandemic crises. The interdisciplinary approach adopted deploys a rule-based NLP method within a political science framework, which collects, processes and analyzes data from social media related to hate speech, and from newspapers reporting hate crimes against specific target groups.

Our analysis focuses on six target groups: ALBANIANS, PAKISTANIS, MUSLIMS, JEWS, MIGRANTS and REFUGEES. We investigate which of these groups, in specific socio-political contexts, were the primary targets of hate speech on Greek Twitter/X during the examined period, and whether distinct linguistic patterns (used per target group for verbal aggression) or stereotypes can be identified. Given that social media discourse shapes the environment in which xenophobic attacks occur, and that incendiary rhetoric often stigmatizes and dehumanizes minorities, we further explore links between online hate speech and offline xenophobic attacks.

The main goals of the research are three: first, our analysis reveals key linguistic patterns reinforcing populist narratives, polarization, and hostility within Greek online discourse, namely verbal aggression expressed over social media. The second goal focuses on the analysis of hate crime (physical aggression) as reported in the press; finally, the third goal corresponds to the investigation of the correlation between hate speech and hate crime.

The results of the first goal of our research are published in (Pontiki et al, 2025), and are summarized in the current paper, for the comparison between Verbal and Physical Aggression to be meaningful. The current paper focuses on the second and third goals, whose results it presents.

By comparing online hate speech (Verbal Aggression) with offline hate crimes (Physical Aggression) reported in the press, we contribute to the discussion for understanding the connection between digital aggression and real-world political violence.

The paper is structured as follows: Section 2 presents the background of the research and the relevant literature; Section 3 summarizes the results of the analysis of Verbal Aggression; Section 4 presents the results of the analysis of Physical Aggression; Section 5 focuses on the comparison of the two produced datasets and investigates their correlation; Section 6 lists the developed resources and Section 7 completes the paper with the conclusions.

2. Background

Hate speech is a general term covering a broad spectrum of extremely negative discourse stretching from hatred and incitement to abusive expression and vilification, as well as to extreme forms of prejudice and bias, and can target on different bases, e.g. religion, disability, social status, politics, race, sex and gender issues, plus others. The spread of hate speech in online platforms has significantly expanded its reach and

Proceedings of The 3rd Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2026) @ LREC 2026, pages 139–148
16 May 2026, Palma, Mallorca (Spain).

impact, deepening polarization and undermining democratic discourse (Sunstein, 2018). It reinforces stereotypes, dehumanizes individuals or groups, and often leads to discrimination, marginalization and enmity against specific groups. The relative anonymity offered to users by social media platforms reduces accountability and facilitates the open expression of prejudiced views (Mondal et al., 2017), thereby intensifying social and ideological divisions. Furthermore, the normalization of hate speech by influential figures, such as political leaders, has further legitimized hateful rhetoric, embedding it in mainstream discourse and, in polarized contexts, enabling online platforms to act as catalysts for offline violence.

Some studies adopt a binary classification schema aiming to distinguish hateful from non-hateful content (e.g. Djuric et al., 2015), other studies attempt to differentiate hate speech and offensive language, while another line of research focuses on specific types/categories of hate speech e.g. racist and sexist hate speech (Waseem and Hovy 2016). Natural Language Processing (NLP) research has made substantial progress in detecting hateful content (e.g., Jurgens et al., 2019; Sap et al., 2020; Caselli et al., 2021; ElSherief et al., 2021; Yoder et al., 2022), laying foundations for interventions such as moderation, debiasing, and counter-speech (Hee et al., 2024). Advances in large language models (LLMs) have further enhanced interpretability and performance, allowing more nuanced analyses. However, applying hate speech detection in real-world contexts remains challenging. The task is highly culturally sensitive (Schmidt and Wiegand, 2017) because hate speech is rooted in the sociocultural contexts from which it emerges (Warner and Hirschberg, 2012; Kennedy et al., 2022). Ethical concerns also arise when labeling communicative practices as hateful, particularly for affected communities (Gagliardone et al., 2022). Hence, models must be rigorously validated to ensure they reflect complex social realities, since erroneous detection of hate speech risks censoring speech and further marginalizing vulnerable groups (Yang et al., 2023). Computational approaches for Greek focus on classifiers for offensive tweet detection (Pitenis et al., 2020) and moderation of abusive content in user generated comments (Pavlopoulos et al., 2017). Perifanos and Goutsos (2021) proposed a multimodal approach to detect abusive contexts in tweets targeting refugees and migrants. Pontiki et al., (2018; 2020; 2025) employed a rule-based framework to identify and categorize specific types of Verbal Aggression against predefined groups. Arcila-Calderón et al., (2022) developed learning models for the detection of online anti-immigration hate speech in Spanish, Greek and Italian.

A further challenge stems from the differing aims of NLP and the social and political sciences (McGillivray et al., 2020). While NLP focuses on building and evaluating computational systems — typically using standardized, English-dominant datasets (Arango et al., 2022) — social and political sciences seek to interpret the political meaning of hate speech and to link it with the current social and political context, emphasizing qualitative as well as quantitative research. This divergence highlights the need for interdisciplinary approaches that adapt computational tools to domain-specific questions and real-world complexities (McGillivray et al., 2020).

3. Verbal Attacks (VA)

3.1 Collection and Processing of VA Data

This section summarizes the method applied and the results of the analysis of Verbal Attacks (first goal of the research), which have already been published (Pontiki et al, 2025).

The research focuses on six predefined target groups: ALBANIANS and PAKISTANIS (the largest migrant nationalities in Greece), MUSLIMS and JEWS (key religious and ethnic minorities in the country), and finally the broader categories of MIGRANTS and REFUGEES.

For each Target Group (TG), relevant tweets were retrieved using related keywords, leading to a total of 4,386,501 tweets covering the period 2015-2022. REFUGEES and MIGRANTS receive the highest number of tweets (Figure 1), with varying intensity over the years.

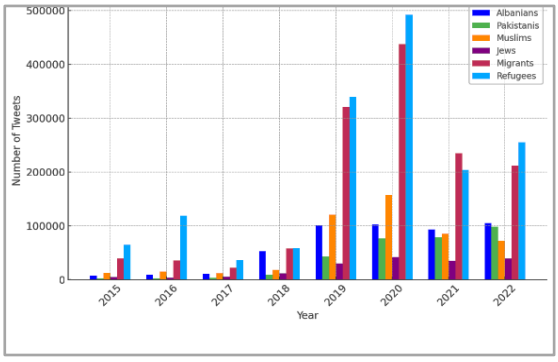

Figure 1: Total amount of tweets per TG and year.

The collected tweets were processed with the GR_VA_Analyzer[1] web service, freely accessible through the CLARIN:EL infrastructure[2] (Gavriilidou et al., 2024).

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

[2] https://inventory.clarin.gr/

The VA analysis tool is a rule-based method comprising lexical resources and linguistic patterns for the detection of explicit verbal attacks against specific targets (Pontiki et al., 2018; 2020), and captures five types of verbal attacks (Pontiki, 2019): **Criticism** (disapproval or negative evaluations of the target), **Swearing** (taboo or profane language used to degrade or insult the target), **Irony** (sarcastic, satirical, or humoristic language), **Ousting** (calls for ouster), and **Physical Abuse** (intentions or calls for physical violence or physical extinction).

The rule-based NLP method was preferred over AI-based techniques, with the aim to better control all steps of the processing, to monitor the performance of the tools, and to adapt them to the requirements of the political scientists.

The method is precision-oriented and focuses on explicitly stated VA; it relies on a set of lexical resources built to capture possible linguistic instantiations of VA towards the TGs of interest. 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). The analyzer initially 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. Each identified attack is represented as a structured tuple containing the target group, the linguistic evidence due to which it was identified as VA, the recognized VA type, and the relevant linguistic evidence. The performance of the VA analyzer was evaluated using a random selection of 500 Tweets per TG (5000 Tweets in total) in terms of Precision (84%), Recall (60%) and FMeasure (68%) (Pontiki et al., 2020). The verbal attacks detected by the tool underwent manual validation of randomly selected samples, as well as deduplication and removal of false positives.

3.2 Analysis of VA Data

As illustrated in Figure 2, MIGRANTS, MUSLIMS and PAKISTANIS are the main targets of verbal attacks over X, for the whole period under examination. It should be noted that the term *Pakistani* is used both as referring to Pakistani nationals, and (by racist groups) as a generic term, covering various Asian ethnic minorities (Afghans, Kurds, etc.).

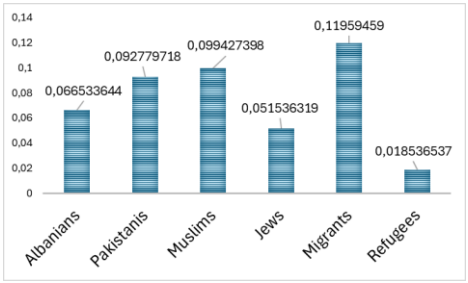

Figure 2: Overall VA rate per TG for the period 2015-2022.

In line with previous research (Pontiki et al., 2018; 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. This means that the selection of one term or the other is based on the opinion holder's general political position: the use of the term *migrant* to refer to recognized *refugees* is not without bias. However, it is obvious that the tool cannot distinguish the correct (or not) uses of the specific terms.

As regards the types of verbal attacks (Figure 3), criticism constitutes the main type of VA detected in our datasets, mostly directed against MUSLIMS, JEWS, ALBANIANS and REFUGEES. PAKISTANIS and MIGRANTS receive the most obscene messages, while the main recipients of ironic tweets are PAKISTANIS, JEWS and REFUGEES. Calls for ousting target mostly MIGRANTS and PAKISTANIS, who are also the main targets of calls for physical violence.

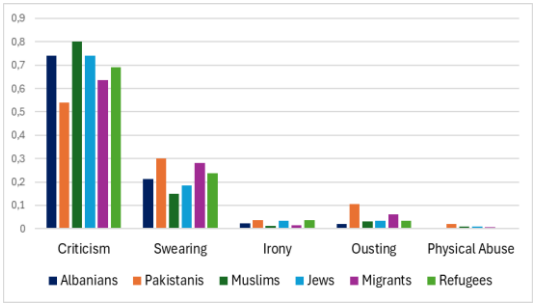

Figure 3: Overall VA type rates per TG for the period 2015-2022.

Criticism, besides being the most frequently used linguistic weapon, also presents relatively stable distribution over the years. However, the decrease of criticism rates does not indicate a general decrease of VA, but rather a shift in the VA type, with aggression moving from criticism towards swearing and irony.

The analysis of VA shows that in social media (X) in Greece, during the period under study:

- specific groups (MIGRANTS, MUSLIMS and PAKISTANIS) are mostly targeted
- different VA types are preferred according to target group
- aggression slides from criticism towards swearing and irony.

4. Physical Attacks (PA)

4.1 PA Data Collection

Greece seems to lack a reliable monitoring system to record and monitor xenophobic physical attacks, despite recent research in the broader field of political violence demonstrating significant fluctuations in physical attacks motivated by cultural variations and expressions of otherness (Georgiadou et al., 2018; Georgiadou, 2020; Pontiki et al., 2022).

For the compilation of a dataset to study xenophobic physical attacks against the same target groups during the same period, we collected material from 10 Greek newspapers, comprising a total of 3,768,597 articles published between 2015 and 2022. To ensure representativeness, we selected newspapers spanning the ideological spectrum (left, center-left, center, center-right, and right). Focused articles (i.e. articles reporting violent incidents against the specific target groups) were retrieved using keyword searches.

4.2 PA Data Processing

Xenophobia-motivated behavior includes several offensive practices against foreigners. For the current research, we rely on previous classification of anti-foreign attacks (Galariotis et al., 2017; Pontiki et al., 2018) following a coding schema (Papanikolaou et al., 2020) that covers a wide spectrum of event types related to xenophobia. In particular, the event taxonomy includes a major event category **Physical Attacks**, encompassing the following event types:

- **Assault** (e.g., *attacked, attack*)
- **Violent Assault** (e.g., *tortured, was beaten, abduction*)
- **Assault Life** (e.g., *murder, killed, injured, stabbed*)
- **Sexual Assault** (e.g., *attempted to rape*)
- **Verbal Attack** (e.g., *verbally abused, verbal assault*)
- **Property Attack** (e.g., *attacked the house/office*)
- **Religious Attack** (e.g., *desecration of religious monument*).

In this schema, the coding unit of the analysis is the Event. An Event comprises a tuple containing information about its structural components:

1. EVENT. The word or phrase representing an event type located within the text.
2. ACTOR. The entity that performs each event instance.
3. TARGET. The entity to whom the action is addressed.
4. LOCATION. The location where the event took place.
5. TIME. The time at which the event happened.
6. CONFIDENCE. The element which captures whether in the article there is any indication that an Actor of an assault may not be the actual perpetrator.

For example, the sentence "*A 24-year-old American is accused of involvement in the arson of the Synagogue at Chania on Thursday*" is coded as follows:

<Actor: *A 24-year-old American*, Confidence: *is accused of involvement*, Event: *in the arson*, Target: *of the Synagogue*, Location: *at Chania,* Time: *on Thursday* >.

This fine-grained schema enables structured and detailed coding of events, supporting both quantitative and qualitative analyses, even though not all information is always available. It is used by a dedicated tool, the *ILSP event extractor for physical attacks in Greek*[3] (Pontiki et al., 2018, Papanikolaou and Papageorgiou, 2020), available through the European Language Grid.[4] The event extractor is a rule-based engine designed to detect different types of physical attacks recorded in Greek news data, along with their structural components, following the above-described event coding schema.

The tool was used to process the newspaper material for the extraction of xenophobic violent attacks. The tool's output is a set of tuples, each depicting an event with its structural elements. These were recorded in the Event Database, and subsequently underwent manual validation by two human annotators, who inspected all automatically detected physical attacks and classified them as (a) Physical attacks constituting xenophobic hate crimes against the predefined target groups in Greece during the period under examination, or (b) Physical attacks not meeting the above criteria, including false positives, incidents outside Greece, events outside the period under study, or other types of crimes. Duplicate entries (e.g., the same event detected multiple times within or across articles) were removed.

A total of 227 hate crimes were identified (Table 1) as reported by the selected sources for the period under examination. It should be noted that this number refers to hate crimes reported in the newspapers and identified by our research, and

[3] https://doi.org/10.57771/mdj5-c740

[4] https://live.european-language-grid.eu/catalogue/

not to the actual number of hate crimes. The number of reported hate crimes might not fully represent reality, due to under-reporting.

TG	#hate crimes
Albanians	7
Pakistanis	19
Muslims	7
Jews	16
Migrants	121
Refugees	57

Table 1: Total number of hate crimes per TG.

4.3 Analysis of PA Results

The distribution of reported hate crimes by year and ideological orientation of newspaper (Figure 4) shows a steady increase until 2018, when most sources reached a peak in reporting.

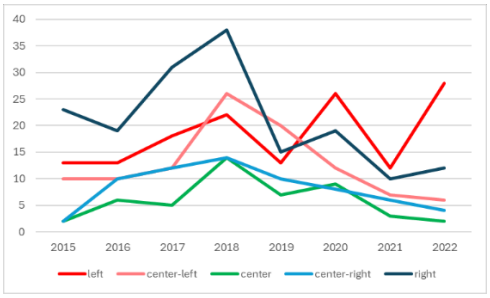

Figure 4: Hate crimes by year and ideological orientation of source.

Another significant rise appears in 2020 and again in 2022, though overall reporting declines after 2019 across several ideological groups. A key exception is left-oriented newspapers, which show a marked resurgence in 2020 and a pronounced peak in 2022, indicating a shift relative to right-oriented outlets during this period, possibly influenced by the 2019 elections and subsequent change in government (currently right-wing). Center-left and center-right outlets maintain moderate levels of reporting, with peaks in 2018 but a noticeable decline thereafter, while center-oriented outlets report at lower levels overall, with smaller peaks in 2018 and 2020. Taken together, these patterns highlight substantial variation in hate crime reporting across ideological orientations, with 2018 standing out as common high point followed by divergent trajectories.

The distribution of the 227 unique hate crimes reported across all sources per year reveals a significant peak for most target groups in 2018 (Figure 5). Overall trends show a steady increase in hate crimes in Greece from 2015, culminating in the 2018 surge, followed by a decline after 2019. By 2022, reported levels return to those observed in 2016. The main actors identified by the tool include the Greek neo-Nazi party Golden Dawn, other extreme-right groups, police, and individuals involved in the employment of irregular immigrants.

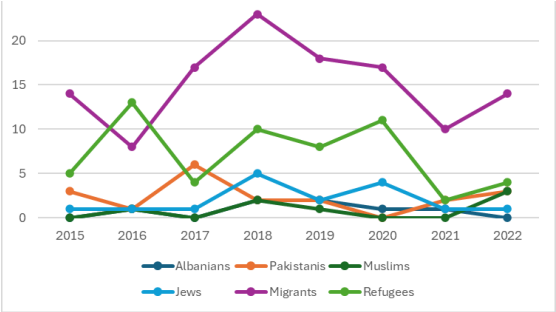

Figure 5: Hate crimes per year and TG.

MIGRANTS (121 cases, 53%) and REFUGEES (57 cases, 25%) are the most frequently targeted groups during the period under examination, together accounting for nearly four out of five reported hate crimes. PAKISTANIS (19 cases, 8%) and JEWS (16 cases, 7%) also appear as significant targets, though at considerably lower levels compared to MIGRANTS and REFUGEES.

Violent assault is the most frequent type of attack, accounting for 43% of all incidents, followed by assault against life (19%), general assault (18%), verbal attacks (9%), and religious attacks (8%).

Overall, based on our datasets (Figure 6), REFUGEES and MIGRANTS receive by far the most media attention, fact which reflects their continuous presence in Greek public discourse during the ongoing migration crisis. MUSLIMS and ALBANIANS follow in third and fourth place, respectively. By contrast, PAKISTANIS and JEWS appear far less frequently, suggesting lower visibility in traditional media. Their coverage is more issue-specific (e.g., anti-Semitic incidents) rather than sustained as a broad topic of discussion—an interpretation supported by the physical aggression rates attested.

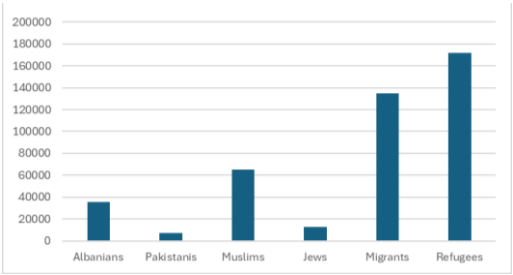

Figure 6: Total focused articles for each Target Group for the period 2015-2022.

As also observed in verbal attacks, different target groups experience distinct forms of hate crimes. MIGRANTS, REFUGEES, and PAKISTANIS are the main victims of verbal abuse, violent physical assaults, and life-threatening attacks. JEWS are predominantly targeted through religiously motivated vandalism, including the desecration of

Jewish cemeteries, Holocaust monuments, and synagogues. MUSLIMS also experience religiously motivated attacks, such as the vandalism of cemeteries, along with verbal assaults on Muslim MPs by members of Golden Dawn.

5. Correlation between Twitter and Newspaper Data

Table 2 presents a comparison (in terms of size) of the two data collections:

	Tweets	Newspaper Articles
Albanians	481,621	35,661
Pakistanis	313,021	7,471
Muslims	493,013	65,330
Jews	170,928	13,004
Migrants	1,359,610	134,517
Refugees	1,568,308	171,468
TOTAL	**4,386,501**	**427,451**

Table 2: Comparison of size per data collection.

This comparison highlights the difference in the volume of discourse about these groups across social media (Tweets) and traditional media (newspaper articles). Tweets reflect spontaneous, real-time reactions, while news articles involve editorial processes, leading to different response patterns. However, the comparison also suggests that, during the period under examination, public discourse in Greece—both in social and traditional media—followed similar trends in focusing on these target groups (Figure 7).

MIGRANTS and REFUGEES dominate both datasets (with a ratio of tweets to newspaper articles approximately 10:1). MUSLIMS have a higher ratio of newspaper articles compared to other groups, indicating that they are more central in formal media discussions as compared to other targets, especially due to global events that often bring religious groups into mainstream news coverage.

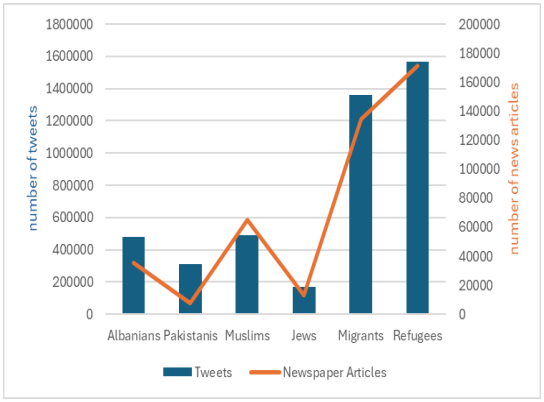

Figure 7: Volume of discourse per TG and source.

For groups like ALBANIANS, PAKISTANIS, and (to a lesser extent) JEWS, the number of tweets visibly exceeds the number of newspaper articles, potentially indicating that these groups are discussed more in informal, unregulated spaces like social media than in traditional media. In particular, the vast number of tweets compared to newspaper articles for ALBANIANS and PAKISTANIS possibly indicates that these groups are disproportionately targeted in informal online spaces where hate speech and verbal aggression can spread more freely, as well as a lack of focus on these groups in formal public discourse.

The correlation between Verbal Aggression rate and Physical Aggression rate per target group (Figure 8) reveals that:

- MIGRANTS experience the highest number of hate crimes (121) and also have the highest hate speech rate (0.1196).
- REFUGEES have the second-highest number of hate crimes (57), but their hate speech rate (0.0185) is the lowest.
- PAKISTANIS also exhibit a relatively high number of hate crimes (19) despite having a moderate hate speech rate (0.0928).
- JEWS have a relatively low hate speech rate (0.0515) but a notable number of hate crimes (16), suggesting that hate-motivated violence against Jewish communities may stem from factors beyond contemporary online discourse.
- ALBANIANS and MUSLIMS both have similar hate crime numbers (7 each) despite moderate hate speech rates (~0.0665 and ~0.0994, respectively), indicating that the online discourse does not necessarily translate into more **reported** hate crimes.

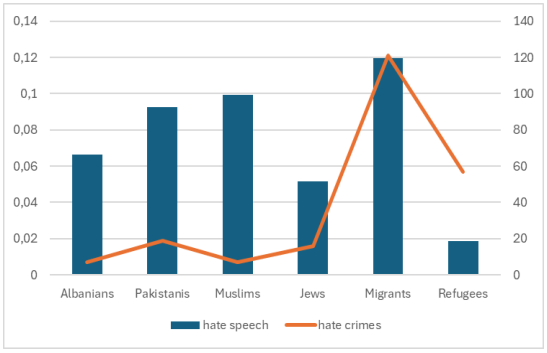

Figure 8: Correlation between VA rate and number of hate crimes per TG.

As depicted in Figure 9, hate speech rates show a gradual increase from 2015 (0.0426) to 2022 (0.0750). The number of hate crimes fluctuates over the years, peaking in 2018 (44 hate crimes) and showing a decline afterwards. Specifically:

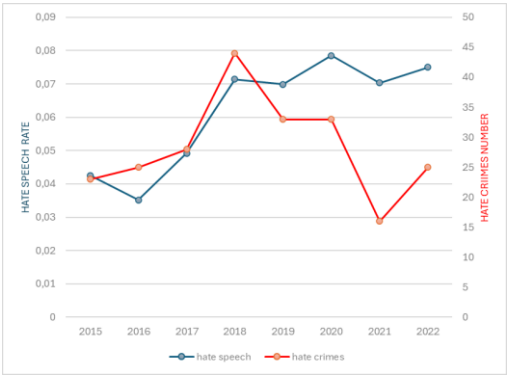

Figure 9: Evolution of hate speech and hate crime during 2015 – 2022.

- **2015-2017**: Both hate speech rates and hate crimes were relatively low and stable.
- **2018**: A notable spike in both hate speech (0.0714) and hate crimes (44) suggests a potential correlation between online discourse and real-world violence.
- **2019-2020**: Hate speech rates remained high (~0.07-0.08), while hate crimes declined slightly but stayed relatively elevated (33 cases in both years).
- **2021**: A sharp drop in hate crimes (16) despite high hate speech rates (0.0703), suggesting that external factors may have influenced real-world outcomes.
- **2022**: Hate speech rates slightly increased (0.0750), and hate crimes also (25).

The Pearson correlation coefficient (r=0.32) suggests a weak positive correlation between hate speech rates and the actual number of hate crimes. Possible explanations for the weak correlation are:

- **Reporting Variability**: Due to underreporting, the number of reported hate crimes does not fully reflect reality. Underreporting has double sense: victims do not report the incident to the police, but even when reported, the incident is not mentioned in the newspapers.
- **Lag Effect**: Hate speech may not result in immediate crimes but could contribute to longer-term normalization of hostility, leading later to hate crime.
- **Other Influencing Factors:** There is a general association between rising hate speech and hate crimes, particularly in the 2015-2018 period. However, deviations in 2019-2022 suggest that hate speech alone may not be the sole factor influencing hate crimes. Offline social, political, economic factors, historical context might impact hate crimes independently of hate speech trends.

6. Resulting Resources – Monitoring Tool

The two datasets constructed in the framework of this research, namely Verbal Aggression database[5] and Hate crimes database[6], as well as the VA analyzer[7] are available through the CLARIN:EL infrastructure.

The results have also served as the basis for the development of a monitoring tool[8] that maps xenophobic violence and tracks its variation over time through a range of visualizations parametrizable by the user (e.g., per year, per group, per geographical area, per aggression type, etc.), making the results explorable, comprehensible, and interpretable for analysts, stakeholders, and policymakers.

7. Conclusions

Our research, situated in the interdisciplinary field combining NLP with social and political sciences, has attempted to investigate the expression of xenophobic attitudes, both verbally and physically. Using as a case study Greece in the turbulent years 2015-2022, we analyzed hostility towards specific target groups, as expressed in the social media and in real life; we furthermore examined the correlation between word and deed at the social level and to comprehend their interaction. Among others, our study has shown the frequent association of Verbal Aggression against migrants with criminality: dehumanizing

[5] http://hdl.handle.net/11500/CLARIN-EL-0000-0000-7B9E-5
[6] http://hdl.handle.net/11500/CLARIN-EL-0000-0000-7BB7-8
[7] http://hdl.handle.net/11500/CLARIN-EL-0000-0000-7B9E-5
[8] https://www.demolish.gr/va-monitoring-tool

specific ethnic groups renders those groups vulnerable targets for physical harm.

Overall, online hate speech does not translate directly into offline hate crimes; rather, it can function as an enabling environment that normalizes hostility and lowers threshold for intolerance. Historical legacies may contribute more heavily than social media activity for some targets, and time-lag effect is plausible: hate speech may not immediately trigger hate crimes but can contribute to their long-term normalization.

Despite its limitations (see relevant section below), the method offers valuable insights into the ways hate speech manifests online in Greece in response to real-world grievances and crises and how its correlated with offline hate crime.

Ultimately, our analysis underscores the need for interdisciplinary approaches that align computational methodologies with sociopolitical contexts, while integrating human oversight. Such approaches are essential for capturing linguistic nuances, speaker intent, and culturally or domain-specific variations, as well as for addressing the ethical complexities inherent in automated detection of hate speech and hate crime.

Future work includes experimentation with AI-based methods and comparative study of the results with those of the NLP pipeline.

8. Limitations

The VA analysis tool, which is a lexicon and rule-based method, may fail to detect implicit or ironic verbal attacks, or tweets using alternative terms or emerging slurs not present in the lexicon. We also recognize the possibility that some of the detected content originates from bots or fake accounts.

The analysis of hate crimes was exhaustive, given the small number of events identified. However, the number of events (as already mentioned) is not considered representative of the real number of hate crimes, due to underreporting. Thus, the correlation investigated is not between hate speech and the real number of hate crimes, but rather with the number of hate crimes reported in the press.

The rule-based NLP method used for the processing of the datasets has its limitations, as described. In future work, experimentation with AI methods on the same research topic would be an interesting comparative experiment.

9. Acknowledgments

The work presented in this paper was 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).

10. Ethics Statement

In accordance with both the General Data Protection Regulation (GDPR)[9] and the Developer Policy of X[10], 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.

11. Bibliographical References

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

Arcila-Calderón, C. Amores J.J., Sánchez-Holgado P., Vrysis, L., Vryzas, N., and Oller Alonso M. 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.

Caselli, T., Basile, V., Mitrović, J., and Granitzer, M. 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.

Cunningham, H., Maynard, D., and Tablan, V. 2000. JAPE: A Java annotation patterns engine. Technical report, University of Sheffield, Department of Computer Science.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. 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.

Djuric N,, Zhou J., Morris R., Grbovic M., Radosavljevic, V., and Bhamidipati, N. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th*

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

[10] https://docs.x.com/developer-terms/policy

International Conference on World Wide Web (WWW 2015), Florence, Italy.

ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., and Yang D. 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.

Galariotis, I., Georgiadou, V., Kafe, A., and Lialiouti Z. 2017. Xenophobic manifestations, Otherness, and violence in Greece: Evidence from an event analysis of Media collections. EUI Working Paper MWP 2017/08.

Gagliardone, I., and Pohjonen, M. 2022. How to Analyze Online Hate Speech and Toxic Communication [How-to Guide]. Sage Research Methods: Doing Research Online.

Gavriilidou, M., Piperidis, S., Galanis, D., Pouli, K., Labropoulou, P., Bakagianni, J., Tsiouli, I., Deligiannis, M., Kolovou, A., Gkoumas, D., Voukoutis, L., and Gkirtzou, K. 2024. The CLARIN:EL infrastructure: Platform, Portal, K-Centre. Selected papers from the CLARIN Annual Conference 2023.

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

Georgiadou, V. 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.

Hee M.S., Sharma S., Cao R., Nandi P., Nakov P., Chakraborty, T. and Ka-Wei Lee R. 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.

Jurgens, D., Hemphill, L., and Chandrasekharan E. 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.

Kennedy, B., Atari, M., Davani, M. 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.

McGillivray, B., Poibeau, T., and Ruiz P. 2020. Digital Humanities and Natural Language Processing: “Je t’aime... Moi non plus”. *Digital Humanities Quarterly, 14* (2).

Mondal, M., Araújo Silva L., and Benevenuto F. 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.

Papanikolaou, K. and Papageorgiou, H. 2020. Protest Event Analysis: A Longitudinal Analysis for Greece. In *Proceedings of Language Resources and Evaluation Conference* (LREC 2020), pp. 57–62.

Pavlopoulos, J., Malakasiotis, P., and Androutsopoulos I. 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.

Pitenis, Z., Zampieri, M., and Ranasinghe T. 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.

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

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

Pontiki, M., Gavriilidou, M., Gkoumas, D., 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.

Pontiki, M., Georgiadou, V., Rori, L., and Gavriilidou, M. 2025. Hate Speech in Times of Crises: a Cross-Disciplinary Analysis of Online Xenophobia in Greece. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 241–253, Vienna, Austria. Association for Computational Linguistics. ISBN 979-8-89176-105-6.

Pontiki, M., Papanikolaou, K. and Papageorgiou H. 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.

Pontiki, M., Saridakis, N., Gkoumas, D., and Gavriilidou, M. 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.

Sap, M., Gabriel S., Qin L., Jurafsky D., Smith, N.A., and Choi Y. 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.

Schmidt, A. and Wiegand, M. 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.

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

Warner, W., and Hirschberg. J. 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.

Waseem, Z., and Hovy, D. 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.

Yang Y., Kim J., Kim, Y., Ho, N., Thorne, J., and Yun, S. 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.

Yoder. M., Ng L., West Brown, D., and Carley., K. 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.

12. Language Resource References

Athena Research Center (2024, November 19). Verbal Aggression (VA) Database. Version 1.0.0 (automatically assigned). [Dataset (Text corpus)]. CLARIN:EL. http://hdl.handle.net/11500/CLARIN-EL-0000-0000-7B9E-5

Athena Research Center (2025, September 30). Hate Crimes Database. Version 1.0.0 (automatically assigned). [Dataset (Text corpus)]. CLARIN:EL. http://hdl.handle.net/11500/CLARIN-EL-0000-0000-7BB7-8

Athena Research Center (2020, January 07). ILSP event extractor for physical attacks in Greek. Version 1.0.0. Athena Research Center. [Software (Tool/Service)]. https://doi.org/10.57771/mdj5-c740

Athena Research Center (2022). Twitter Verbal Aggression Analyzer (English). Version 1.0.0 (automatically assigned). [Software (Tool/Service)]. CLARIN:EL. http://hdl.handle.net/11500/CLARIN-EL-0000-0000-7596-3