*Legal Knowledge and Information Systems*
*G. Sileno et al. (Eds.)*


*doi:10.3233/FAIA230981*

# Multilevel Hate Speech Classification Based on Multilingual Case-Law

Monica PALMIRANI[a1], Chiara CATIZONE[a],
Giulia VENDITTI[a], Salvatore SAPIENZA[a]

[a] *CIRSFID-ALMA-AI, University of Bologna*

**Abstract.** This paper presents classification tools to detect hate speech topics using NLP tools in four different languages (Italian, Spanish, Germany, English) using the selected case-law from national and international jurisdiction. The research is conducted inside the FAST-LISA European project with the aim to classify the hate speech in online public debate.

**Keywords.** Classification AI, NLP, hate speech, legal concepts.

## 1. Introduction

Digital tools facilitate the exchange of information and ideas, promoting freedom of expression and human rights. However, they also allow for the spread of harmful content, such as online hate speech, due to anonymity and filter bubbles [9]. Online hate speech takes various forms and targets individuals or groups with discriminatory or pejorative language[2]. Hate incidents have been on the rise, fuelled by factors like migration, economic crises, anti-gender narratives, and conspiracy theories, especially during the COVID-19 pandemic, with relevant impacts on freedom of expression and democratic participation. To combat this, international institutions like the European Commission[3], the Council of Europe[4], the United Nations and its member states have been working to define and address hate speech and hate crime [12]. Despite the adoption of AI tools for moderation, automated systems have proven insufficient, and further efforts are needed to effectively tackle online hate speech [2, 3, 7]. The FAST LISA project[5] addresses the challenges of combating hate speech through a human-AI approach, integrating different competences and approaches: legal, linguistic, technical, ethical. Its main objective is to define, detect, monitor, and raise awareness against hate speech in the public debate for favouring a correct and effective debate during the public consultations opened by the Municipalities involved in the project[6]. This paper presents a Natural Language Processing (NLP) study conducted for creating an innovative classification about hate speech starting from linguistic and legal knowledge in four different languages with the goal of matching legal crimes with common *day-by-day* language used in the public discourse.

---

[1] E-mail: {monica.palmirani, chiara.catizone, giulia.venditti, salvatore.sapienza}@unibo.it
[2] United Nations. 'UN Strategy and Plan of Action on Hate Speech 18 June SYNOPSIS'. UN. Nations United. 2023. 'What Is Hate Speech?' United Nations. United Nations. 2023.
[3] Communication from the Commission to the European Parliament and the Council, A more inclusive and protective Europe: extending the list of EU crimes to hate speech and hate crime
[4] https://www.coe.int/en/web/combating-hate-speech/council-of-europe-on-hate-speech
[5] FAST LISA project is funded by the European Commission's CERV program.
[6] Municipality of Ravenna and Santa Coloma de Gramenet, association of Offenbach am Main. See the official portal: https://fastlisa.eu

## 2. Related Work

In the rapidly evolving digital landscape, the importance of developing robust NLP techniques for detecting hate speech across diverse languages and social contexts is emphasized. Various approaches, including lexicon-based methods, supervised machine learning models, multilingual NLP models, and domain-specific techniques, have shown promise in effectively identifying hate speech. Several studies [e.g., 10, 12] have utilized lexicon-based methods, leveraging annotated corpora to create taxonomies of offensive language. The use of TF-IDF measures and pre-trained models has been instrumental in identifying hate speech nuances. Addressing low-resource languages, [8] introduced a multilingual model, demonstrating the superiority of multilingual models over monolingual ones. Domain-specific hate speech classification has been explored in studies such as [4] and [6], which developed specialized models for specific social contexts. Supervised machine learning models have been extensively examined by studies like [1] and [13], achieving promising results with various classifiers, including transformer-based language models. These findings highlight diverse avenues for effectively detecting hate speech, underscoring the need for continued research to address evolving trends and promote online safety. The projects conducted by [3] and [7] are considered a valuable starting point, approaching the problem with different NLP and AI techniques.

## 3. Methodology

FAST LISA aims to create an interdisciplinary solution for combating hate speech online through a combination of AI tools and human-centred approaches oriented to educate digital citizens, in particular during public consultations supported by Municipalities or associations, that take place to discuss problems related to the territory and the day-by-day life of citizens, such as immigrates, building regulation, etc. The project develops an innovative methodology using four different levels of knowledge: 1) **Political level**: the language of the political and public discourse. The proposed NLP tools are tailored for institutions, citizens, and for the context, with media as the main channel; 2) **User-centred[7] level**: user-personas focused use-cases, centred around citizens and employees that are subject of supposed hate speech actions; 3) **Legal level**: legal domain-oriented lexicon, connecting legal language with concepts for classifying criminal speech and language misuse; 4) **Linguistic level**: consideration of language pragmatics and situations for hate speech classification, beyond just relying on lexicon.

Using these four axes as inputs, our goal is building a classification of the typologies of hate speech from the bottom words included in the short posts of citizens in the public debate discourse. We utilize text mining techniques for Multilingual Applications of NLP to support the creation of a trans-border classification of online hate speech, focusing on identifying involved abusive behaviours, targets, and determining whether they qualify as hate crimes in the respective partner countries. To achieve this, we first recognize the need to comprehend the differences and similarities between the partner countries' norms, social behaviours, and language. The process follows includes the following steps: i) legal analysis of the case-law and legislation by legal experts in each country and at international level; ii) linguistic analysis in the light of pragmatic approach using *slang* databases (e.g., Musixmatch lyrics of the songs); iii) development of the model and the preliminary vocabulary in KNIME; iv) validation of the correctness of the application of the

[7] In particular, we use Human Computer Interaction methodology in conjunction also with the AI guidelines principles of the "Ethics Guidelines for Trustworthy AI" of the European Commission.

classification categories under legal, social, linguistic points of view; iv) graphical visualisation; v) evaluation of the results with all the experts of the consortium.

## 4. KNIME Pipeline and Agile Prototyping

We use KNIME Analytics Platform, utilizing various text mining techniques to extract relevant textual features from five legal corpora[8]. Starting from previous literature we selected the following technical approaches to extract important classes from legal documents retrieved by domain experts in the Italian, German, Spanish and English languages. Our pipeline consists of the following steps:

1) **Corpora exploration and cleaning**. Bearing in mind the definition of hate speech provided by the UN in the absence of consolidated definitions, we used five legal corpora for extracting the hate speech vocabulary. As presented in Table 1, they are in line with the judiciaries systems and combine the facts with the crime of hate speech.

| CORPUS/SOURCE | N. DOCUMENTS | DESCRIPTION | APPLICABLE LAW |
|---|---|---|---|
| European Courts of Human Rights (ECHR) | 144 | Judgments mainly related to freedom of expression, freedom of religion, private life, and discrimination against sexual orientations | European Convention on Human Rights art. 8, 10, |
| United Kingdom, Court of Appeal (EWCA) | 21 | Cases address fear arousing, material dissemination, defamation, menaces, and homophobic behaviours | Public Order Act 1986; Public Order Act (Northern Ireland) 1987; Protection from Harassment Act 1997, Section 4; Crime and Disorder Act 1998; Terrorism Act 2000, Section 58 |
| Germany, German Federal Court of Justice (BGH) and the German Federal Constitutional Court (BVerfG). | 24 | Sedition (Volksverhetzung), trivialization (Verharmlosens) and denial (Leugnens), depiction of violence (Gewaltdarstellung), and hate speech on the internet (Hassrede auf Internetplatform). | German Civil Code (BGB); Law on the Federal Constitutional Court (BVerfGG); Network Enforcement Act (NetzDG); German Penal Code (StGB); Telecommunications Telemedia Data Protection Act (TTSDG). |
| Italy, Italian Supreme Court of Cassation (Corte Suprema di Cassazione) and local Courts of Appeal | 22 | Judgments on hate speech perpetrated on the internet, involving ethnic, racial, and religious hatred, incitement to violence, propaganda, and defamation. | Italian Constitution, articles 21 and 49; Law 654/1975. |
| Spain, Supreme Court of Spain (Tribunal Supremo) | 85 | Cases on apologizing for crimes, making threats, spreading false statements, insulting individuals, inciting hate and discrimination, and promoting or justifying terrorism crimes | Spanish Penal Code |

Table 1: Corpora description

In the pre-processing stages the text underwent several key NLP techniques. Firstly, they removed noise factors such as URLs, numerical values, and stop words to improve the clarity of the results. The use of OpenNLP NER and Parts of speech (POS) tagging were employed to filter out personal names, places, and dates, ensuring compliance with data protection laws. tagging was conducted to label words in the text with their respective parts of speech, such as adjectives, nouns, and verbs. Different tokenization models were used for English, German, Spanish, and Italian corpora, with specific models for each language. For instance, English used the Stanford Tagger with the left3words model, while German and Spanish employed Universal Dependencies (UD) models. Italian text was processed using the POS Tagger (Italian) metanode in KNIME, which relies on a list of tagged Italian terms. Then, the Stanford Lemmatizer was applied to group similar terms, and custom filter lists were used to remove legal

[8] KNIME is a user-friendly open-source platform for data integration, processing, analysis, and exploration. It's ideal for quick prototyping and allows non-technical experts to review and adjust results. Nodes are the core building blocks used to create workflows, making the entire analysis process documented and shareable with ease.

vocabulary terms and other specific categories. Finally, the pre-processed documents underwent stemming to group similar terms together, which enhances the quality of analysis.

2) **TF-IDF**. After pre-processing we applied Term Frequency-Inverse Document Frequency (TF–IDF) to detect non-relevant words that appear frequently in the corpus. In our KNIME workflow, we computed the TF–IDF values for unigrams in each corpus, setting a lower bound value between 0.01 and 0.02 due to the previous heavy filtering of input documents. We then filtered out the resulting terms from the text and computed bigrams, also extracting their TF–IDF values in the filtered corpora.
3) **Topic modelling**. Topic modelling is an unsupervised NLP task used to detect main themes in text data. It involves representing each topic with a list of relevant and weighted words. To determine the optimal number of clusters-topics for each corpus, we employed the Elbow Method. Then, we used Gensim's Latent Dirichlet Allocation (LDA) model for computing topic models. LDA assumes documents are mixtures of topics, and topics are mixtures of words. It finds topics by looking at word co-occurrence patterns in documents. After selecting the optimal number of topics using the Elbow Method, we extracted and visualized the topics and their co-occurrence values. Table 2 displays the above-mentioned parameters.

| CORPUS | ITERATIONS | CHUNK SIZE | PASSES | PERPLEXITY | CLUSTERS SQUARED ERROR |
|---|---|---|---|---|---|
| EHRC | 5000 | 145 | 100 | -6.58 | 471.162, 225.829, 122.99, **70.218**, 48.397 |
| UK | 10000 | 21 | 500 | -6.56 | 315.864, 140.551, 106.101, 53.438, 48.774, **45.582**, 37.762 |
| Germany | 10000 | 23 | 500 | -7.48 | 809.622, 459.92, 238.095, **190.896**, 181.193 |
| Italy | 10000 | 22 | 500 | -6.69 | 213.025, 115.415, 68.694, **51.487**, 42.153 |
| Spain | 50000 | 85 | 500 | -7.25 | 712.019, 332.941, 205.265, **156.424**, 136.117 |

Table 2: Table with LDA model values. “Iterations” is the maximum number of iterations for inferring topic distribution; “chunk size” is the number of documents per training chunk; “passes” is the number of passes through the corpus during training; “perplexity” measures predictive ability using log likelihood; “clusters squared errors” is the difference between observed and predicted values for each corpus, representing t+1 (t = number of selected topics).

4) **Word2Vec and Doc2Vec**. Word vectors, also called word embeddings, are real-valued vectors representing words in multi-dimensional space to capture their semantic meaning. We used the Word2Vec Learner node in KNIME to train a Word2Vec model on unlabelled documents, focusing on the skip-gram model, which predicts surrounding words. t-SNE reduced dimensionality and created scatterplots for word embeddings. Doc2Vec creates numeric representations of documents using the PV-DM model. In our KNIME workflow, we used the Doc2Vec Learner node to extract document vectors, which we combined with topic modeling results to assign documents to prominent topics. Finally, t-SNE reduced dimensionality to create 3-D and 2-D scatterplots from words and documents embeddings.
5) **Distance Matrix**. In NLP, the Euclidean distance is a common similarity measure used with word embeddings. It quantifies the dissimilarity or similarity between two points in Euclidean space based on their Cartesian coordinates. By calculating the length of the line segment connecting these points, we can determine how far apart or close together words are in the embedding space. In our KNIME workflow, we used the Distance Matrix Calculate and Distance Matrix

## 5. Evaluation and Results

The following table describes and comments detailed findings from each corpus. Overall, they share common themes of addressing hate and discrimination. Words between quotation marks represent significant terms for each topic. Topic labels have been selected from the taxonomy entitled "The Future of Free Speech" [9] created for the ECHR jurisdiction and used in similar studies, like [15].

| | | | | | |
|---|---|---|---|---|---|
| **ECHR** | **Religious Hatred** | **Ideologies** | **Genocide Denial** | **Ethnic Hatred** | **Homophobia/Gender Hatred** |
| | Involves public statements expressing "violence" and "offences" against religion, including physical and "damaging" activities | Involves public "opinion" and information expressing political ideologies and activities related to "interference" and political "parties" | Genocide denial focuses on public communication via internet and the role played by internet "internet" as regards "respect" and "crime" | Abusive behaviours are identified as: 1) prejudice, also against specific minorities (e.g., "Armenians"); 2) "violence", "incitement", and "defamation" | Violence, incitement, degradation, defamation, and harassment are the identifiable abusive behaviours towards gender, sexual orientation, and "family" |
| **UNITED KINGDOM** | **Religious Hatred** | **Religious Extremism** | **Harassment** | **Ethnic Hatred** | **Homophobia/Gender Hatred** |
| | Religious hatred is related to the public discourse with reference to specific religious beliefs. It involves "education" and "challenges" | Religious extremism is connected to xenophobia. "Jihad" "terrorist" "muslim" and "Islam" are some of the most relevant terms. | Harassment relates to personal communication ("messages") with offences or threats, via personal devices ("telephone") or other communication media ("post") | Characteristics of an ethnicity, and race ("racial", "blsck"), and a victim-centric approach, as well as effects ("abuse") are considered | Offences against the LGBTQ+ community and sexual preferences emerge from advertisement and religious discourse. Online content ("comment") is considered. |
| **GERMANY** | **Violent Attacks** | **Xenophobia** | **Harassment** | **Anti-Semitism** | |
| | Cases revolve around violent communication and opinions against the society or disadvantaged people ("benachteil") with the use of force ("gewalt") as its peculiar trait. Surprisingly, "schiff" ("boat") emerges as a relevant term. After further analysis, it seems that some cases discuss communication against migrants travelling by boats | It relates to critiques and hate messages ("hassbotschaft") in social networks and their functioning ("profile", "Facebook", "platformbetrieb") against migrants ("fluchtling") crossing borders ("grenz") | Social networks play a major role in cases pertaining to harassment. Words such as "fear" ("fried"), "killing" ("mord"), "attack" ("angegriff") appear to be relevant | Social networks are involved in cases related to hate against Jews ("jude") and, specifically women ("frau"). Also, genocide-related contents are displayed ("volkermord", "konzetrationslager","historisch" ) , but it seems specific towards the Jewish community | |
| **ITALY** | **Ethnic Hatred** | **Genocide Denial** | **Online Harassment** | **Religious Hatred** | |
| | Cases of hate ("odio") around ethnicity ("etnia", "razza") involve both social media ("Facebook") and public discourse ("politica", "fascismo"). Freedom of expression ("espressione", "manifestazione") is mentioned. Likely, a right-wing specific political party is mentioned. | Historical ("storico") discourse seems to play a role in Holocaust ("Olocausto") denial. Moreover, ethnic minorities different from Jews (e.g., "clandestini", "razza") seem to play a role | The lexicon used is more specific towards online content and users ("post", "utente"). Online harassement is involves racist attacks ("razzista", "etnico") including concepts such as discrimination ("discriminazione"), propaganda ("propaganda"), violence ("violenza"), incitement ("incitamento") | Religious ("religione") hatred targets religious groups, such as Jews ("ebreo", "anti-semita") and Islamic communities ("Islam"), mixed with racism ("discriminazione", "razziale") | |
| **SPAIN** | **Religious Hatred** | **Harassment** | **Hate incitement** | **Racial Hatred** | |
| | Religious ("religion") hatred ("odio") has a significant dimension in the public sphere ("libertad", "publico", "expresion"). It is connected to religious sentiment ("sentimiento"), belief ("credencia") values ("valor"), in relation to dignity ("dignidad"). It consists of violent ("violencia"), discriminatory ("discriminación") or inciting ("incitación") | Harassment consists of threatening ("amenazar") violence and injuries ("lesion", "injuria") towards victims ("victima"). Words such as sexuality ("sexualidad") and dignity ("dignidata") suggest a broad interpretation of this category in the Spanish courts. Interestingly, videos ("video") and messages ("mensaje") are also mentioned as harassment media | It primarily focuses on the context of hate ("odio") incitement ("concerto", "cancion", "musical") or expression ("diffusion", "provocation"). Race ("raza") emerges as the hate target, as well as disability ("inhabilitad") | "People" ("pueblo"), race ("negro", "blanco"), and ethnicity ("judios") appear to be closely related to hate ("odio") and associated to danger ("peligroso"), war ("guerra") and politics ("politica"). | |

[9] Available at https://futurefreespeech.com/hate-speech-case-database/

## 6. Conclusions

The present research work demonstrated the necessity of having a multi-level classification tool to the complex phenomena of hate speech because the common-sense vocabulary is not enough. First, not all the harming words have been classified as hate speech legally speaking, so the first level of the classification must distinguish what is really *hate speech* and what is a mis-usage of the language or foul language (HateSpeech-HS- or no-HS). Secondly, we distinguish the phenomena considering the behaviour/fact (e.g., sexual insulting, trolling, etc.) and the target of the hate speech (e.g., women, minor). Finally, mapping the misconduct with the type of crimes could help to assign a different weight to the content and so to permit to evaluate counter measurement and remedies.

**Acknowledgements.** We thank FAST LISA project co-funded by the Citizens, Equality, Rights and Values programme (CERV), contract number 101049342 by European Commission.

## References

[1] Ababu TM, and Michael MW. Afaan Oromo Hate Speech Detection and Classification on Social Media. In Proceedings of the Thirteenth Language Resources and Evaluation Conference; 6612–19; 2022

[2] Alraddadi RA, El-Khalil Ghembaza, MI. Automatic Detection and Classification of Anti-Islamic Web Text-Contents. In:(?) Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering; 2022;LNICST, 438 LNICST, pp. 162-181. doi: https://doi.org/10.1007/978-3-031-04409-0_33

[3] Ollagnier A, Cabrio E, Villata S. Unsupervised fine-grained hate speech target community detection and characterisation on social media. 2023; Soc. Netw. Anal. Min. 13(1): 58. doi: https://doi.org/10.1007/s13278-023-01061-4

[4] Beyhan F, Carık B, Arın İ, Terzioğlu A, Yanikoglu B, Yeniterzi R. A Turkish Hate Speech Dataset and Detection System. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 4177–85; 2022

[5] Dietz C, Berthold MR. KNIME for Open-Source Bioimage Analysis: A Tutorial. In: De Vos WH, Munck S, Timmermans JP, editors. Focus on Bio-Image Informatics. 2016; 219:179–97. Advances in Anatomy, Embryology and Cell Biology. Cham: Springer International Publishing. Doi: 10.1007/978-3-319-28549-8_7

[6] Armend D, Casadei C, Tosi M, Celli F. Hate versus Politics: Detection of Hate against Policy Makers in Italian Tweets. 2021; SN Social Sciences 1 (9): 223. doi: 10.1007/s43545-021-00234-2

[7] Corazza M, Menini S, Cabrio E, Tonelli S, Villata S. A Multilingual Evaluation for Online Hate Speech Detection. 2020 ; ACM Trans; Internet Techn;20(2): 10:1-10:22 . doi: https://doi.org/10.1145/3377323Nozza D, Bianchi F, Attanasio G. HATE-ITA: Hate Speech Detection in Italian Social Media Text. In: Proceedings of the Sixth Workshop on Online 164 Abuse and Harms (WOAH), 252–60. Seattle, Washington (Hybrid): Association for Computational Linguistics. 2022. doi: 10.18653/v1/2022.woah-1.24

[8] Ruiter D, Reiners L, Geet D'Sa A, Kleinbauer T, Fohr D, Illina I, Klakow D, Schemer C, Monnier A. Placing M-Phasis on the Plurality of Hate: A Feature-Based Corpus of Hate Online. 2022. doi: 10.48550/arXiv.2204.13400

[9] Salminen J, Almerekhi H, Milenković M, Jung S, An J, Kwak H, Jansen B. Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. Proceedings of the International AAAI Conference on Web and Social Media 12 (1). 2018. doi: https://doi.org/10.1609/icwsm.v12i1.15028

[10] De Varennes F.Effective Guidelines on Hate Speech, Social Media and Minorities. 2022.

[11] Vargas F, Carvalho I, Rodrigues de Góes F, Pardo T, Benevenuto F. HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 7174–83; 2022; Marseille, France; European Language Resources Association. doi: 10.48550/arXiv.2103.14972Vrysis L, Vryzas N, Kotsakis K, Saridou T, Matsiola M, Veglis A, Arcila-Calderón A, and Dimoulas C. A Web Interface for Analyzing Hate Speech. 2021; Future Internet 13 (3): 80. doi: https://doi.org/10.3390/fi13030080