future internet

MDPI

Article

A Web Interface for Analyzing Hate Speech

Lazaros Vrysis [1], Nikolaos Vryzas [1], Rigas Kotsakis [1], Theodora Saridou [1], Maria Matsiola [1], Andreas Veglis [1,*], Carlos Arcila-Calderón [2] and Charalampos Dimoulas [1]

[1] School of Journalism & Mass Communication, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; lvrysis@auth.gr (L.V.); nvryzas@auth.gr (N.V.); rkotsakis@auth.gr (R.K.); saridout@jour.auth.gr (T.S.); mmat@jour.auth.gr (M.M.); babis@eng.auth.gr (C.D.)

[2] Facultad de Ciencias Sociales, Campus Unamuno, University of Salamanca, 37007 Salamanca, Spain; carcila@usal.es

* Correspondence: veglis@jour.auth.gr

Abstract: Social media services make it possible for an increasing number of people to express their opinion publicly. In this context, large amounts of hateful comments are published daily. The PHARM project aims at monitoring and modeling hate speech against refugees and migrants in Greece, Italy, and Spain. In this direction, a web interface for the creation and the query of a multi-source database containing hate speech-related content is implemented and evaluated. The selected sources include Twitter, YouTube, and Facebook comments and posts, as well as comments and articles from a selected list of websites. The interface allows users to search in the existing database, scrape social media using keywords, annotate records through a dedicated platform and contribute new content to the database. Furthermore, the functionality for hate speech detection and sentiment analysis of texts is provided, making use of novel methods and machine learning models. The interface can be accessed online with a graphical user interface compatible with modern internet browsers. For the evaluation of the interface, a multifactor questionnaire was formulated, targeting to record the users' opinions about the web interface and the corresponding functionality.

Keywords: hate speech detection; natural language processing; web interface; database; machine learning; lexicon; sentiment analysis; news semantics

Citation: Vrysis, L.; Vryzas, N.; Kotsakis, R.; Saridou, T.; Matsiola, M.; Veglis, A.; Arcila-Calderón, C.; Dimoulas, C. A Web Interface for Analyzing Hate Speech. *Future Internet* **2021**, *13*, 80. https://doi.org/10.3390/fi13030080

Academic Editor: Devis Bianchini

Received: 26 February 2021
Accepted: 18 March 2021
Published: 22 March 2021

1. Introduction

In today's ubiquitous society, we experience a situation where digital informing and mediated communication are dominant. The contemporary online media landscape consists of the web forms of the traditional media along with new online native ones and social networks. Content generation and transmission are no longer restricted to large organizations and anyone who wishes may frequently upload information in multiple formats (text, photos, audio, or video) which can be updated just as simple. Especially, regarding social media, which since their emergence have experienced a vast expansion and are registered as an everyday common practice for thousands of people, the ease of use along with the immediacy they present made them extremely popular. In any of their modes, such as microblogging (like Twitter), photos oriented (like Instagram), etc., they are largely accepted as fast forms of communication and news dissemination through a variety of devices. The portability and the multi-modality of the equipment employed (mobile phones, tablets, etc.), enables users to share, fast and effortless, personal or public information, their status, and opinions via the social networks. Thus, communications nodes that serve many people have been created minimizing distances and allowing free speech without borders; since more voices are empowered and shared, this could serve as a privilege to societies [1–8]. However, in an area that is so wide and easily accessible to large audiences many improper intentions with damaging effects might be met as well, one of which is hate speech.

It is widely acknowledged that xenophobia, racism, gender issues, sexual orientation, and religion among others are topics that trigger hate speech. Although no universally agreed definition of hate speech has been identified, the discussion originates from discussions on freedom of expression, which is considered one of the cornerstones of a democracy [9]. According to Fortuna and Nunes (2018, p. 5): "Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used" [10]. Although international legislation and regulatory policies based on respect for human beings prohibit inappropriate rhetoric, it finds ways to move into the mainstream, jeopardizing values that are needed for societal coherence and in some cases relationships between nations, since hate speech may fuel tensions and incite violence. It can be met towards one person, a group of persons, or to nobody in particular [11] making it a hard to define and multi-dimensional problem. Specifically, in Europe as part of the global North, hate speech is permeating public discourse particularly subsequent to the refugee crisis, which mainly - but not only- was ignited around 2015 [12]. In this vein, its real-life consequences are also growing since it can be a precursor and incentive for hate crimes [13].

Societal stereotypes enhance hate speech, which is encountered both in real life and online, a space where discourses are initiated lately around the provision of free speech without rules that in some cases result to uncontrolled hate speech through digital technologies. Civil society apprehensions led to international conventions on the subject and even further social networking sites have developed their own services to detect and prohibit such types of expressed rhetoric [14], which despite the platforms' official policies as stated in their terms of service, are either covert or overt [15]. Of course, a distinction between hate and offensive speech must be set clear and this process is assisted by the definition of legal terminology. Mechanisms that monitor and further analyze abusive language are set in efforts to recognize aggressive speech expanding on online media, to a degree permitted by their technological affordances. The diffusion of hateful sentiments has intrigued many researchers that investigate online content [11,13,15,16] initially to assist in monitoring the issue and after the conducted analysis on the results, to be further promoted to policy and decision-makers, to comprehend it in a contextualized framework and seek for solutions.

Paz, Montero-Díaz, and Moreno-Delgado (2020, p.8) refer to four factors, media used to diffuse hate speech, the subject of the discourse, the sphere in which the discourse takes place, and the roots or novelty of the phenomenon and its evolution that each one offers quantification and qualification variables which should be further exploited through diverse methodologies and interdisciplinarity [17]. In another context, anthropological approaches and examination of identities seek for the genealogy through which hate speech has been created and sequentially moved to digital media as well as the creation of a situated understanding of the communication practices that have been covered by hate speech [13]. Moreover, on the foundation to provide a legal understanding of the harm caused by hateful messages, communication theories [18] and social psychology [19] are also employed. Clearly, the hate speech problem goes way back in time, but there are still issues requiring careful attention and treatment, especially in today's guzzling world of social media and digital content, with the vast and uncontrollable way of information publishing/propagations and the associated audience reactions.

1.1. Related Work: Hate Speech in Social Media and Proposed Algorithmic Solutions

Hate speech has been a pivotal concept both in public debate and in academia for a long time. However, the proliferation of online journalism along with the diffusion of user-generated content and the possibility of anonymity that it allows [20,21] has led to the increasing presence of hate speech in mainstream media and social networks [22,23].

During the recent decades, media production has often been analyzed through the lens of citizen participation. The idea of users' active engagement in the context of mainstream media was initially accompanied by promises of enhancing democratization and strengthening bonds with the community [24,25]. However, the empirical reality of user participation was different from the expectations, as there is lots of dark participation, with examples ranging from misinformation and hate campaigns to individual trolling and cyberbullying; a large variety of participation behaviors are evil, malevolent, and destructive [22]. Journalists identify hate speech as a very frequently occurring problem in participatory spaces [8]. Especially comments, which are considered an integral part of almost every news item [26], have become an important section for hate speech spreading [27].

Furthermore, an overwhelming majority of journalists argue that they frequently come upon hate speech towards journalists in general, while most of them report a strong increase in hate speech personally directed at them [28]. When directed at professionals, hate speech can cause negative effects both on journalists themselves and journalistic work: it might impede their ability to fulfill their duties as it can put them under stark emotional pressure, trigger conflict into newsrooms when opinions diverge on how to deal with hateful attacks or even negatively affect journalists' perception of their audience [28]. Hence, not rarely, professionals see users' contributions as a necessary evil [27] and are compelled to handle a vast amount of amateur content in tandem with their other daily tasks [29].

To avoid problems, such as hate speech, and protect the quality of their online outlets, media organizations adopt policies that establish standards of conduct and restrict certain behaviors and expressions by users [30]. Community managers are thus in charge of moderating users' contributions [31], by employing various strategies for supervising, controlling, and enabling content submission [32]. When pre-moderation is followed, every submission is checked before publication and high security is achieved. However, this method requires considerable human, financial, and time resources [27]. On the other hand, post-moderation policies lead to a simpler and more open approach but can lower the quality [33], exposing the platform to ethical and legal risks. Apart from manual moderation, some websites utilize artificial intelligence techniques to tackle this massive work automatically [34], while others implement semi-automatic approaches that assist humans through the integration of machine learning into the manual process [35].

The automation of the process of hate speech detection relies on the training and evaluation of models, using annotated corpora. The main approaches include lexicon-based term detection and supervised machine learning. Lexicons contain a list of terms, along with their evaluation concerning the relation to hate speech. The terms are carefully selected and evaluated by experts on the field, and they need to be combined with rule-based algorithms [36–38]. Such algorithms are based on language-specific syntax and rules. Computational models such as unsupervised topic modeling can lead to insight regarding the most frequent terms that allow further categorization of the hate-related topics [39,40]. In supervised machine learning approaches, models are trained using annotated corpora. Baseline approaches rely on bag-of-words representations combined with machine learning algorithms [36–41]. More recent methods rely on deep learning and word embeddings [42,43]. The robustness of a supervised machine learning algorithm and its ability to generalize for the detection of hate in unseen data relies on the retrieval of vast amounts of textual data.

Big data analytics of social media contents is an emerging field for the management of the huge volumes that are created and expanded daily [44]. Most social media services offer dedicated application programming interfaces (APIs) for the collection of posts and comments, to facilitate the work of academics and stakeholders. Using a dedicated API, or a custom-made internet scraper makes it easy to retrieve thousands of records automatically. Twitter is the most common choice, due to the ease-of-use of its API, and its data structure that makes it easy to retrieve content relevant to a specific topic

[36,37,41]. While textual analysis is the core of hate speech detection, metadata containing information about the record (e.g., time, location, author, etc.) may also contribute to model performance. Hate speech detection cannot be language-agnostic, which means that a separate corpus or lexicon and methodology needs to be formed for every different language [36,37,45]. Moreover, a manual annotation process is necessary, which, inevitably introduces a lot of human effort, as well as subjectivity [36]. Several annotation schemes can be found in literature, differing in language, sources (e.g., Twitter, Facebook, etc.), available classes (e.g., hate speech, abusive language, etc.), and ranking of the degree of hate (e.g., valence, intensity, numerical ranking, etc.) [37]. The selected source itself may influence the robustness of the algorithmic process. For instance, Twitter provides a maximum message length, which can affect the model fitting in a supervised training process [36]. Multi-source approaches indicate the combination of different sources for analytics [46]. In [47] for example, Twitter data from Italy are analyzed using computational linguistics and the results are visualized through a Web platform to make them accessible to the public.

1.2. Project Motivation and Research Objectives

Based on the preceding analysis, there is missing a multilingual hate-speech detection (and prevention) web-service, which individuals can utilize for monitoring informatory streams with questionable content, including their own user-generated content (UGC) posts and comments. More specifically, the envisioned web environment targets to offer an all-in-one service for hate speech detection in text data deriving from social channels, as part of the Preventing Hate against Refugees and Migrants (PHARM) project. The main goal of the PHARM project is to monitor and model hate speech against refugees and migrants in Greece, Italy, and Spain to predict and combat hate crime and also counter its effects using cutting-edge algorithms. This task is supported via intelligent natural language processing mechanisms that identify the textual hate and sentiment load, along with related metadata, such as user location, web identity, etc. Furthermore, a structured database is initially formed and dynamically evolving to enhance precision in subsequent searching, concluding in the formulation of a broadened multilingual hate-speech repository, serving casual, professional, and academic purposes. In this context, the whole endeavor should be put into test through a series of analysis and assessment outcomes (low-/high-fidelity prototypes, alpha/beta testing, etc.) to monitor and stress the effectiveness of the offered functionalities and end-user interface usability in relation to various factors, such as users' knowledge and experience background. Thus, standard application development procedures are followed through the processes of rapid prototyping and the anthropocentric design, i.e., the so-called logical-user-centered-interactive design (LUCID) [48–52]. Therefore, audience engagement is crucial, not only for communicating and listing the needs and preferences of the targeted users but also for serving the data crowdsourcing and annotating tasks. In this perspective, focusing groups with multidisciplinary experts of various kinds are assembled as part of the design process and the pursued formative evaluation [50–52], including journalists, media professionals, communication specialists, subject-matter experts, programmers/software engineers, graphic designers, students, plenary individuals, etc. Furthermore, online surveys are deployed to capture public interest and people's willingness to embrace and employ future Internet tools. Overall, following the above assessment and reinforcement procedures, the initial hypothesis of this research is that it is both feasible and innovative to launch semantic web services for detecting/analyzing hate speech and emotions spread through the Internet and social media and that there is an audience willing to use the application and contribute. The interface can be designed as intuitively as possible to achieve high efficiency and usability standards so that it could be addressed to broader audiences with minimum digital literacy requirements. In this context, the risen research questions (RQ) elaborated to the hypotheses are as follows:

RQ1: Is the PHARM interface easy enough for the targeted users to comprehend and utilize? How transparent the offered functionalities are?
RQ2: What is the estimated impact of the proposed framework on the journalism profession and the anticipated Web 3.0 services? Are the assessment remarks related to the Journalism profession?

2. Materials and Methods

As a core objective of the PHARM project is to build a software environment for querying, analyzing, and storing multi-source news and social media content focusing on hate speech against migrants and refugees, a set of scripts for Natural Language Processing (NLP) has been developed, along with a web service that enables friendly user interaction. Let the former be called the PHARM Scripts, the latter the PHARM Interface, and both of them the PHARM software. All these implementations are constantly elaborated and updated as the project evolves, the source code of the PHARM Scripts, along with the required documentation, is publicly available as a GitHub repository (http://github.com/ thepharmproject/set_of_scripts, accessed on 18 March 2021), while the PHARM Interface has the form of a website (http://pharm-interface.usal.es, accessed on 18 March 2021). The detailed documentation of the algorithms is out of the scope of the current work, so only a brief presentation of the relevant functionality follows. Comprehensive documentation and use instructions for the interface are available online (http://pharm-interface.usal. es/instructions, accessed on 18 March 2021) in English, Greek, Italian and Spanish.

2.1. Data Collection

The core outcome of the PHARM software concern a multi-source platform for the analysis of unstructured news and social media messages. On the one hand, it is very important for hate speech texts to include both data (texts of the news or social media messages) and metadata (location, language, date, etc.). On the other hand, the diversity of the sources is unquestionable and thus, mandatory. Therefore, several sources have been selected for the collection of content related to hate speech, while all the required technical implementations have been made to collect the necessary data from these sources. The sources include websites in Greek, Italian, Spanish as well as Twitter, YouTube, and Facebook and concern articles, comments, tweets, posts, and replies. The list of the sources was initialized and updated by the media experts. The list of the sources includes 22 Spanish, 12 Italian, and 16 Greek websites that are prone to publishing hate speech content in the articles or the comments section. Site-specific scripts for scraping have been developed for the collection of semi-structured content (including the accompanying metadata) from the proposed websites, while content from open Facebook groups and pages, as well as websites that are not included in the list, are supported using a site-agnostic scraping method. Tweets are gathered using a list of hashtags and filters containing terms relevant to anti-immigration rhetoric and YouTube comments are collected using search queries relevant to immigration. To store, query, analyze and share news and social media messages, PHARM software adopts a semi-structured format based on JSON (JavaScript object notation), adapted to media features.

2.2. Data Format

Taking into account the requirements of the project (i.e., the use of some relevant extra information for hate speech analysis), the sources that are used for scraping content (i.e., website articles and comments, YouTube comments, tweets), interoperability and compatibility considerations for importing and exporting data between the PHARM Interface, PHARM Scripts, and third-party applications, some general specifications for the data format have been set. The main field is the text (i.e., content), accompanied by the id, annotations, and meta fields. The meta field is a container that includes all metadata.

A minimum set of metadata is used for all platforms (i.e., type, plang, pdate, phate, psent, pterms, ploc). These fields are found for all records across different sources. Table 1 presents the proposed data scheme.

Table 1. The common fields of the specified data format.

Field	Description
id	unique identifier
text	content
annotations	hate speech and sentiment annotations
meta/type	type of text (tweet, article, post, comment, etc.)
meta/plang	language detection via PHARM Scripts
meta/pdate	datetime estimation via PHARM Scripts
meta/phate	hate speech detection via PHARM Scripts
meta/psent	sentiment analysis via PHARM Scripts
meta/pterms	frequent terms collection via PHARM Scripts
meta/ploc	geolocation estimation via PHARM Scripts
meta/meta	unsorted metadata

In the cases of web scraping, metadata depends on the available data provided by each site, whereas for YouTube comments and tweets, where the corresponding API is used, specific metadata have been selected and are stored along with the text. Table 2 demonstrates the fields that are used for data that originate from the Twitter and YouTube social media platforms.

Table 2. The metadata fields that are exploited for the YouTube and Twitter records.

Twitter	YouTube
tweet id	comment id
is retweet	reply count
is quote	like count
user id	video id
username	video title
screenname	channel
location	video description
follower count	author id
friend count	author name
date	date

2.3. Data Analysis

The most notable analysis methods that are used in the PHARM software concern date, time, and geolocation estimation, language detection, hate speech detection, and sentiment analysis. Various software libraries have been deployed for implementing the supported analysis methods, along with custom algorithms that have been developed specifically for the PHARM software. A brief description of these methods follows.

Language Detection: The PHARM software mainly processes text produced in Greek, Italian, and Spanish languages but many of the sources may contain texts in other languages or local dialects [53]. To work with these three national languages, a procedure to detect the language of the media text when it is not properly declared has been specified. An ensemble approach for improved robustness is adopted, querying various language detection libraries simultaneously. Amongst the used libraries are the textblob and googletrans Python libraries [54,55].

Geolocation Estimation: Geolocation identification of the collected texts is considered useful for analysis [53]. Therefore, a method for detecting geolocation from text data has

been implemented. Named entities are extracted from texts and geocoded i.e., the geographical coordinates are retrieved for the found entities. The named entities include geopolitical entities (GPE) (i.e., countries, cities, states), locations (LOC) (i.e., mountains, bodies of water), faculties (FAC) (buildings, airports, highways, etc.), organizations (ORG) (companies, agencies, institutions, etc.). For this method, the Nominatim geocoder, along with openstreetmap data are used [56].

Datetime Estimation: Besides location and language, when metadata is available, relevant extra information for hate speech analysis can be used. Some of this extra information, such as date or time, may be available in different formats, introducing the necessity of standardization. Therefore, a method for detecting and standardizing date and time information from meta- and text- data has been implemented. A couple of Python libraries (e.g., dateparser, datefinder, and parsedatetime) are exploited for detecting datetime objects in texts. This is based on metadata analysis, where date information is commonly present. If datetime detection fails for the metadata, the same workflow is applied to the text data.

Hate Speech Detection: Undoubtedly, hate speech detection is a core algorithmic asset for the project. Therefore, a couple of methods for detecting hate speech have been implemented, based on both an unsupervised and a supervised approach. The former concerns a lexicon-based method relying on a dictionary containing static phrases, along with dynamic term combinations (i.e., adjectives with nouns), while the latter refers to a machine learning procedure. For both methods, a language model is loaded (according to the language of the text) and common normalization practices are taking place (lower-casing, lemmatization, stop-word and punctuation removal). In the first case, the targeted terms are being searched and the text, while in the second, a recurrent neural network (RNN) undertakes the task of detecting hate speech [41,57]. For this reason, a pretrained tokenizer and a deep network consisting of an embedding layer, a gated recurrent unit (GRU) layer, and a fully connected layer are deployed. The models and the tokenizer have been trained using the Keras framework [58].

Sentiment Analysis: Similarly, two methods for sentiment analysis in the context of hate speech against refugees have been embedded in the interface [45,59]. These follow the same concepts as in hate speech detection but exploiting different lexicons and training data. The unsupervised method adopts many key aspects of the SentiStrength algorithm, such as the detection of booster, question, and negating words [60]. The supervised model for sentiment analysis follows the same architecture as the one for hate speech detection, trained on a different corpus.

Topic Modeling: The lexicons for both hate speech and sentiment analysis were developed by a team of experts in the field of journalism, communication, and media. To facilitate the process with automated text analysis, exploratory content processing techniques for topic modeling and entity collection have been deployed as well. The dictionaries have been built using frequent words, boosted by entity extraction based on ter frequency (TF) for absolute entity counting and term frequency-inverse document frequency (TF-IDF) for proportional counting, showing how important an entity is for the document or even the entire corpus [39,59].

2.4. Project Analysis and Usability Evaluation

The defined research questions and the elongated conclusions were supported by an empirical survey regarding the evaluation of the developed interface, while its statistical results are presented in the respective section of the manuscript. In this context, a multifactor questionnaire was formulated, targeting to record the users' opinions about the web interface and the corresponding functionalities. It has to be noted that in this section, an overview of the questionnaire is exhibited, along with the survey identity, while a detailed description can be accessed in the manuscript appendix. The evaluation process was categorized into 8 major factors, namely the efficiency (7 items), usability (12 items), learnability (5 items), satisfaction (11 items), navigation (4 items), content (6 items),

interactivity (5 items) and design (3 items) of the web interface. Furthermore, the participants were called to answer about the area that the web interface usage could cover according to their interests (5 items) and the scope of utilization in public information and awareness (6 items). Taking into account the main functionalities of the web interface, a stand-alone question was posed regarding the assessment of the contribution of the project towards the identification of hate speech mechanisms and sentiment loads in text data. All the aforementioned metrics were measured on a 5-level Likert scale, ranging from 1—strongly disagree to 5—strongly agree. The demographic questions that were involved in the survey addressed gender (male, female, other, no answer), age (18–22, 23–30, 31–40, 41–50, over 50), education (high school, vocational learning, bachelor, master, Ph.D.), computer familiarity (Likert scale 1–5), Internet familiarity (Likert scale 1–5), news awareness (Likert scale 1–5) and the profession (in 9 categories) of the participants. However, a dedicated binary inquiry was inserted recording if the participant works/ has worked as a journalist (yes-no), because of the additive value in the assessment of the web service that focuses on hate speech detection. Table 3 summarizes the set of questions that were implicated in the current survey.

Table 3. Overview of the formulated questionnaire.

#	Questions/Factors	Measure
1	Efficiency-7 items	Likert Scale 1–5
2	Usability-12 items	Likert Scale 1–5
3	Learnability 5 items	Likert Scale 1–5
4	Satisfaction-11 items	Likert Scale 1–5
5	Navigation-4 items	Likert Scale 1–5
6	Content-6 items	Likert Scale 1–5
7	Interactivity-5 items	Likert Scale 1–5
8	Design-3 items	Likert Scale 1–5
9	"Use for" scenarios-5 items	Likert Scale 1–5
10	Public Information and Awareness-6 items	Likert Scale 1–5
11	Contribution to hate speech/ sentiment detection	Likert Scale 1–5
12	Gender	Male/Female/Other/No answer
13	Age	18–22, 23–30, 31–40, 41–50, 50+
14	Education	Highschool, Vocational Learning, Bachelor, Master, PhD
15	Computer Familiarity	Likert Scale 1–5
16	Internet Familiarity	Likert Scale 1–5
17	News Awareness	Likert Scale 1–5
18	Profession	9 Categories
19	Working/has worked as Journalist	Binary Yes-No

The survey was conducted mainly via the social media channels of the authors, while the final number of gathered responses reached n = 64. The temporal length upon the completion of the survey ranged from 7 to 10 min, while in this duration the participants had to navigate and interact with the website interface and at the same time to answer the projected inquiries regarding its evaluation. The moderate number of responses can be justified on the multidisciplinary nature of the conducted research, which prerequisites more expertise and of course more time since several tasks were involved during the assessment process. Nevertheless, the aforementioned argument favors the reliability of the evaluation results because of the volunteered engagement of the 64 users in the survey. Finally, it has to be highlighted that this is the first time that the project is assessed, aiming at preliminary evaluation remarks for further optimizing crucial aspects of the interface based on the participants' opinions. A reliability test was conducted on the questionnaire,

based on Cronbach's alpha, revealing the respective coefficient a = 0.87, therefore supporting confident statistical results in the following section. Table 4 presents the basic demographic information of the respondents.

Table 4. Demographic Information of Participants.

#	Question	Answers-Distribution
1	Gender	Male (28.1%), Female (71.9%)
2	Age	18–22 (12.5%), 23–30 (43.8%), 31–40 (34.4%), 41–50 (4.7%), 50+ (4.7%)
3	Education	Highschool (14.1%), Vocational Learning (1.6%), Bachelor (31.3%), Master (45.3%), PhD (7.8%)
4	Computer Familiarity	1 (3.1%), 2 (29.7%), 3 (20.3%), 4 (34.4%), 5 (12.5%)
5	Internet Familiarity	1 (1.6%), 2 (18.8%), 3 (9.3%), 4 (54.7%), 5 (15.6%)
6	News Awareness	1 (0%), 2 (3.1%), 3 (21.9%), 4 (37.5%), 5 (37.5%)
7	Working/has worked as Journalist	Yes (46.9%), No (53.1%)

During the survey preparation, all ethical approval procedures and rules suggested by the "Committee on Research Ethics and Conduct" of the Aristotle University of Thessaloniki were followed.

3. Results

The results of the current research concern the presentation of the functionality and usability, along with the multi-faceted evaluation of the implemented web interface.

3.1. The Implemented Web Interface

The PHARM Interface serves as the front-end of the PHARM software. It is the graphical interface for exposing data and functionality to the users and relies on the back-end, which consists of the PHARM scripts. For the development of the interface, the Python web framework Flask has been used. The choice is justified, as the NLP and data analysis scripts are also written in Python and, following this approach, all the functionality of the interface can be included within a common software project. The graphical user interface (GUI) has been mainly designed in Bootstrap, a popular HTML, CSS, and JavaScript library. Additional HTML, CSS, and JavaScript blocks have been added where needed. The Flask project has been deployed on a virtual machine and is served using the Waitress, a production-quality pure-Python web server gateway interface (WSGI) with very acceptable performance. The PHARM Interface is accessible at http://pharm-interface.usal.es (accessed on 18 March 2021). The home screen of the Interface gives some basic information about the PHARM project and provides a starting point for accessing the supported NLP methods (Figure 1).

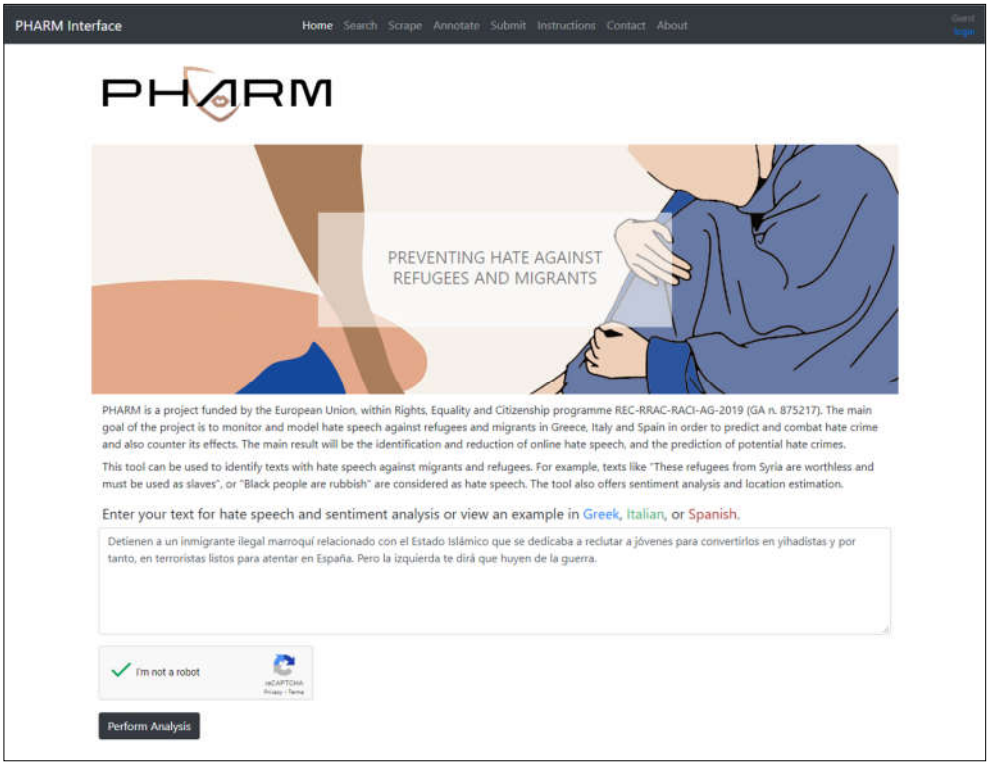

Figure 1. The home screen of the Preventing Hate against Refugees and Migrants (PHARM) web interface.

Let us further describe the software by analyzing the types of users and the actions that have been formally specified. The core functions of the Interface are five: search, analyze, scrape, annotate, and submit, whereas two types of users have been defined: the visitor and the contributor. A visitor can search and analyze hate speech data, while the contributor can also scrape, annotate and submit relevant content. The functions that are available to all users can be considered as public, while the rest as private. The private functionality is only accessible by registered users which are intended to be media professionals. Figure 2 demonstrates the succession of all functions in a single workflow, divided into the two aforementioned groups. As the current work focuses on the public functionality of the interface, only a brief description of the private functions is given.

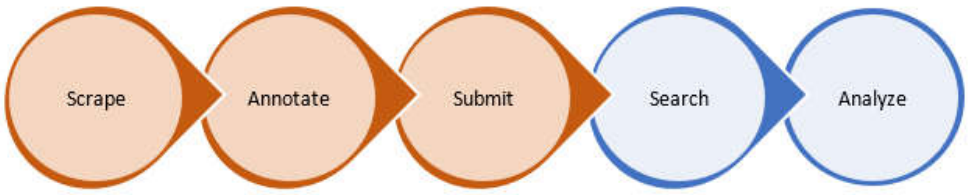

Figure 2. The private (orange) and public (blue) functionality of the PHARM Interface.

Search: One of the main functionalities of the interface is the navigation through the hate speech records contained in the database. The user can view records for any supported language (English, Greek, Italian, Spanish) or, additionally, filter the results by applying a variety of filters. The available filters are:

- Source selection (tweets, Facebook posts and replies, website articles and comments).
- Date and time selection (show results inside a specific period).
- Annotation filtering (hate/no hate, positive/neutral/negative sentiment).

- Keyword filtering (a search query for finding occurrences to texts).

The user can preview the records as a list, download them as a CSV or a JSON file, or display detailed information for each item. The search results can be viewed and downloaded either in the "simple" or the "scientific" form, disabling or enabling the presence of metadata, respectively. Figure 3 presents the search results screen of the interface.

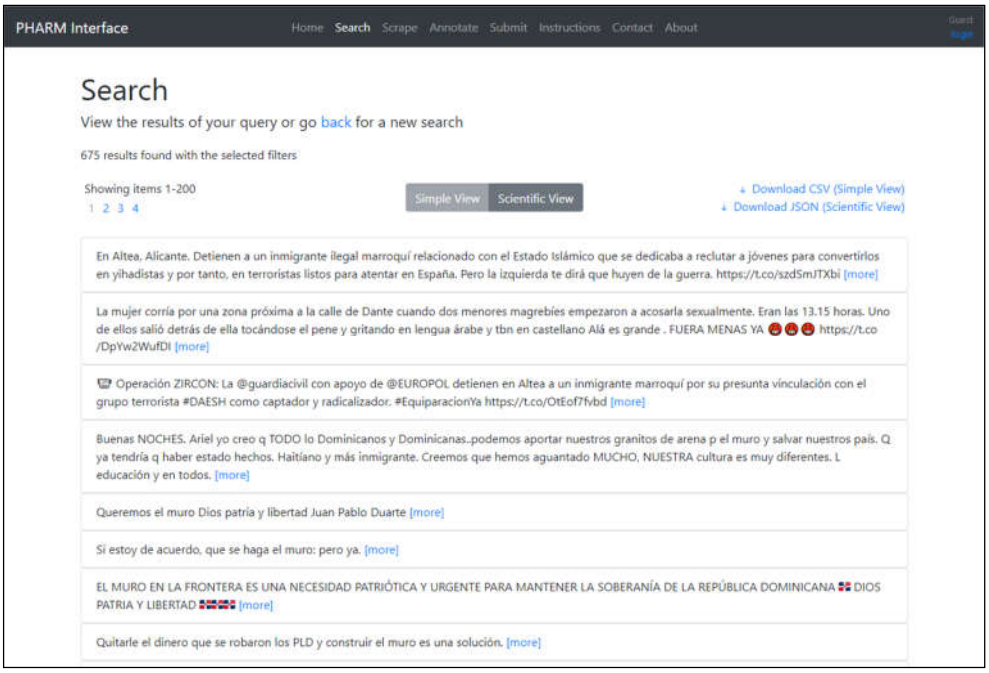

Figure 3. The search results screen of the PHARM Interface.

Analyze: When a record is selected, or a text is placed on the home screen, a detailed report appears. The location is marked on a map and the results of various text analysis algorithms are presented with graphics (icons, bars, etc.). The results concern hate speech detection and sentiment analysis (for both unsupervised and supervised classification methods), frequent entity detection, and geolocation estimation. Figure 4 depicts the analysis screen of the PHARM interface.

Scrape: The PHARM interface enables the mass-collection of text data from two popular platforms: Twitter and YouTube. A user can collect hate speech data from Twitter, by selecting language (Greek, English, Italian, or Spanish) and invoking the process. The tweets are collected based on language-specific lexicons that have been developed in the context of the project. The process stops after a user-configurable time interval and a link is provided for downloading a JSON file that contains the data. These data may be further used for annotation, submission to the PHARM database, or any other NLP task. In the case of YouTube, instead of selecting the language, a search query should be set. The search query can include individual search terms or a combination of them, separated by a comma. The resulting data can be downloaded as a CSV or JSON file.

Annotate: The annotation process is powered by the Doccano tool [61]. Doccano is an annotation management system for text data and can be used for developing datasets for facilitating classification, entity tagging, or translation tasks. In the context of the PHARM project, it is used for text classification and each record should be labeled with specific tags denoting hate speech and sentiment load.

Submit: Data entry can be executed either one by one or massively. Concerning the first method, the user should set all data (text) and metadata (source, language, date, hate,

sentiment, etc.) via the corresponding input forms (i.e., text fields, radio buttons, etc.). If data are already formed appropriately, they can be imported as a JSON file too.

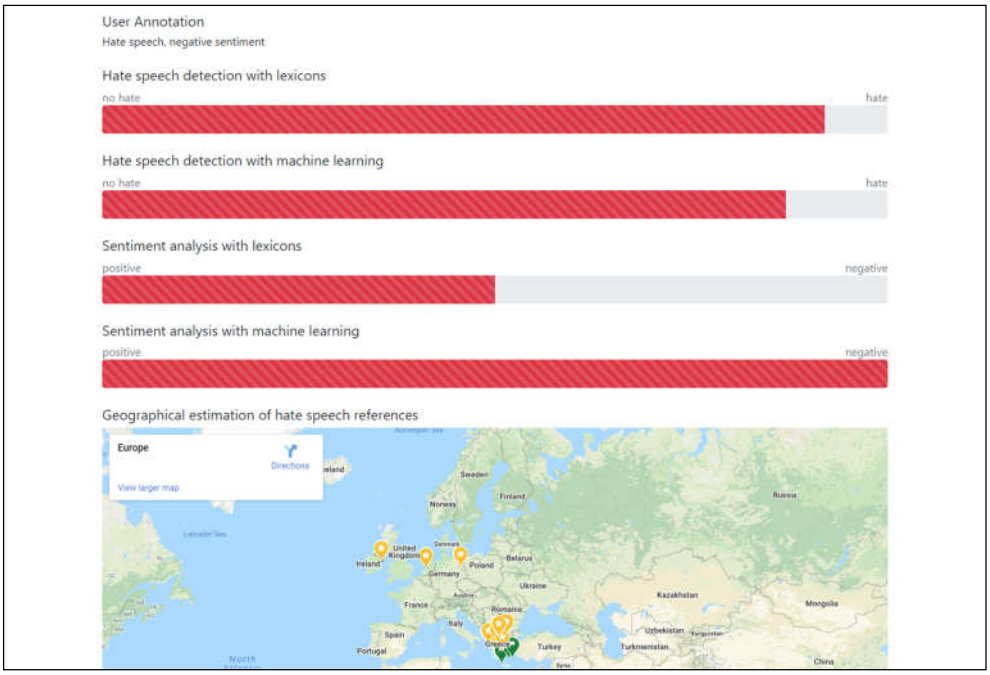

Figure 4. The analysis screen of the PHARM Interface.

3.2. Analysis and Usability Evaluation Results

The web interface was developed during the last year, while critical alpha evaluation tests were conducted inside the research team. Furthermore, the implemented interface was subjected to a beta assessment process by experts in the scientific fields of web design, web graphics, etc., aiming at the detection and correction of problematic aspects/ flaws at an early stage. Consequently, the final step of the evaluation was the conducted broadened empirical survey via the formulated questionnaire of Section 2.4, while in this section the extracted results are presented. Specifically, Figure 5 exhibits the responses regarding the suitability of the web interface towards the hate speech and sentiment loads detection mechanisms in text data, gathering 75% of agree/strongly agree evaluation score.

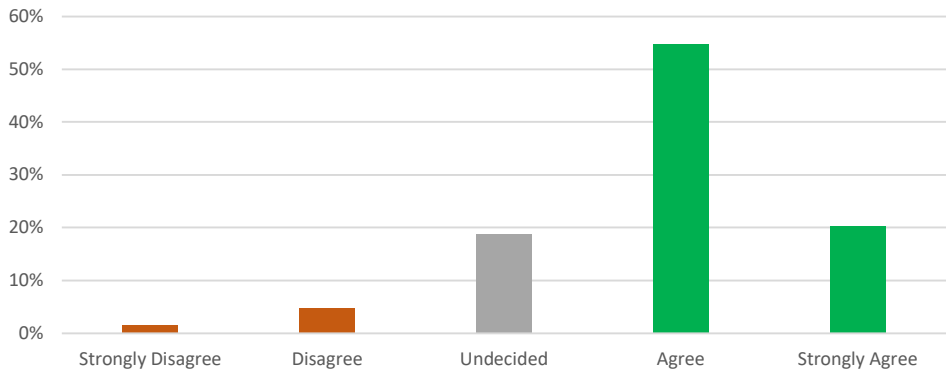

Figure 5. Responses to the suitability of the PHARM interface.

As Table 5 presents, the eight evaluation factors are constituted by various numbers of items, therefore for homogeneity and statistical purposes, the average scores of the

Likert-scaled items were computed for each metric. Table 5 exhibits the mean and standard deviation values of the assessment factors based on the responses of the $n = 64$ participants. The results showed that in all aspects the web interface grades are prone to four with a lower to one standard deviation, therefore indicating a well-designed web interface, with increased usability, presenting comprehensive content, navigation, and interaction mechanisms, and being easy-to-learn. Of course, these general results are further processed towards the determination of inner class correlations and group scores differentiations to address the defined research questions of the survey.

Table 5. Statistical values of evaluation factors.

#	Factor	Mean	Standard Deviation
1	Efficiency	3.72	0.73
2	Usability	3.95	0.64
3	Learnability	3.97	0.68
4	Satisfaction	3.71	0.73
5	Navigation	3.93	0.82
6	Content	3.74	0.55
7	Interactivity	3.85	0.67
8	Design	3.66	0.88

One of the main concerns for the effective engagement of users into a web interface is their knowledge background, possible previous experience in similar services, etc. Therefore, while addressing RQ1, the evaluation metrics were examined in relation with computer and Internet familiarity of the implicated users, towards the extraction of meaningful results. Taking into consideration that the factors of computer familiarity and Internet familiarity are answered in a 5-level Likert scale, the subjective judgment of knowledge background is unavoidably inserted in the statistical analysis. Because of possible error propagation due to the moderate number of participants and also the preliminary nature of the conducted survey, the responses in these two factors are grouped into two major categories. Specifically, the recoded categories are poor familiarity (including Answers 1 and 2) and good familiarity (including Answers 4 and 5), leaving out the moderate/ ambiguous level of computer and Internet familiarity (Level 3 in Likert scale), therefore functioning in a binary mode. Taking into consideration the continuous-form variables of the evaluation factors and the nominal two-scaled grouping of computer and Internet familiarity items, the statistical analysis proceeded into independent samples t-test methods, in order to compute the average scores differentiations of the assessment values into the formulated groups of participants.

Figure 6 graphically exhibits the crosstabulation matrices of average evaluation scores of the groups, while Table 6 presents the calculated values of the conducted t-tests, with significance level $a = 0.05$. In this context, statistically significant differentiations in average scores between groups were computed for usability, learnability, navigation in the computer familiarity groups, while only for the first two ones in the Internet familiarity groups.

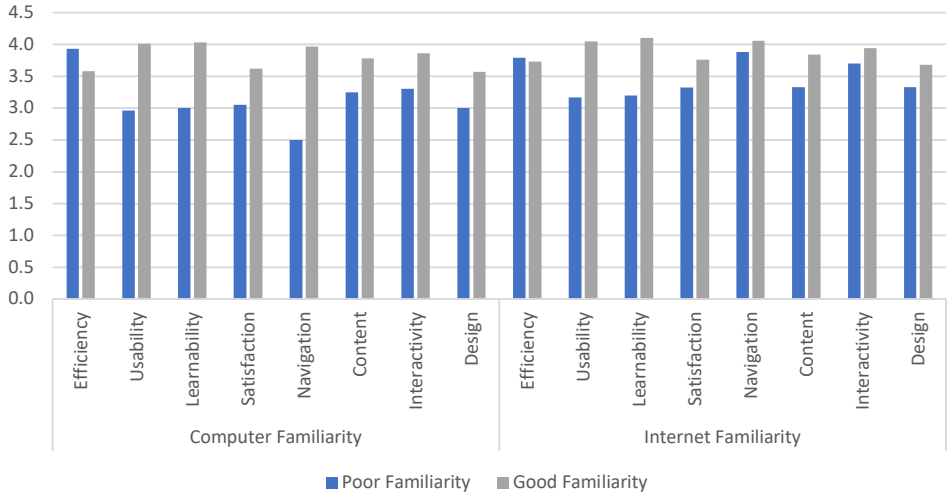

Figure 6. Average evaluation scores of the two groups for computer and internet familiarity variables.

Table 6. T-tests results for correlation between the evaluation factors and the groups of computer and Internet familiarity variables.

#	Factor	Computer Familiarity		Internet Familiarity	
		t-Value	*p*-Value	t-Value	*p*-Value
1	Efficiency	0.598	0.554	0.105	0.917
2	Usability	−2.514	0.018 *	−2.067	0.045 *
3	Learnability	−2.283	0.030 *	−2.217	0.032 *
4	Satisfaction	−1.015	0.318	−0.821	0.416
5	Navigation	−2.265	0.031 *	−0.349	0.728
6	Content	−1.267	0.215	−1.347	0.185
7	Interactivity	−1.231	0.228	−0.525	0.602
8	Design	−0.764	0.451	−0.517	0.608

* Statistically significant difference between groups at a = 0.05 significance level

Because of the specific contribution of the PHARM Project in hate speech detection in textual data, affecting public news and awareness, the second research question (RQ2) refers to the assessment of the web interface from participants that work (or have worked) in journalism compared to simple users with other professions. For this reason, a dedicated independent samples *t*-test was conducted for the evaluation scores of these two groups (answering YES if they work/have worked as journalists and NO if not the case).

Figure 7 presents the average values of the eight evaluation factors for the two subsets of participants, while Table 7 exhibits the related *t*-test outputs, again in a = 0.05 significance level. As can be observed, there was statistical significant difference in the average scores of the two groups only for the efficiency evaluation factor.

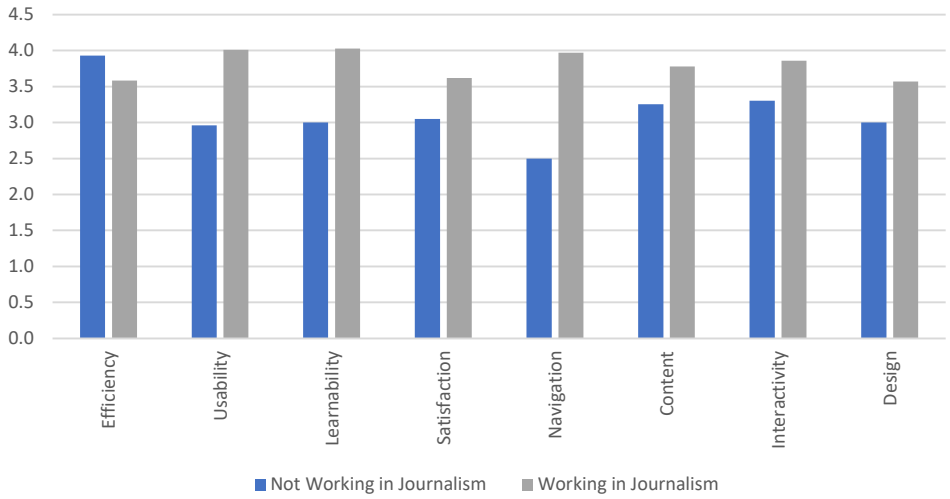

Figure 7. Average evaluation scores of groups in Journalism variable.

Table 7. Independent samples *t*-test of evaluation scores for the groups of working/ have worked as journalists or not.

#	Factor	t-Value	*p*-Value
1	Efficiency	2.230	0.029 *
2	Usability	−0.399	0.691
3	Learnability	0.096	0.924
4	Satisfaction	0.275	0.784
5	Navigation	−0.033	0.974
6	Content	0.425	0.672
7	Interactivity	0.750	0.456
8	Design	−0.278	0.782

* Statistically significant difference between groups at a = 0.05 significance level

4. Discussion

Overall, as Figure 6 presents, the PHARM interface was positively evaluated by the engaged participants (75%), while special attention was given to the remaining 25% concerning possible problems or negative aspects and functionalities of the platform, for subsequent developed versions. While referring to RQ1, the second column of Table 6 indicates that there is statistical significant difference between the mean scores of the groups of computer familiarity with respect to usability ($p = 0.018$), learnability ($p = 0.030$), and navigation ($p = 0.031$) evaluation factors of the interface. This fact is also validated in Figure 6, since the average values for the computer poor familiarity group are substantial lower versus the good familiarity one, in the factor of usability (2.96 vs. 4.01), learnability (3.00 vs. 4.03) and navigation (2.50 vs. 3.97). These results imply that amateurs in computer science participants confronted potential difficulties while navigating or learning how to utilize the web interface. In this context, some modifications are necessary for further optimizing the end-user interface to become more comprehensive, with increased usability. Nevertheless, for the rest of five evaluation factors, there was no statistically significant difference between the groups of computer familiarity, which is in accordance with the similar average values in these cases of Figure 6. Consequently, the conducted tests indicated towards the users' satisfaction and web service efficiency, along with the inclusion of adequate content, design, and interactivity mechanisms.

Almost the same results were retrieved for the two groups of Internet familiarity (RQ2), since there was statistically significant difference only for usability ($p = 0.045$) and

learnability (p = 0.032) metrics (without any difference in navigation). The average evaluation scores of Internet poor familiarity group were substantially lower compared to the good familiarity one for usability (3.17 vs. 4.05) and learnability (3.20 vs. 4.09), while there were no crucial differentiations for the remaining six evaluation factors. Taking into consideration the aforementioned results, the web interface was, in general, positively evaluated by all participants for most of its aspects, while specific actions are required in further optimizations/evolvement of the platform, to address the low usability and learnability scores for the less technologically experienced users. For instance, short "how-to" videos and simplified versions of manuals are already discussed among the research team members, to address the usability, navigation, and learnability deficiencies for potential amateur users.

With regard to RQ2, the extracted p values of Table 7 indicate that there is a statistically significant difference of evaluation scores between the two groups only for the factor of the efficiency (p = 0.029) of the web platform, while the exact average values are 3.51 for the subset who work/ have worked in journalism compared to 3.91 for those who have no relationship with it. This fact implies that the first group remains somehow skeptical about the effectiveness of the web service towards the detection of hate speech and emotional load in text, which mainly relies on human-centric approaches due to the implicated subjectivity. Therefore, the integrated natural language processing modules will be further evolved to achieve maximum precision, persuading for applicability of the innovative automations without human intervention. However, it has to be highlighted that in all other assessment metrics there was no substantial difference in the average evaluation scores, validating the high-quality content, design, navigation mechanisms, etc., either for professionals or simple users (with scores usually close to four for both groups).

The web service that has been presented in the current paper is part of a software framework for the collection and analysis of texts from several social media and websites, containing hate speech against refugees. The web service covers the functionality of web scraping, annotating, submitting annotated content and querying the database. It supports multi-language and multi-source content collection and analysis. This allows the formulation of a comprehensive database that can lead to the development of generalized hate speech and sentiment polarity modeling. This is expected to contribute significantly in the enhancement of semantic aware augmentation of unstructured web content. Future plans include an in-depth evaluation of state-of-the-art technologies in the big data volumes that are collected and annotated constantly through the PHARM software. Providing the functionality and the database online, makes it accessible to the public and allows more people to get involved. The results of the formative evaluation that are presented validate the appeal of the project to the target audience, and provide important feedback for the improvement of future versions. Subsequent larger scale and more generalized evaluation phases will follow, according to the adopted human-centered LUCID design.

Author Contributions: Conceptualization, C.A.-C., A.V., and C.D.; methodology, L.V., N.V., and R.K.; software, L.V., and N.V.; formal analysis, R.K.; investigation, T.S., M.M., and R.K.; resources, L.V., N.V., and R.K.; data curation, T.S., and M.M.; writing—original draft preparation, L.V., N.V., R.K., T.S., M.M., C.D., and A.V.; writing—review and editing, L.V., N.V., R.K., T.S., M.M., C.D., and A.V.; visualization, L.V., N.V., and R.K.; supervision, A.V., C.D. and C.A.-C.; project administration, A.V., C.A.-C., and C.D.; funding acquisition, C.A.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union's Right Equality and Citizenship Programme (2014–2020). REC-RRAC-RACI-AG-2019 Grant Agreement 875217.

Data Availability Statement: All data that are not subjected to institutional restrictions are available through the links provided within the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Matsiola, M.; Dimoulas, C.A.; Kalliris, G.; Veglis, A.A. Augmenting User Interaction Experience Through Embedded Multimodal Media Agents in Social Networks. In *Information Retrieval and Management*; IGI Global: Hershey, PA, USA, 2018; pp. 1972–1993.
2. Siapera, E.; Veglis, A. *The Handbook of Global Online Journalism*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
3. Katsaounidou, A.; Dimoulas, C.; Veglis, A. *Cross-Media Authentication and Verification: Emerging Research and Opportunities*; IGI Global: Hershey, PA, USA, 2018.
4. Dimoulas, C.; Veglis, A.; Kalliris, G. Application of mobile cloud based technologies in news reporting: Current trends and future perspectives. In *Joel Rodrigues*; Lin, K., Lloret, J., Eds.; Mobile Networks and Cloud Computing Convergence for Progressive Services and Applications; Chapter 17; IGI Global: Hershey, PA, USA, 2014; pp. 320–343.
5. Dimoulas, C.A.; Symeonidis, A.L. Syncing Shared Multimedia through Audiovisual Bimodal Segmentation. *IEEE MultiMedia* **2015**, *22*, 26–42.
6. Sidiropoulos, E.; Vryzas, N.; Vrysis, L.; Avraam, E.; Dimoulas, C. Growing Media Skills and Know-How in Situ: Technology-Enhanced Practices and Collaborative Support in Mobile News-Reporting. *Educ. Sci.* **2019**, *9*, 173, doi:10.3390/educsci9030173.
7. Dimoulas, C.A.; Veglis, A.A.; Kalliris, G.; Khosrow-Pour, D.M. Semantically Enhanced Authoring of Shared Media. In *Encyclopedia of Information Science and Technology, Fourth Edition*; IGI Global: Hershey, PA, USA, 2018; pp. 6476–6487.
8. Saridou, T.; Veglis, A.; Tsipas, N.; Panagiotidis, K. Towards a semantic-oriented model of participatory journalism management. Available online: https://coming.gr/wp-content/uploads/2020/02/2_2019_JEICOM_SPissue_Saridou_pp.-27-37.pdf (accessed on 18 March 2021).
9. Cammaerts, B. Radical pluralism and free speech in online public spaces. *Int. J. Cult. Stud.* **2009**, *12*, 555–575, doi:10.1177/1367877909342479.
10. Fortuna, P.; Nunes, S. A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.* **2018**, *51*, 1–30, doi:10.1145/3232676.
11. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI Conference on Web and Social Media, Palo Alto, CA, USA, 25–28 June 2017.
12. Ekman, M. Anti-immigration and racist discourse in social media. *Eur. J. Commun.* **2019**, *34*, 606–618, doi:10.1177/0267323119886151.
13. Burnap, P.; Williams, M.L. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making. In Proceedings of the 2014 Internet, Policy & Politics Conferences, Oxford, UK, 15–26 September 2014.
14. Pohjonen, M.; Udupa, S. Extreme speech online: An anthropological critique of hate speech debates. *Int. J. Commun.* **2017**, *11*, 1173–1191.
15. Ben-David, A.; Fernández, A.M. Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *Int. J. Commun.* **2016**, *10*, 1167–1193.
16. Olteanu, A.; Castillo, C.; Boy, J.; Varshney, K. The effect of extremist violence on hateful speech online. In Proceedings of the twelfth International AAAI Conference on Web and Social Media, Stanford, CA, USA, 25–28 June 2018.
17. Paz, M.A.; Montero-Díaz, J.; Moreno-Delgado, A. Hate Speech: A Systematized Review. *SAGE Open* **2020**, *10*, doi:10.1177/2158244020973022.
18. Calvert, C. Hate Speech and Its Harms: A Communication Theory Perspective. *J. Commun.* **1997**, *47*, 4–19, doi:10.1111/j.1460-2466.1997.tb02690.x.
19. Boeckmann, R.J.; Turpin-Petrosino, C. Understanding the harm of hate crime. *J. Soc. Issues* **2002**, *58*, 207–225.
20. Anderson, P. *What Is Web 2.0? Ideas, Technologies and Implications for Education*; JISC: Bristol, UK, 2007.
21. Kim, Y.; Lowrey, W. Who are Citizen Journalists in the Social Media Environment? *Digit. J.* **2014**, *3*, 298–314, doi:10.1080/21670811.2014.930245.
22. Quandt, T. Dark Participation. *Media Commun.* **2018**, *6*, 36–48, doi:10.17645/mac.v6i4.1519.
23. Schmidt, A.; Wiegand, M. A Survey on Hate Speech Detection using Natural Language Processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Valencia, Spain, 3 April 2017; pp. 1–10.
24. Bruns, A. The active audience: Transforming journalism from gatekeeping to gatewatching. In *Making Online News: The Ethnography of New Media Production*; Paterson, C., Domingo, D., Eds.; Peter Lang: New York, NY, USA, 2008; pp. 171–184.
25. Gillmor, D. *We the media. Grassroots Journalism by the People, for the People*; O'Reilly: Sebastopol, CA, USA, 2004.
26. Hanitzsch, T.; Quandt, T. Online journalism in Germany. In *The Handbook of Global Online Journalism*; Siapera, E., Veglis, A., Eds.; Wiley-Blackwell: West Sussex, UK, 2012; pp. 429–444.
27. Singer, J.B.; Hermida, A.; Domingo, D.; Heinonen, A.; Paulussen, S.; Quandt, T.; Reich, Z.; Vujnovic, M. *Participatory Journalism. Guarding Open Gates at Online Newspapers*; Wiley-Blackwell: Malden, MA, USA, 2018; doi:10.1002/9781444340747.
28. Obermaier, M.; Hofbauer, M.; Reinemann, C. Journalists as targets of hate speech. How German journalists perceive the consequences for themselves and how they cope with it. *Stud. Commun. Media* **2018**, *7*, 499–524, doi:10.5771/2192-4007-2018-4-499.
29. Boberg, S.; Schatto-Eckrodt, T.; Frischlich, L.; Quandt, T. The Moral Gatekeeper? Moderation and Deletion of User-Generated Content in a Leading News Forum. *Media Commun.* **2018**, *6*, 58–69, doi:10.17645/mac.v6i4.1493.
30. Wolfgang, J.D. Pursuing the Ideal. *Digit. J.* **2015**, *4*, 764–783, doi:10.1080/21670811.2015.1090882.

31. Wintterlin, F.; Schatto-Eckrodt, T.; Frischlich, L.; Boberg, S.; Quandt, T. How to Cope with Dark Participation: Moderation Practices in German Newsrooms. *Digit. J.* **2020**, *8*, 904–924, doi:10.1080/21670811.2020.1797519.
32. Masullo, G.M.; Riedl, M.J.; Huang, Q.E. Engagement Moderation: What Journalists Should Say to Improve Online Discussions. *Journal. Pract.* **2020**, 1–17, doi:10.1080/17512786.2020.1808858.
33. Hille, S.; Bakker, P. Engaging the social news user: Comments on news sites and Facebook. *Journal. Pract.* **2014**, *8*, 563–572.
34. Wang, S. Moderating Uncivil User Comments by Humans or Machines? The Effects of Moderation Agent on Perceptions of Bias and Credibility in News Content. *Digit. J.* **2021**, *9*, 64–83, doi:10.1080/21670811.2020.1851279.
35. Risch, J.; Krestel, R. Delete or not delete? Semi-automatic comment moderation for the newsroom. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, Santa Fe, NM, USA, 25 August 2018; pp. 166–176.
36. MacAvaney, S.; Yao, H.-R.; Yang, E.; Russell, K.; Goharian, N.; Frieder, O. Hate speech detection: Challenges and solutions. *PLoS ONE* **2019**, *14*, e0221152, doi:10.1371/journal.pone.0221152.
37. Ayo, F.E.; Folorunso, O.; Ibharalu, F.T.; Osinuga, I.A. Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Comput. Sci. Rev.* **2020**, *38*, 100311, doi:10.1016/j.cosrev.2020.100311.
38. Gitari, N.D.; Zhang, Z.; Damien, H.; Long, J. A Lexicon-based Approach for Hate Speech Detection. *Int. J. Multimed. Ubiquitous Eng.* **2015**, *10*, 215–230, doi:10.14257/ijmue.2015.10.4.21.
39. Arcila-Calderón, C.; de la Vega, G.; Herrero, D.B. Topic Modeling and Characterization of Hate Speech against Immigrants on Twitter around the Emergence of a Far-Right Party in Spain. *Soc. Sci.* **2020**, *9*, 188.
40. Arcila-Calderón, C.; Herrero, D.B.; Frías, M.; Seoanes, F. Refugees Welcome? Online Hate Speech and Sentiments in Twitter 2 in Spain during the reception of the boat Aquarius. *Sustainability* **2021**, *13*, 2728.
41. Arcila-Calderón, C.; Blanco-Herrero, D.; Apolo, M.B.V. Rechazo y discurso de odio en Twitter: Análisis de contenido de los tuits sobre migrantes y refugiados en español/Rejection and Hate Speech in Twitter: Content Analysis of Tweets about Migrants and Refugees in Spanish. *Rev. Española Investig. Sociol.* **2020**, *172*, 21–40, doi:10.5477/cis/reis.172.21.
42. Badjatiya, P.; Gupta, S.; Gupta, M.; Varma, V. Deep Learning for Hate Speech Detection in Tweets. In Proceedings of the 26th International Conference on Compiler Construction, Austin, TX, USA, 5–6 February 2017; pp. 759–760.
43. Pitsilis, G.K.; Ramampiaro, H.; Langseth, H. Effective hate-speech detection in Twitter data using recurrent neural networks. *Appl. Intell.* **2018**, *48*, 4730–4742, doi:10.1007/s10489-018-1242-y.
44. Ghani, N.A.; Hamid, S.; Hashem, I.A.T.; Ahmed, E. Social media big data analytics: A survey. *Comput. Hum. Behav.* **2019**, *101*, 417–428, doi:10.1016/j.chb.2018.08.039.
45. Sánchez-Holgado, P.; Arcila-Calderón, C. Supervised Sentiment Analysis of Science Topics: Developing a Training Set of Tweets in Spanish. *J. Infor. Technol. Res.* **2020**, *13*, 80–94.
46. Korkmaz, G.; Cadena, J.; Kuhlman, C.J.; Marathe, A.; Vullikanti, A.; Ramakrishnan, N. Multi-source models for civil unrest forecasting. *Soc. Netw. Anal. Min.* **2016**, *6*, 1–25, doi:10.1007/s13278-016-0355-8.
47. Capozzi, A.T.; Lai, M.; Basile, V.; Poletto, F.; Sanguinetti, M.; Bosco, C.; Patti, V.; Ruffo, G.; Musto, C.; Polignano, M.; et al. Computational linguistics against hate: Hate speech detection and visualization on social media in the "Contro L'Odio" project. In Proceedings of the 6th Italian Conference on Computational Linguistics, CLiC-it 2019, Bari, Italy, 13–15 November 2019; Volume 2481, pp. 1–6.
48. Dimoulas, C.A. Multimedia. In *The SAGE International Encyclopedia of Mass Media and Society*; Merskin, D.L. Ed.; SAGE Publications, Inc.: Saunders Oaks, CA, USA, 2019.
49. Dimoulas, C.A. *Multimedia Authoring and Management Technologies: Non-Linear Storytelling in the New Digital Media*; Association of Greek Academic Libraries: Athens, Greece, 2015. Available online: http://hdl.handle.net/11419/4343 (accessed on 18 March 2021). (In Greek)
50. Chatzara, E.; Kotsakis, R.; Tsipas, N.; Vrysis, L.; Dimoulas, C. Machine-Assisted Learning in Highly-Interdisciplinary Media Fields: A Multimedia Guide on Modern Art. *Educ. Sci.* **2019**, *9*, 198, doi:10.3390/educsci9030198.
51. Psomadaki, O.; Dimoulas, C.; Kalliris, G.; Paschalidis, G. Digital storytelling and audience engagement in cultural heritage management: A collaborative model based on the Digital City of Thessaloniki. *J. Cult. Herit.* **2019**, *36*, 12–22.
52. Katsaounidou, A.; Vrysis, L.; Kotsakis, R.; Dimoulas, C.; Veglis, A. MAthE the Game: A Serious Game for Education and Training in News Verification. *Educ. Sci.* **2019**, *9*, 155, doi:10.3390/educsci9020155.
53. Graham, M.; Hale, S.A.; Gaffney, D. Where in the World Are You? Geolocation and Language Identification in Twitter. *Prof. Geogr.* **2014**, *66*, 568–578, doi:10.1080/00330124.2014.907699.
54. De Vries, E.; Schoonvelde, M.; Schumacher, G. No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications. *Politi. Anal.* **2018**, *26*, 417–430, doi:10.1017/pan.2018.26.
55. Loria, S. Textblob Documentation. Available online: https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf (accessed on 18 March 2021).
56. Clemens, K. Geocoding with openstreetmap data. In Proceedings of the GEOProcessing 2015, Lisbon, Portugal, 22–27 February 2015; p. 10.
57. Arcila-Calderón, C.; Amores, J.; Blanco, D.; Sánchez, M.; Frías, M. Detecting hate speech against migrants and refugees in Twitter using supervised text classification. In Proceedings of the International Communication Association's 71th Annual Conference, Denver, CO, USA, 27–31 May 2021.
58. Chollet, F. Keras: The Python Deep Learning Library. Available online: http://ascl.net/1806.022 (accessed on 18 March 2021).

59. Spiliotopoulou, L.; Damopoulos, D.; Charalabidis, Y.; Maragoudakis, M.; Gritzalis, S. Europe in the shadow of financial crisis: Policy Making via Stance Classification. In Proceedings of the 50th Hawaii International Conference on System Sciences (2017), Hilton Waikoloa Village, HI, USA, 4–7 January 2017.
60. Thelwall, M.; Buckley, K.; Paltoglou, G.; Cai, D.; Kappas, A. Sentiment strength detection in short informal text. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 2544–2558, doi:10.1002/asi.21416.
61. Nakayama, H.; Kubo, T.; Kamura, J.; Taniguchi, Y.; Liang, X. Doccano: Text Annotation tool for Human. 2018. Available online: https://github. com/doccano/doccano (accessed on 18 March 2021).