Heliyon 9 (2023) e16084

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A web framework for information aggregation and management of multilingual hate speech

Rigas Kotsakis [a], Lazaros Vrysis [b], Nikolaos Vryzas [b], Theodora Saridou [b], Maria Matsiola [c], Andreas Veglis [b], Charalampos Dimoulas [b,*]

[a] *International Hellenic University, Greece*
[b] *Aristotle University of Thessaloniki, Greece*
[c] *University of Western Macedonia, Greece*

ARTICLE INFO

Keywords:
Hate speech
Sentiment analysis
Information retrieval
Content management and moderation
Social media
Multilingual
Natural language processing
Machine learning

ABSTRACT

Social media platforms have led to the creation of a vast amount of information produced by users and published publicly, facilitating participation in the public sphere, but also giving the opportunity for certain users to publish hateful content. This content mainly involves offensive/discriminative speech towards social groups or individuals (based on racial, religious, gender or other characteristics) and could possibly lead into subsequent hate actions/crimes due to persistent escalation. Content management and moderation in big data volumes can no longer be supported manually. In the current research, a web framework is presented and evaluated for the collection, analysis, and aggregation of multilingual textual content from various online sources. The framework is designed to address the needs of human users, journalists, academics, and the public to collect and analyze content from social media and the web in Spanish, Italian, Greek, and English, without prior training or a background in Computer Science. The backend functionality provides content collection and monitoring, semantic analysis including hate speech detection and sentiment analysis using machine learning models and rule-based algorithms, storing, querying, and retrieving such content along with the relevant metadata in a database. This functionality is assessed through a graphic user interface that is accessed using a web browser. An evaluation procedure was held through online questionnaires, including journalists and students, proving the feasibility of the use of the proposed framework by non-experts for the defined use-case scenarios.

1. Introduction

News aggregation and automated extraction of semantic information from unstructured text has become an essential task for adding value to existing massive data volume. Opinion mining and sentiment analysis (OMSA) is a natural language processing task that uses an algorithmic formulation to identify opinionated content and categorize it as having “positive”, “negative”, or “neutral” polarity [46,47]. Stance detection is another important task, with a blurry difference from sentiment analysis, since it focuses on the stance concerning an object of evaluation; “favour”, “against”, “none” [1]. Popular approaches include different classification techniques (e.g., machine learning and lexicon-based approaches) and levels of analysis (e.g., document, sentence, or aspect-level). It is

[*] Corresponding author.
E-mail address: babis@eng.auth.gr (C. Dimoulas).

https://doi.org/10.1016/j.heliyon.2023.e16084
Received 28 November 2022; Received in revised form 21 April 2023; Accepted 4 May 2023
Available online 9 May 2023

observed that reviews constitute the most frequently worked on dataset for OMSA research followed by tweets and news articles [68].

Hate speech is defined as “any speech, which attacks an individual or a group with an intention to hurt or disrespect based on identity” (gender, race, religion, disabilities, etc.) [13]. In the case of hate speech detection, automated techniques can be applied to moderate content, but also for socio-political understanding and content analysis. A strong connection is considered to exist between hate speech and actual hate crime [38]. Thus, it has a solid impact on the lives of vulnerable communities that are victimized. Two dangerous types of hate speech are identified; those directed at the victims, and those directed at like-minded individuals [10]. In direct hate speech, the victims are injured immediately by the hate content, while in indirect hate speech harm may be immediate or delayed, and perpetrated by the agents, rather than the original actor [13]. In Ref. [64] the results of an online experiment show that participants who are exposed to hate, or civil negativity in user comments concerning the refugee crisis, are significantly affected, and are less likely to develop pro-social behavior. Such conclusions support the need for moderation. The real-time moderation of huge volumes of content that are published constantly indicates the need for automated workflows. The PHARM project introduces techniques and an interface for the automated detection of hate speech and it can also support users in checking their own texts before publishing. Moderation does not necessarily result in rejecting the publication of hateful content. In the US, hateful content is covered by freedom of speech law, while in the EU its removal is required [25]. Some social media companies automatically remove messages using a hate speech reference database [20]. This requires high-quality databases, manually annotated by experts. Human moderation is expensive and, in addition, often implies a significant negative psychological impact on the moderators [52].

2. Related work

2.1. User generated content and hate speech

With the advent of Web 2.0, users can contribute to a vast number of online frameworks through several forms of user-generated content (UGC), creating enormous data volumes with positive and negative contributions [17,60]. The information produced is difficult to be monitored and handled even for media professionals, thus new models are developed [44], in order -among others-to increase user awareness against harmful online communication [19]. In the same framework, research on people’s ability to distinguish real from fake images is performed in the context of “crowdsourced validation” [29]. However, in the field of journalism, the trustworthiness of reporting with incorporated UGC is disputable and presents negative or no overall effect [4,22]. Research conducted among social media users suggests that hate speech is expressed at higher levels in social media platforms [8,24], potentially leading to personal distress [49].

Exposure to hateful content online is more common among younger ages and is more frequently used to stereotype groups, centred on race or ethnicity [16,40], although with some variations among countries [24]. Furthermore, other factors, such as income and education, have been linked to inequalities in access, skills, and patterns of the Internet and social media use [18]. In a similar vein, lower social-grade individuals who use fewer sources of online news on average, are less likely to go directly to news organisations for news online and are consequently more reliant on the distributed discovery of news via social media and search engines [27]. Research results also suggest that digital media rather reinforce social inequalities in cultural consumption [65] and, while moving from a relatively low-choice offline media environment to a high-choice online media environment with more intense competition for attention, social inequalities in news consumption are likely to increase [27].

In order to prevent and counter the spread of hate speech online, the European Commission has agreed with Facebook, Microsoft, Twitter, and YouTube in a “Code of conduct on countering illegal hate speech online”, while other Internet giants, such as Instagram and TikTok, have also joined in it later. In parallel, a series of European Union (EU) Directives aim to control racist and xenophobic behaviors in the media (e.g. The Audiovisual Media Services Directive 64). Within mainstream media platforms, professionals try to counter problematic UGC using moderation methods [12], often engaging users in the process [59,69]. Moreover, some organisations utilize artificial intelligence techniques to tackle this massive work automatically [63] or implement semi-automatic approaches that integrate machine learning into the manual process [51], while considering the impact of the conversational context along with the combination of text content with social context in the process of automated hate speech detection [39,45].

Twitter is a major choice for social network studies research, thanks to its popularity, as well as its simple data model and Application Programming Interface (API) [2]. One of the advantages of exploring data from Twitter in social studies is that it is possible, besides content, to process user-related data, as well as the social graph, containing network-related data, like user interactions and connections. Sentiment analysis is one of the most promising techniques for textual content analysis in the context of social studies.

Several techniques exist for the support of such tasks. Lexical features that can be used in the computational analysis include word stems, Part Of Speech (POS), and n-grams. Named Entity Recognition is the extraction of the semantic identity of a word (person, place, etc.) and can also play an important role in the analysis. Emoticons are an element of a message that often carries affective information. Several language-dependent tools exist for the extraction of such tags [2]. In the work of [66] memes, a popular format on the modern web, are taken into consideration for the expression of hateful ideas. Another issue noted by Ref. [46] is that user-generated texts in the Web 2.0 tend to be noisy. The evaluation shows that for all text-mining approaches (unsupervised, supervised, semi-supervised), text pre-processing is a crucial part of the analysis pipeline.

2.2. Affective computing and sentiment analysis: tackling hate speech

As emotions play an integral role in communication, relative surveys are conducted via a variety of spectra and implementation techniques, such as the involved emotional aspects in terms of Quality of Experience (QoE) on the communication process through the influence of audiovisual content encoding decisions and properties [26,31], and the degree of agreement between the perceived emotion and the intended expressed emotion in a framework of applied Speech Emotion Recognition (SER) for theatrical performance and social media communication [62]. Affective computing and sentimental analysis is an interdisciplinary research area associated with advancements in machine learning relating computational interpretation and human emotions. On the field's presented challenges, adversarial training appears to be helpful in the development of generative and related discriminative models [23]. Furthermore, as convictions and feelings are spread among social media users, research on community detection based on users' behavioral similarity in personal neighborhood communities is performed while processing social interaction data [55,56].

A plethora of techniques (Machine/Deep Learning, Rule-based and Fuzzy Logic) can be found in the literature for hate speech detection, as well as hybrid methods that result in combinations of the above [7]. Likewise, a model employing a multichannel convolutional bidirectional gated recurrent unit was developed to identify toxicity in social networking sites [32]. Such techniques aim at the classification of texts regarding the existence of hateful or abusive language, depending on the classification scheme and the specific task. Similarly, regarding sentiment analysis, machine learning models are used either for classification schemes (positive--negative), or as regression models that calculate the degree of polarity in a given scale [67]. In the work of [6]; to overcome the limitations of a binary hate-speech classification scheme, an unsupervised clustering approach, combined with fuzzy logic rules is proposed to discover emotional patterns in tweets.

A major bottleneck in the field is set by the existence of a few benchmark datasets, limited in volume and with no uniform annotation schema [28]. follow a deep multi-task learning (MTL) framework in order to address several text classification tasks to improve performance for the individual task of hate speech detection. The formulation of large amounts of high-quality manually labelled data is often the most challenging task in the automated detection of abusive content. Creating a dataset with variations, may increase the likelihood of mining more informative instances [37]. This can be the work of experts or crowdsourcing [2]. Data augmentation refers to the actions of increasing the number of available data, using several techniques, depending on the data type. This can lead to better robustness and generalization of the trained deep learning models. In Ref. [9] an augmentation technique for text is presented, using a Back Translation and a Paraphrasing technique to produce similar versions of the available texts. Evaluation on different datasets proves increased performance. According to Ref. [3]; July), the experimental results of state-of-the-art models contradict with the performance of automated detection in actual applications. This is due to the evaluation on specific datasets. Better user distribution is important for better generalization in English models and, ultimately, robust multilingual models. Annotators can play also an important role in the integrity of datasets, and consequently in the effectiveness of models. Demographic characteristics and personal beliefs are important aspects of the subjective perception of offensive language. Some approaches consider these aspects for model training [30].

Multilingual text mining (MLTM) includes a set of techniques in order to address multilingual text input to extract semantic information, by applying some kind of automated process to discover the relationships between different languages. Also, although several models achieve good performance in benchmark datasets, cross-dataset model generalization remains a challenge. The feasibility of reusing models in different contexts, as well as the overfit of generated features in smaller corpora, can set the bottleneck of real-world applications [20]. This is what makes the creation of big, heterogeneous, multi-source datasets essential for the field. In the research of [42]; multilingual and cross-domain detection is investigated. Several models are trained for language-specific misogyny detection in social media. The performance of these models in datasets of general offensive language and multiple languages is evaluated, showing that deep learning models can generalize better in cross-domain and multilingual tasks [43]. evaluate several methodologies for transferring knowledge between corpora in different languages, in the direction of multi-lingual hate speech detection. Hate speech detection models, as well as most text recognition-related tasks, are much more advanced for the English language since more resources are available [48]. evaluate transfer learning of pre-trained models for hate speech detection in Spanish.

Social Network Analysis based on the social media messages' metadata can play an important role in abusive language detection [41]. investigated the behavior patterns of certain users and their effect on the propagation of toxic content in social networks in the topic-specific context of COVID-19 related abusive comments. They applied topic modelling based on Latent Dirichlet allocation (LDA) to discover topics related to the COVID-19 umbrella topic [11]. [36]. propose that social media posts should not be considered as standalone text. It is interactive, and posts are context-dependent, which means that knowledge concerning preceding posts can be crucial for recognizing aggression [7]. propose a JSON-based Metadata Representation Structure Model (MRSM) that includes metadata information concerning the temporal, spatial, and semantic attributes of every social media object.

The Perspective API created by Google developers can assist moderation by suggesting the detection of toxic content in a conversation, based on a ground truth created by experts and crowdsourcing [50]. Another interface based on web browser plugins to visualize aggressive content expressed either implicitly or explicitly is presented in the work of [36]. The plugins are developed for the two most popular media: Facebook and Twitter. Hatemeter is a European project that aims at the documentation, analysis, and prevention of hate speech on social media, connecting researchers and non-governmental organisations (NGOs) in Italy, France, and Great Britain. The delivered Hatemeter Platform allows automatic data gathering and real-time analysis. Social scientists may use the platform to explore large amounts of data, while NGOs can monitor speech trends and produce counter-narratives [33]. In CyberAid [58], a number of features (textual, user-related, and network-related) are combined in a mobile application that aims at cyberbullying detection, improving accuracy compared to existing approaches [14]. propose an architecture for the regulation of the circulation of hate speech in mobile environments. Their architecture consists of different layers that match the different stages of content creation, distribution, and consumption and include users' (sender/receiver) intermediaries, at national and international levels.

3. Research aims and project motivation

Based on the elaboration of the related work, there is a lack of user-friendly services that can address the full work cycle of analyzing hateful content online through a user-friendly graphic interface. This realization fuels the motivation to provide a framework that can be used by individuals with no technological background and training to collect content from Social Media and the web, access state-of-the-art models to analyze it in terms of sentiment polarity, hateful content, and context (e.g. geographical information), interpret the results, monitor hateful content publication real-time, access a well documented structured repository through highly configurable queries, annotate content and contribute to this repository. Therefore, a web interface that provides the aforementioned functionality, has been envisioned in the scope of Preventing Hate against Refugees and Migrants (PHARM) project [61]. The goal of the PHARM project is to monitor and model hate speech against refugees and migrants in Greece, Italy, and Spain. Ultimately, the project's anticipation is to predict and combat hate crime and also counter its effects using text-processing algorithms. Therefore, computing interfaces for Natural Language Processing (NLP) have been designed, along with the implementation of a graphical web tool for allowing a more intuitive user interaction. These have been named the "PHARM Scripts" and "PHARM Interface", respectively, forming the "PHARM Software".

This particular work concerns the design process of the PHARM Interface, aiming at validating the targeted usability and impact of the analyzed application. Hence, standard application development procedures are followed through the processes of rapid prototyping and anthropocentric design, i.e., the so-called logical-user-centred-interactive design (LUCID). Audience engagement is crucial, not only for communicating and listing the needs and preferences of the targeted users but also for serving crowdsourcing and data 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, including journalists, media professionals, communication specialists, subject-matter experts, 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 hate speech and emotions spread through the Internet and social media and that there is an audience willing to use the corresponding software application. 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 on the hypotheses are as follows.

[RQ1] What are the targeted usefulness and estimated usability of the application?
[RQ2] What are the use scenarios that the application addresses (personal use, contribution, personal publication, literacy)?

4. Materials and methods: design and evaluation framework

The methodology of the current research includes the documentation of the design and evaluation protocol of the required web interface.

4.1. The PHARM web interface: functionality, information processing, and management flow

The PHARM Interface should serve as the front-end of the PHARM Software. The core functions of the Interface are the following: monitor, search, analyze, scrape, annotate, and submit. Moreover, two types of users have been defined: the visitor and the contributor. A visitor can monitor, search for and analyze hate speech data, while the contributor can also scrape, annotate and submit relevant content. The functions that are available to all users are public, while the rest are private. The private functionality is only accessible by registered users which are intended to be media professionals (Fig. 1).

These functional requirements are summarized in Fig. 2. All presented functions concern the user group of the contributors, while

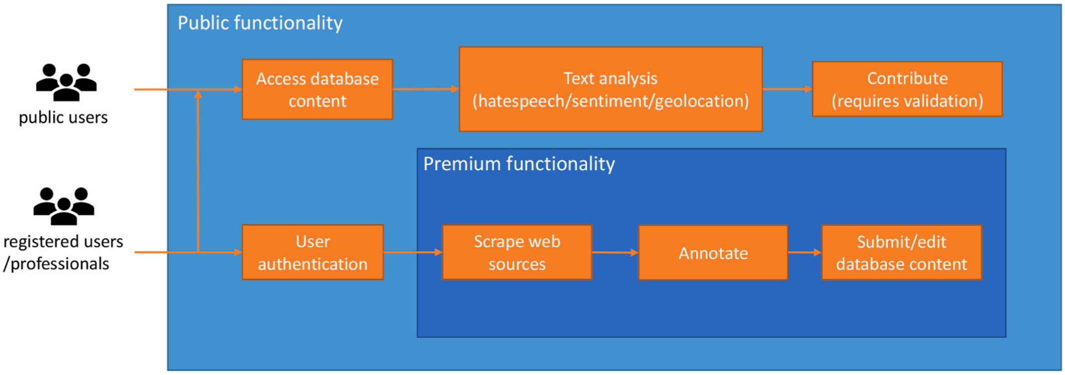

Public functionality
public users
registered users /professionals
Access database content
Text analysis (hatespeech/sentiment/geolocation)
Contribute (requires validation)
Premium functionality
User authentication
Scrape web sources
Annotate
Submit/edit database content

Fig. 1. The PHARM web interface roles and workflow.

the first three (highlighted in blue) are available for anyone who visits the website of the PHARM Interface.

In the scope of the current research, these three functions are under evaluation, and more detailed descriptions follow.

Monitor: The spatiotemporal visualization of the detected hate-speech content should be a distinctive feature of the user interface (UI). In order to enable such a visualization, an interface needs to be created and put into use.

Search: Navigation through the hate speech records present in the PHARM database must be one of the interface's primary features. Users should have the ability to search, receive a list of entries that match their criteria, and examine specific details for each record. The following filters must be accessible.

- Language (English, Greek, Italian, Spanish)
- Source selection (Facebook posts, tweets, YouTube comments, web articles, and comments).
- Date selection (show results inside a specific period).
- Annotation filtering (hate/no hate, positive/neutral/negative sentiment).
- Keyword filtering (a search query for finding occurrences in texts).

The user may also download the results as CSV or JSON files.

Analyze: A thorough report ought to show up if a record is chosen or if text is entered in the home screen. The results of the text analysis algorithms should be shown with graphics, and the estimated location of where the text has been composed may be highlighted on a map (icons, bars, etc.). The results concern sentiment analysis and hate speech detection (for both unsupervised and supervised classification methods). An early, low-fidelity mockup of the interface's analysis screen is shown in Fig. 3. When the Interface was first released, this screen served as its home page. Later, a different home screen was created and put into use.

The functionality and the architecture of the web interface were designed in the context of addressing several scenarios, and purposes of use of the application. The main use cases include checking one's texts for hate speech detection and analysis, texts from other users, contributing to the project's database, retrieving relative comments from the project's database, and conducting geolocation monitoring of articles. These use cases can be proved essential in addressing several scenarios like secondary and university education, multidisciplinary research, vocational training, and digital literacy, events in social media, and participatory infotainment.

Regarding the backend of the system, PHARM's data management and analysis features are referred to as PHARM Scripts. The source code of the PHARM Scripts, along with their documentation, is publicly available as a GitHub repository (https://github.com/thepharmproject/set_of_scripts). As for data management, it is worth mentioning that information is arranged in records of a single format with the text (i.e., content) as the primary field and the id, annotations, and meta fields as secondary fields. The meta field is a container for all metadata, such as source, language, date, location, hate, and sentiment load. Concerning data analysis, the most noteworthy methods include sentiment and hate speech analysis, language detection, time, and geolocation estimation. These methods have been implemented by exploiting a variety of programming libraries as well as coding custom scripts. As mentioned, the PHARM Interface primarily reads and analyses texts written in Greek, Italian, and Spanish. However, much of the scraped data may also include texts written in other languages or regional dialects. Therefore, a method for identifying the language of a text when it is not explicitly defined has been implemented. Python's textblob and googletrans libraries are two of the ones that are often utilized. Next, to estimate geolocation, named entities are taken out of texts and geocoded. Geopolitical entities (GPE), locations (LOC), faculties (FAC), and organisations (ORG), are among the named entities (e.g., countries, mountains, buildings, companies, etc.). This approach makes use of openstreetmap data and the Nominatim geocoder [15]. In addition to these, a technique for extracting date and time information from text has also been put into practice. To find datetime objects in texts, a variety of Python libraries are utilized, including dateparser, datefinder, and parsedatetime. As one of the project's main objectives is hate speech detection, two different methods have been implemented: a lexicon-based one and a machine-learning one. Both methods rely on the utilization of a language-specific model

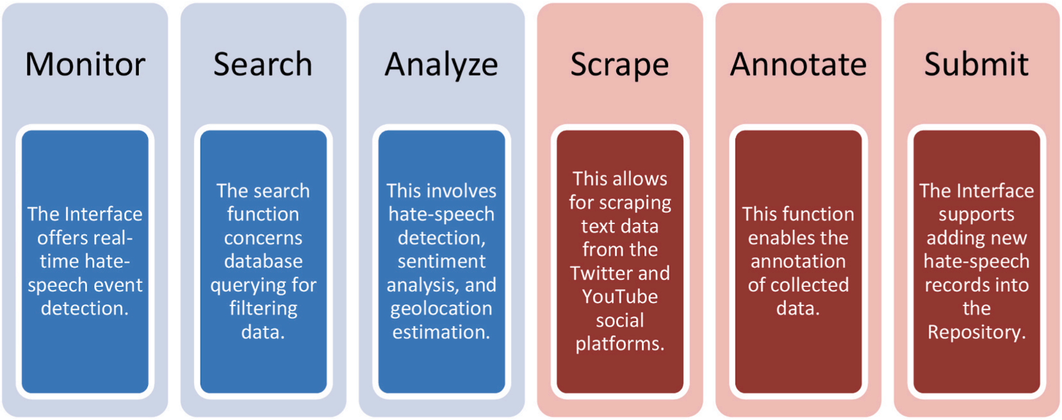
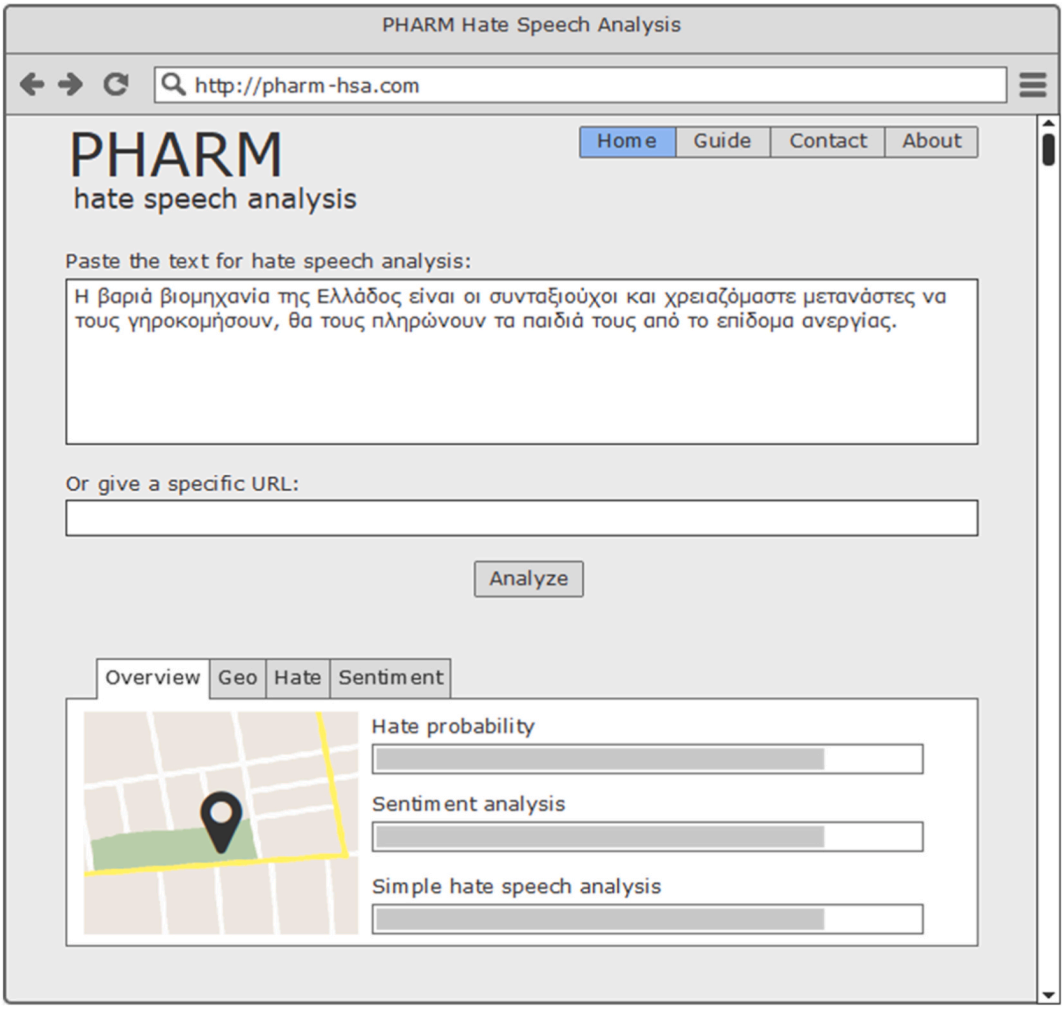

Monitor
The Interface offers real-time hate-speech event detection.
Search
The search function concerns database querying for filtering data.
Analyze
This involves hate-speech detection, sentiment analysis, and geolocation estimation.
Scrape
This allows for scraping text data from the Twitter and YouTube social platforms.
Annotate
This function enables the annotation of collected data.
Submit
The Interface supports adding new hate-speech records into the Repository.

Fig. 2. The main functional requirements for the PHARM Interface.



Η βαριά βιομηχανία της Ελλάδος είναι οι συνταξιούχοι και χρειαζόμαστε μετανάστες να τους γηροκομήσουν, θα τους πληρώνουν τα παιδιά τους από το επίδομα ανεργίας.

Fig. 3. A low-fidelity mockup of the analysis screen that served as the homepage of the Interface at the early stages of development.

and make use of common normalisation techniques for pre-processing (e.g., lower-casing, lemmatisation, and stop-word removal). In the first case, terms are searched in the text, while, in the second, hate speech detection is handled by a recurrent neural network (RNN) [5]. Similarly, the PHARM interface incorporates two methods for sentiment analysis [53]. Many essential components of the SentiStrength algorithm, including the recognition of booster, question, and negating words, are adopted by the lexicon-based method [57]. The supervised model for sentiment analysis, trained on a separate corpus, has the same architecture as the model for hate speech identification.

4.2. *Interactive design and assessment framework*

The questions are formed as part of the user experience (UX) design process, aiming at validating the targeted usability and impact of the analyzed application, as it will be explained further in the methodology section (typical system evaluation through surveys). The overall evaluation of the web interface along with the usability scenarios were supported by an empirical survey, concluding in the statistical results in the respective section of the manuscript. This survey was based on a guided interaction of users with the service, and afterwards, their opinions were recorded via a multi-factor questionnaire. “In this context, Nielson model [21,34] was mainly employed to evaluate aspects of the platform such as the Learnability (easy to learn), the Efficiency (meeting users’ expectations, effectiveness), Satisfaction (pleasant in use, potential entertaining character). However, the assessment criteria were augmented according to Quality in Use Measurement (QUIM) model [35,54] to involve factors such as Navigation (for accessible and organized navigation in the platform environment), Content (productive and informative material in web service, without errors). The final set of the 8 assessment factors is defined below, while a more detailed description is exhibited in Table 1.

- Efficiency (7 questions)

Table 1
Questionnaire of the conducted survey.

#	Factors	Measure
1	Efficiency – 7 items (effectiveness, productivity, usefulness, helpful, timesaving, covers user's needs, meets expectations)	Likert Scale for ordinal variables: 1-Strongly Disagree to 5-Strongly Agree
2	Usability – 12 items (easy to use, simple, user-friendly, low number of steps for execution, flexible, no special effort for use, use without manual, use without consulting, no inconsistencies, oriented to occasional/systematic users, minimum number of errors and easy recovering, successful using every time)	
3	Learnability – 5 items (quick learning, perceiving of use without memorizing options, easy to learn, user can easily be skillful in using it, helps the user to easily be productive into it)	
4	Satisfaction – 11 items (satisfied with it, recommendation to a friend, entertaining, functions as the user wants, excellent, feels to obtain, pleasant, good function integration, inspires confidences while using, effective completion of posed use scenarios, easy completion of posed use scenarios)	
5	Navigation – 4 items (clear hierarchical web structure, comprehensive web page, clear and precise navigation and processing options, easy web interface orientation)	
6	Content – 6 items (comprehensive terminology, comprehensive information structure, comprehensive information highlighting, adequate and precise content, no content errors and inaccuracies, clear details and instructions)	
7	Interactivity – 5 items (adaptivity and flexibility, options visibility, control handling during interaction, consistent control options with menus and buttons, minimizes the needed memory for using it)	
8	Design – 3 items (colors, graphics, web pages organization, and appearance)	
9	Hate Speech Detection	
10	Purposing Scenarios – 5 items (check my texts, check others' texts, contribute to the project's database, retrieve relative content from the project's database, conduct geo-location of articles)	
11	Addressing Scenarios – (secondary education, academic education, multidisciplinary research, vocational training, and digital literacy, events in social media, participatory infotainment scenarios)	
12	Gender (Male, Female, Other)	Binary Nominal
13	Age (18–22, 23–30, 31–40, 41–50, 50+)	5-level Ordinal
14	Education (Highschool, Bachelor, Master, PhD)	4-level Ordinal
15	Computer Familiarity (Low – folders/files management and Office software knowledge, High – specialized software and data analysis and programming)	Binary Nominal
16	Internet Familiarity (Low – search engines, emails, social networking, High – web page developing and skills in advanced web services)	Binary Nominal
17	News Awareness	Ordinal 5-level Likert Scale
18	Profession (Student, Academic Staff, Non Academic Profession)	3-class Nominal
19	Working/has worked as Journalist	Binary Nominal

- Usability (12 questions)
- Learnability (5 questions)
- Satisfaction (11 questions)
- Navigation (4 questions)
- Content (6 questions)
- Interactivity (5 questions)
- Design (3 questions)

As described in the functionalities section of the PHARM project, the main target of the developed interface is the detection of hate speech and the NLP-based sentiment load implication. For this reason, a dedicated question/item was involved in the survey assessing the fulfilment/contribution of the service towards the identification of hate speech mechanisms and sentiment loads in text data. Moreover, the involved participants posed their attitudes about possible purpose-of-use scenarios of the platform according to their knowledge, beliefs, and interests, along with potential area/scope of utilization in public services.

- Hate speech detection (1 item)
- Purposing scenarios (5 items)
- Addressing scenarios (6 items)

All the above questions/items were monitored in a Likert scale with 5 levels, ranging from Strongly Disagree (1) to Strongly Agree (5).

Furthermore, the demographic information of the respondents was recorded in categorical form concerning their Age, Gender, Educational level, and Profession. Since the evaluation process refers to a specialized web-based NLP service, the participants' Computer and Internet Familiarity were measured in separate binary variables referring to Low and High competencies/skills. Finally, taking into consideration that the dynamically evolving database of the PHARM project derives from social platforms and news sites, the respondents were asked about their News Awareness (Likert scale 1–5) and if they work/have worked as Journalists (Yes-No). Table 1 presents a detailed description of the questionnaire that supported the current survey.

The survey was conducted on a volunteer basis via the LimeSurvey online tool during the period February 1, 2021–March 31, 2021, while the final number of respondents reached to N = 298. Each participant was asked to visit the web platform and navigate/interact with the various pages, buttons, and processing modules and observe the returned results. Afterwards, each user had to complete the formulated questionnaire regarding the web service evaluation inquiries along with the possible use-case scenarios and demographics.

The overall length of the survey ranged from 8 to 11 min, terminated at the point of recording the user's response.

It has to be highlighted that a preliminary assessment procedure was conducted by Ref. [61]; extracting initial feedback in a beta evaluation mode from a lower number of participants (N = 64), since the questionnaire was disseminated via the authors' social networks. The pool of participants in the pilot (alpha) testing of the platform derived from the Schools of Electrical and Computer Engineering, and Journalism and Mass Communications of the Aristotle University of Thessaloniki. The aim was to form a multidisciplinary assessment group with prior experience both in the implementation and evaluation of user interfaces as well as in the development of backend services (machine/deep learning architectures analyzing audiovisual and textual content in terms of sentiment analysis, fake news, hate speech, etc.). Hence, the motivation behind this choice was to initially evaluate the purposing scenarios of the web service (how useful would be for journalists that want to detect hate speech/fake news, if the platform is easy to be handled by non-technical personnel, UX-design, etc.), also taking into consideration engineering and computing aspects (data flow, real-time geolocation, algorithmic optimizations, etc.). The users' remarks and opinions were taken thoroughly into consideration towards the optimization of crucial aspects of the web interface, while more processing modules are currently integrated to facilitate a user-friendly and efficient web environment that lives up to the users' needs and standards. In this context, during the current survey, the project is subjected to a generic evaluation framework, implemented via the LimeSurvey platform, in which the increased pool of participants were free to navigate in the "upgraded" version of the web interface.

Before proceeding into the results section, it has to be mentioned that reliability tests and internal consistency tests were imposed on the assessment factors (inner class and generic) based on the calculation of Cronbach's alpha measure, revealing in all cases coefficients ranging from a = 0.91 to a = 0.96. Hence, it can be supported that the subsequent statistical analysis supports confident statistical results and conclusions. Finally, Table 2 presents the basic demographic information of the respondents.

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

5. Results and discussion

5.1. The implemented web interface

The Python web framework Flask has been used to build the interface. This choice was made because all functionality can be bundled into a single software project using Python, as the NLP algorithms have been coded in Python too. The widely used HTML, CSS, and JavaScript package Bootstrap were used to create the graphical user interface (GUI). Waitress, a high-performance pure-Python web server gateway interface (WSGI), is used to serve the web tool. The PHARM Project and the aims that serves is presented at https://pharmproject.usal.es. The five following figures depict the main graphical interfaces for the monitor, search, and analyze functionalities. The home screen of the Interface provides some basic information about the PHARM project and input for performing the analyses on user-defined text. It also features sample content for analysis in the Greek, Italian, and Spanish languages. These three languages are natively supported, but content in all languages can be analyzed as well, via translation. The next screen concerns the "monitor" capability (Fig. 4). The Interface features real-time hate speech monitoring capabilities across the regional area of Europe. In specific, geolocated tweets are constantly being analyzed for containing hate speech, and detected hate-speech events are stored and displayed on a map (Fig. 4). In order to make the platform's functionality more comprehensible, the user may interact and experiment with self-imported content and extract the semantic description of the respective text in the following link: http://pharm-interface.usal.es/. Furthermore, the platform's presentation along with its inner properties can be accessed in http://pharm-interface.usal.es/static/quick_start_guide.pdf, while a detailed guide/instructions for all steps that are conducted during semantic analysis process can be found in http://pharm-interface.usal.es/instructions.

Next, the search screen is presented. Fig. 5 demonstrates the available filters and their corresponding controls. Search results are presented as a list of records. The user can preview the records, and display details for each one of them. Either the "Simple" or the "Scientific" view can be chosen. The "scientific" view presents metadata along with data. The list can be downloaded as a CSV or JSON file. Sample files in both formats can be previewed via the following links: Standard View and Scientific View.

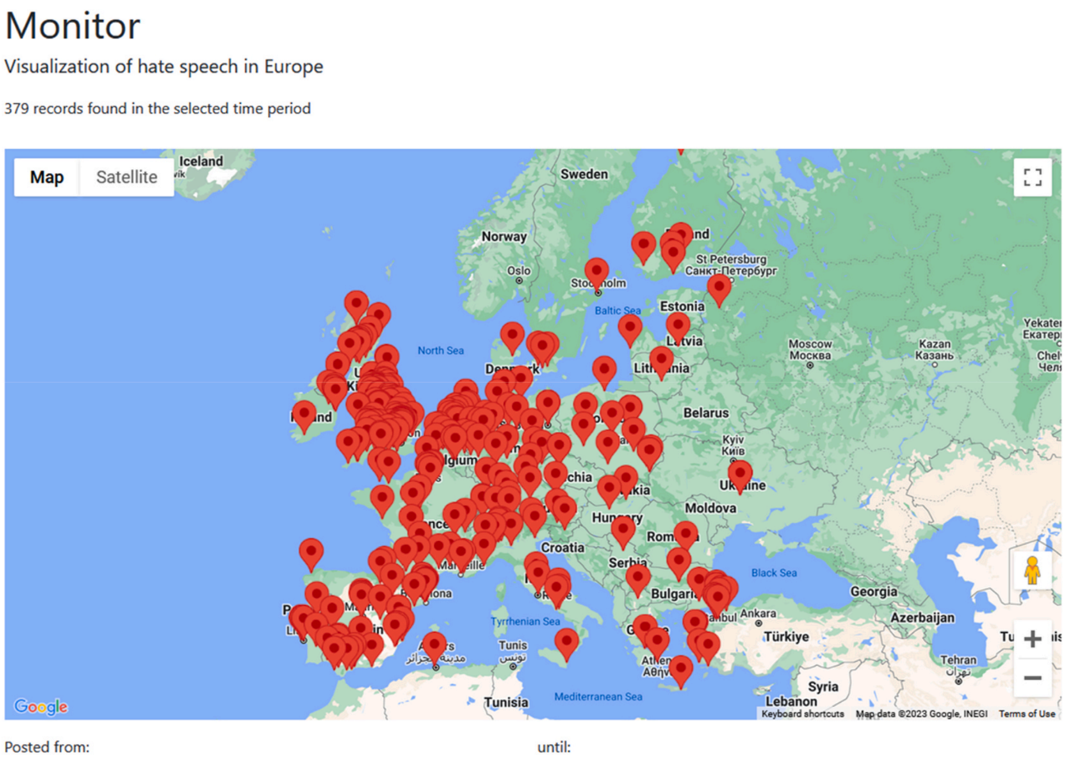
When a record is selected (or for a text that is placed on the home screen), a view presenting detailed information appears (Fig. 6).

5.2. User experience and usability assessment

The development of the web platform launched on October 2020, while major optimizations and flaws corrections took place,

Table 2
Distributions of demographic data.

#	Factor	Distribution
1	Gender	Men (36%), Women (64%)
2	Age	18-22 (38.3%), 23–30 (25.5%), 31–40 (18.1%), 41–50 (12.1%), 50+ (6%)
3	Education	Highschool (40.9%), Bachelor (24.5%), Master (22.8%), PhD (11.8%)
4	Computer Familiarity	Low (54.6%), High (45.4%)
5	Internet Familiarity	Low (37.1%), High (62.9%)
6	News Awareness	Not at all (4.4%), Rarely (13.1%), Sometimes (27.5%), Often (26.2%), Very Often (28.9%)
7	Working/has worked as a Journalist	Yes (16.2%), No (83.8%)



Monitor

Visualization of hate speech in Europe

379 records found in the selected time period

Posted from: until:

Fig. 4. The real-time monitoring screen of the PHARM Interface.

based on preliminary alpha evaluation testing by domain experts (in fields of web design and graphics, software engineering, etc.) and the subsequent beta evaluation of web interface. In this section, the experimental results of the conducted survey (N = 298) are presented regarding an overall usability assessment along with the purpose-of-use scenarios of the PHARM project.

To begin with the evaluation of the usefulness of the approach, Fig. 7 answers to the suitability of the web service to meet the main designated aim of supporting the detection of hate speech and sentiment load in text data, concluding in an increased acceptance of 60% of Agree/Strongly, while only 19.2% of the participants Disagree/Strongly Disagree with it.

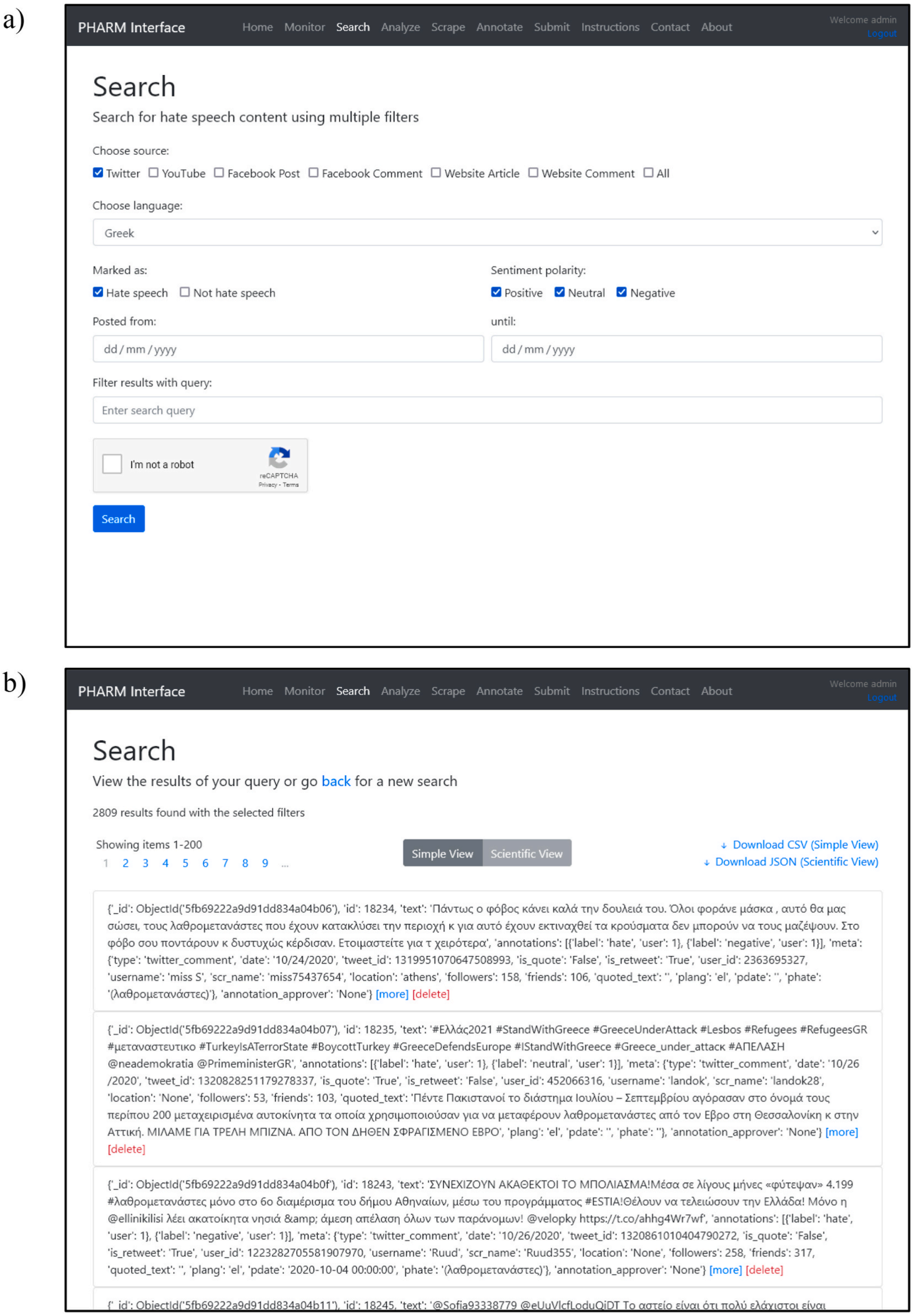
Furthermore, as Table 1 exhibited, the 8 assessment factors were represented by a different number of items, therefore, the mean scores of Likert-scale responses were computed, towards the computation of average evaluation metrics, since the Cronbach's alpha measure validates the respective factor integrity (above 0.91 in all cases). Fig. 8 exhibits the mean score and standard deviation of every factor for the N = 298 participants, revealing in almost all aspects satisfactory results that are prone to 4 (with standard deviations lower than 1). In this context, early conclusions are drawn, indicating a well-designed web interface with increased usability, withholding comprehensive and easy-to-learn content, interaction, and navigation mechanisms. Further experiments/tests within groups of the correspondents were subsequently employed and presented to pinpoint differences among the formulated clusters.

Towards a more thorough evaluation of the developed web service, the mean scores were compared within the 5 groups of participants, according to their agreement regarding the fulfilment of the main purpose of the platform for hate speech detection. Taking into consideration that Levene homogeneity of variance test and Shapiro-Wilk normality test indicated values of $p < 0.001$, the non-parametric method of Kruskal Wallis H method was utilized for detecting the mean scores differentiations among the 5 groups (df = 4). Table 3 presents the H and p-values values that were calculated for each assessment factor along with the effect size of the statistically significant differences, while Fig. 9 exhibits the mean evaluation scores for the groups of participants.

In all cases there is a large effect that combined with the subsequent post-hoc Mann Whitney tests pointed in the direction that the participants that didn't agree that the web service support its main task, evaluated it in statistically significant lower scores ($p < 0.001$) compared to those who agreed to the respective question. This fact is also validated in Fig. 9, where the average values range between 1.5 (for efficiency) to 2.6 (for learnability), while all the assessment metrics are scored around 4 for those who strongly Agree/Agree to the question under examination. The above remarks lead to the conclusion that platform optimizations are essential to convince for its orientation and therefore increase the evaluation scores.

Because of the core functionality of the web platform towards aggregating and checking news feeds from Twitter, YouTube, and News webpages, an interesting relation had to be examined between the assessment factors and the variables of users' News Awareness and that representing whether they work/have worked as Journalists. For the first analysis the Kruskal-Wallis method was utilized because of the five participants groups (based on answers: Not at all, Rarely, Sometimes, Often, Very Often – df = 4), while the Mann-Whitney test for the Journalist binary variable (Yes, No).

One crucial aspect of the whole implementation refers to the potential use case and scenarios of the developed platform. For this



Search

Search for hate speech content using multiple filters

Choose source:

☑ Twitter □ YouTube □ Facebook Post □ Facebook Comment □ Website Article □ Website Comment □ All

Choose language:

Greek

Marked as:

☑ Hate speech □ Not hate speech

Sentiment polarity:

☑ Positive ☑ Neutral ☑ Negative

Posted from:

dd / mm / yyyy

until:

dd / mm / yyyy

Filter results with query:

Enter search query

Search

View the results of your query or go back for a new search

2809 results found with the selected filters

Showing items 1-200

1 2 3 4 5 6 7 8 9 ...

Simple View
Scientific View
↓ Download CSV (Simple View)
↓ Download JSON (Scientific View)

{'_id': ObjectId('5fb69222a9d91dd834a04b06'), 'id': 18234, 'text': 'Πάντως ο φόβος κάνει καλά την δουλειά του. Όλοι φοράνε μάσκα , αυτό θα μας σώσει, τους λαθρομετανάστες που έχουν κατακλύσει την περιοχή κ για αυτό έχουν εκτιναχθεί τα κρούσματα δεν μπορούν να τους μαζέψουν. Στο φόβο σου ποντάρουν κ δυστυχώς κέρδισαν. Ετοιμαστείτε για τ χειρότερα', 'annotations': [{'label': 'hate', 'user': 1}, {'label': 'negative', 'user': 1}], 'meta': {'type': 'twitter_comment', 'date': '10/24/2020', 'tweet_id': 1319951070647508993, 'is_quote': 'False', 'is_retweet': 'True', 'user_id': 2363695327, 'username': 'miss S', 'scr_name': 'miss75437654', 'location': 'athens', 'followers': 158, 'friends': 106, 'quoted_text': '', 'plang': 'el', 'pdate': '', 'phate': '(λαθρομετανάστες)'}, 'annotation_approver': 'None'} [more] [delete]

{'_id': ObjectId('5fb69222a9d91dd834a04b07'), 'id': 18235, 'text': '#Ελλάς2021 #StandWithGreece #GreeceUnderAttack #Lesbos #Refugees #RefugeesGR #μεταναστευτικο #TurkeyIsATerrorState #BoycottTurkey #GreeceDefendsEurope #IStandWithGreece #Greece_under_attack #ΑΠΕΛΑΣΗ @neademokratia @PrimeministerGR', 'annotations': [{'label': 'hate', 'user': 1}, {'label': 'neutral', 'user': 1}], 'meta': {'type': 'twitter_comment', 'date': '10/26 /2020', 'tweet_id': 1320828251179278337, 'is_quote': 'True', 'is_retweet': 'False', 'user_id': 452066316, 'username': 'landok', 'scr_name': 'landok28', 'location': 'None', 'followers': 53, 'friends': 103, 'quoted_text': 'Πέντε Πακιστανοί το διάστημα Ιουλίου – Σεπτεμβρίου αγόρασαν στο όνομά τους περίπου 200 μεταχειρισμένα αυτοκίνητα τα οποία χρησιμοποιούσαν για να μεταφέρουν λαθρομετανάστες από τον Εβρο στη Θεσσαλονίκη κ στην Αττική. ΜΙΛΑΜΕ ΓΙΑ ΤΡΕΛΗ ΜΠΙΖΝΑ. ΑΠΟ ΤΟΝ ΔΗΘΕΝ ΣΦΡΑΓΙΣΜΕΝΟ ΕΒΡΟ', 'plang': 'el', 'pdate': '', 'phate': ''}, 'annotation_approver': 'None'} [more] [delete]

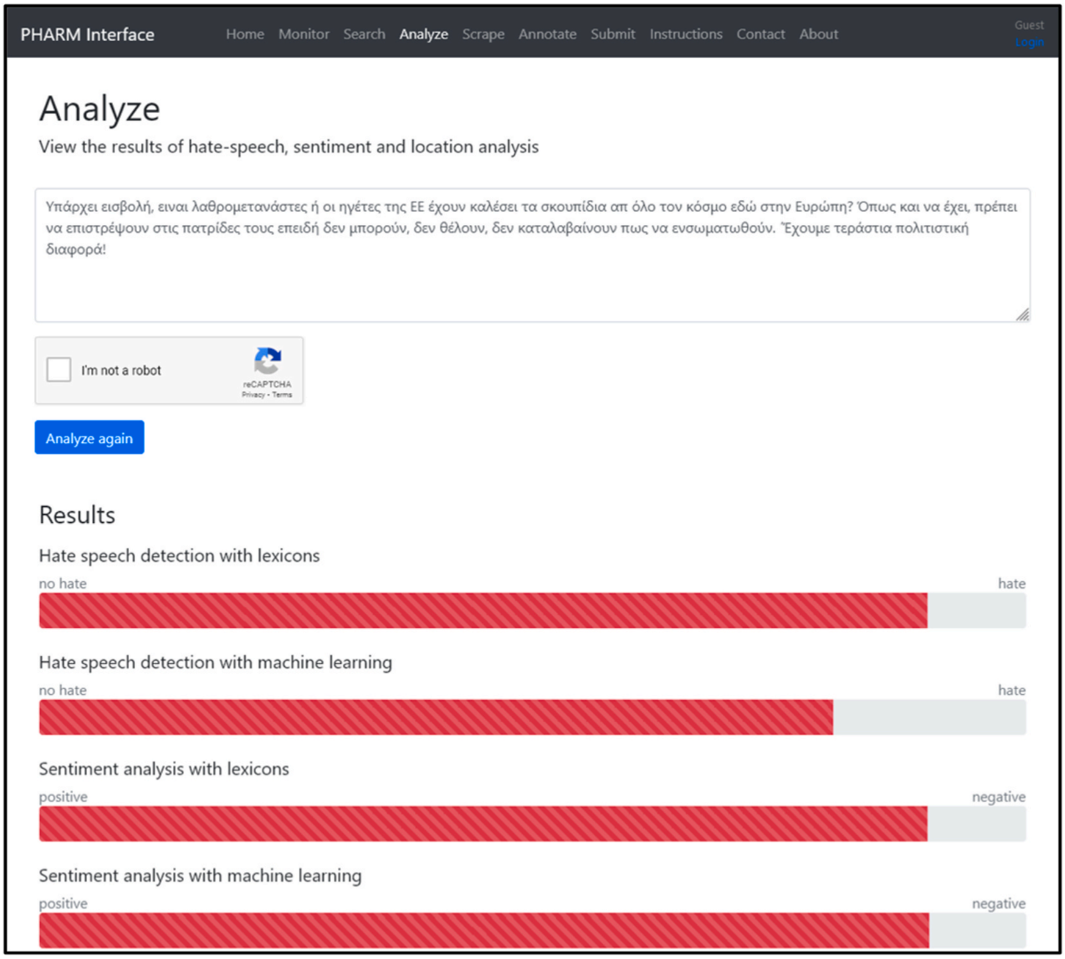
{'_id': ObjectId('5fb69222a9d91dd834a04b0f'), 'id': 18243, 'text': 'ΣΥΝΕΧΙΖΟΥΝ ΑΚΑΘΕΚΤΟΙ ΤΟ ΜΠΟΛΙΑΣΜΑ!Μέσα σε λίγους μήνες «φύτεψαν» 4.199 #λαθρομετανάστες μόνο στο 6ο διαμέρισμα του δήμου Αθηναίων, μέσω του προγράμματος #ESTIA!Θέλουν να τελειώσουν την Ελλάδα! Μόνο η @ellinikilisi λέει ακατοίκητα νησιά &amp; άμεση απέλαση όλων των παράνομων! @velopky https://t.co/ahhg4Wr7wf', 'annotations': [{'label': 'hate', 'user': 1}, {'label': 'negative', 'user': 1}], 'meta': {'type': 'twitter_comment', 'date': '10/26/2020', 'tweet_id': 1320861010404790272, 'is_quote': 'False', 'is_retweet': 'True', 'user_id': 1223282705581907970, 'username': 'Ruud', 'scr_name': 'Ruud355', 'location': 'None', 'followers': 258, 'friends': 317, 'quoted_text': '', 'plang': 'el', 'pdate': '2020-10-04 00:00:00', 'phate': '(λαθρομετανάστες)'}, 'annotation_approver': 'None'} [more] [delete]

{'_id': ObjectId('5fb69222a9d91dd834a04b11'), 'id': 18245, 'text': '@Sofia93338779 @eUuVlcfLoduQiDT Το αστείο είναι ότι πολύ ελάχιστοι είναι

Fig. 5. The search input form (a) and search results (b) screens of the PHARM Interface.

reason, the main purpose of the hate speech detection along with the purpose and addressing scenarios (expressed in Likert scale as stated in Table 1) of the web service were statistically analyzed in overall terms and within groups of participants.

Figs. 10 and 11 present the respondents' opinions (%) about the suitability and use-case potentials, showing that participants would



Analyze

View the results of hate-speech, sentiment and location analysis

Υπάρχει εισβολή, ειναι λαθρομετανάστες ή οι ηγέτες της ΕΕ έχουν καλέσει τα σκουπίδια απ όλο τον κόσμο εδώ στην Ευρώπη? Όπως και να έχει, πρέπει να επιστρέψουν στις πατρίδες τους επειδή δεν μπορούν, δεν θέλουν, δεν καταλαβαίνουν πως να ενσωματωθούν. Έχουμε τεράστια πολιτιστική διαφορά!

Results

Hate speech detection with lexicons

Hate speech detection with machine learning

Sentiment analysis with lexicons

Sentiment analysis with machine learning

Fig. 6. The analysis screen of the PHARM Interface.

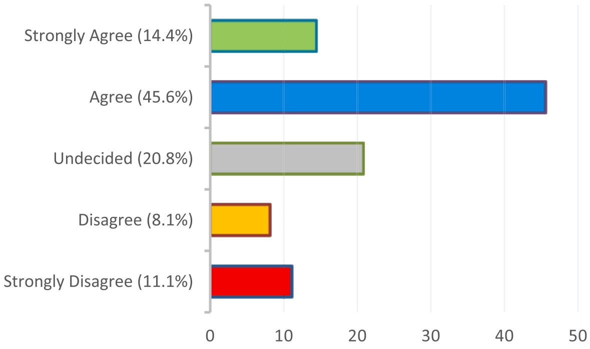


Fig. 7. Suitability of web service for hate speech detection.

mostly utilize the platform for checking other's texts (64% agreement) and retrieving relative content from the project's database (61% agreement). Furthermore, they suggested that the web service would be more suitable in Academic Education and Multidisciplinary research (72% agreement), closely followed by events in social media (71% agreement). However, the respective percentages are above 65% for all scenarios, proving the versatile nature that the web service withholds, offering various/adaptive ways for its utilization. The subsequent statistical analysis investigates the evaluated mean scores differentiations within groups of variables, aiming to extract more specific knowledge about users' preferences.

In the beginning, the relations between Computer and Internet Familiarity were examined with the agreement on the aforementioned scenarios. Table 4 presents the results of the Mann-Whitney tests, while Figs. 12 and 13 the respective group scores. For

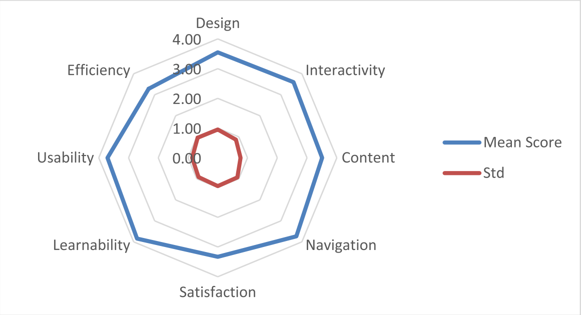


Fig. 8. Mean scores of the eight (8) assessment factors.

Table 3
Kruskal Wallis H tests and effect sizes for mean scores differences into the groups of participants for purpose fulfillment of the web service.

Factor	H	p-value	η^2
Efficiency	149.801	<0.001[a]	0.498
Usability	85.889	<0.001[a]	0.279
Learnability	52.967	<0.001[a]	0.167
Satisfaction	126.307	<0.001[a]	0.417
Navigation	58.201	<0.001[a]	0.185
Content	100.174	<0.001[a]	0.328
Interactivity	84.957	<0.001[a]	0.276
Design	58.671	<0.001[a]	0.187

[a] Statistically significant difference between groups at a = 0.05 significance level.

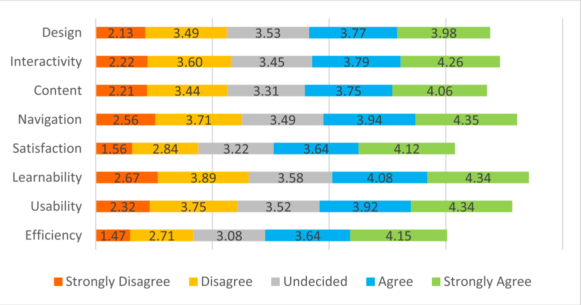


Fig. 9. Mean evaluation scores of the groups of participants for purpose fulfillment of the web service.

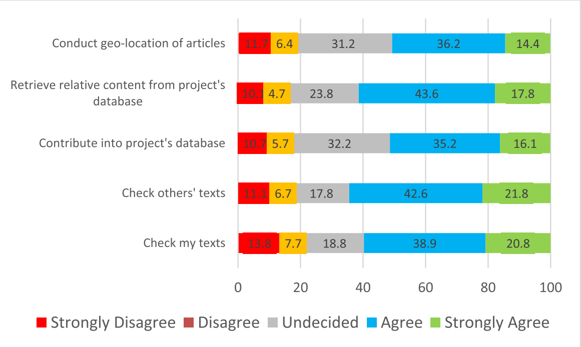


Fig. 10. Participants' agreement on purposing scenarios of the developed web platform.

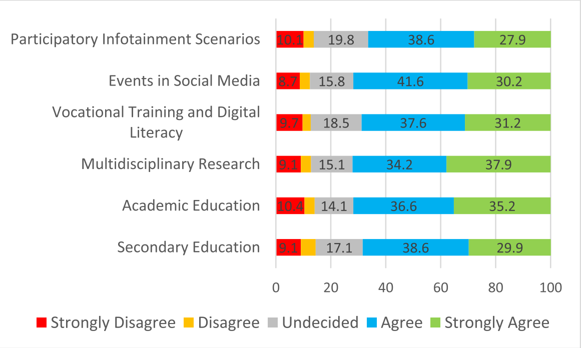


Fig. 11. Participants' agreement on addressing-to scenarios of the developed web platform.

Computer Familiarity groups there was no statistical difference in use-case evaluation scores, but when moving into more focused knowledge of Internet orientation the situation changes. Specifically, the High Familiarity group was evaluated with a statistically significant increased score compared to the Low one in the addressing scenarios of Vocational Training and Digital Literacy ($p = 0.009$, $\mu_{high} = 3.86$, $\mu_{low} = 3.63$), Events in Social Media ($p = 0.017$, $\mu_{high} = 3.86$, $\mu_{low} = 3.72$) and Participatory Infotainment Scenarios ($p = 0.010$, $\mu_{high} = 3.79$, $\mu_{low} = 3.56$), however with small effect sizes ($\eta^2 \approx 0.020$ in all cases). In general, the results imply that the users' technical background/knowledge brings a small impact on the multivariate utilization/character of the developed web platform.

Moving forward, the participants' news relationship is investigated towards their respective evaluation/agreement on the suitability of web service in the proposed potential utilizations. Specifically, Table 5 exhibits the statistical Kruskal Wallis test for News Awareness factor (5 groups, df = 4), while the Mann-Whitney method was employed for the variable of Working/has Worked as Journalist (binary variable). Again, statistically significant differentiations of scenarios evaluation appear solely among the respondents' groups in News Awareness and only for the addressing cases of Events in Social Media ($p = 0.022$, $\eta^2 = 0.025$) and Participatory Infotainment Scenarios ($p = 0.026$, $\eta^2 = 0.024$), with small to intermediate effect sizes. The subsequent Mann-Whitney post-hoc tests determined that these differences are located for both factors only between the groups of participants that replied Often and Very Often to the News Awareness inquiry. Specifically, the group of Often respondents evaluated statistically higher (compared to the Very Often one), the case of Events in Social Media ($p = 0.011$, $\mu_{often} = 4.13$, $\mu_{very\ often} = 3.43$) and the Participatory Infotainment Scenarios ($p = 0.010$, $\mu_{often} = 4.03$, $\mu_{very\ often} = 3.33$). On the other hand, the journalistic profession doesn't appear to affect the general, purpose and addressing scenarios, while it has to be highlighted that in all cases the group related to Journalism participants noted a pattern of slightly decreased scores.

Finally, the statistical analysis proceeded into the investigation of the scenarios' evaluation rate within the demographic groups of participants, while Tables 6 and 7 gather the respective results. Starting from the Profession attribute, statistically significant differences appeared for the purpose scenario of Contributing to the project's database ($p = 0.032$) and the addressing one related to Secondary Education ($p = 0.048$), however with small effect sizes. Specifically, the corresponding Mann-Whitney post-hoc tests revealed that this difference appeared only between of 31–40 with 50+ age clusters for both scenarios, with the first one evaluating statistically higher the suitability of the web service for Contributing to the project's database ($p = 0.047$, $\mu_{31\text{-}40} = 3.70$, $\mu_{50+} = 2.83$) and utilization into the Secondary Education ($p = 0.048$, $\mu_{31\text{-}40} = 3.91$, $\mu_{50+} = 3.06$). Furthermore, it has to be highlighted that the Age group of 31–40 noted higher scoring for all use cases compared to the rest participants' groups. On the other hand, the Educational level seems to impact more strongly the proposed utilization scenarios as Table 6 exhibits, specifically into 3 Purposing variables and 3 Addressing-to ones. Master's degree holders agree higher on all the use-case scenarios of the web service, compared to the rest of the

Table 4
Statistical tests for mean scores of use scenarios into the groups of Computer and Internet Familiarity.

	Use Scenarios	Computer Familiarity		Internet Familiarity		
		U	p	U	p	η^2
Main task	The web interface helps into hate speech and sentiment load detection	10,776	0.746	9517	0.255	
Purpose Scenarios	Check my texts	10447.5	0.435	10,278	0.992	
	Check others' texts	10914.5	0.901	9159	0.098	
	Contribute into project's database	10,077	0.192	9399.5	0.196	
	Retrieve relative content from project's database	10392.5	0.384	9204.5	0.111	
	Conduct geo-location of articles	10888.5	0.872	9897	0.571	
Addressing Scenarios	Secondary Education	10515.5	0.491	9410	0.200	
	Academic Education	10655.5	0.622	9268	0.134	
	Multidisciplinary Research	10,872	0.853	9020	0.062	
	Vocational Training and Digital Literacy	10,986	0.981	8513	0.009[a]	0.021
	Events in Social Media	10624.5	0.590	8666	0.017[a]	0.017
	Participatory Infotainment Scenarios	10,723	0.693	8533	0.010[a]	0.020

[a] Statistically significant difference between groups at a = 0.05 significance level.

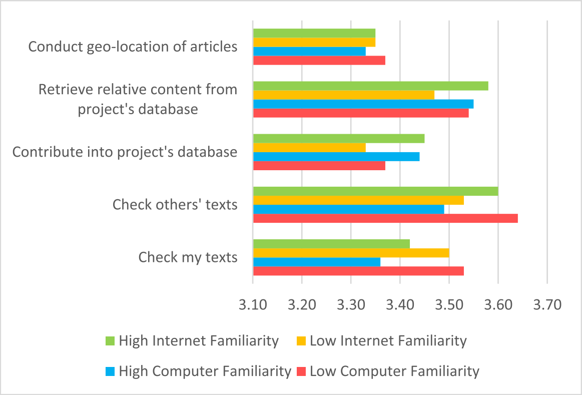


Fig. 12. Purpose scenarios scores in Computer and Internet Familiarity groups.

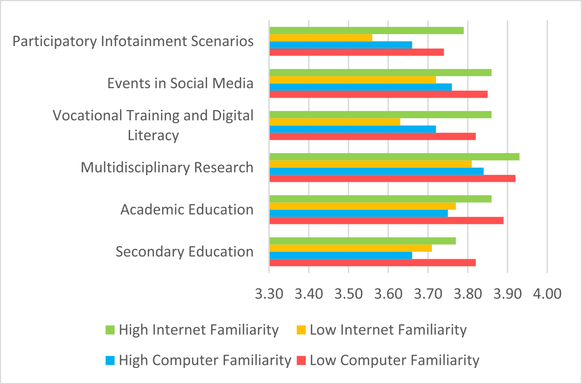


Fig. 13. Addressing scenarios scores in Computer and Internet Familiarity groups.

Table 5
Statistical tests for mean scores of use scenarios into the groups of news awareness and working/has worked as Journalist.

	Use Scenarios	News Awareness		Working/has Worked as Journalist		
		H	p	η^2	U	p
Main task Purpose Scenarios	The web interface helps with hate speech and sentiment load detection	3.565	0.468		5688.5	0.546
	Check my texts	2.102	0.717		5927	0.889
	Check others' texts	6.713	0.152		5900.5	0.848
	Contribute to the project's database	3.779	0.437		5959.5	0.938
	Retrieve relative content from the project's database	4.933	0.294		5812.5	0.717
	Conduct geo-location of articles	6.084	0.193		5981.5	0.972
Addressing Scenarios	Secondary Education	4.282	0.369		5762.5	0.649
	Academic Education	3.206	0.524		5412.5	0.258
	Multidisciplinary Research	6.536	0.163		5607	0.449
	Vocational Training and Digital Literacy	6.354	0.174		5764	0.651
	Events in Social Media	11.423	0.022[a]	0.025	5633.5	0.479
	Participatory Infotainment Scenarios	11.007	0.026[a]	0.024	5946.5	0.918

[a] Statistically significant difference between groups at a = 0.05 significance level.

categories of educational background. However, the biggest effect sizes yield in the Contribution to the project's database ($p = 0.007$, $\eta^2 = 0.031$) and Vocational Training/Digital Literacy ($p = 0.009$, $\eta^2 = 0.029$), while the post-hoc tests revealed differentiations for both scenarios between Highschool with Master ($p = 0.027$ and $p = 0.019$ respectively) and PhD with Master educational levels ($p = 0.013$ and $p = 0.027$ respectively). For the rest 4 use cases, the mean agreement scores differentiations are detected with small effect sizes between Master and PhD groups of educational levels ($p = 0.033$ for Checking others' texts, $p = 0.049$ for Conducting articles geolocation, $p = 0.021$ for Events Social Media orientation and $p = 0.017$ for Participatory Infotainment Scenarios). Proceeding into Gender effects, female participants expressed increased agreement rates for all the utilization scenarios compared to male ones, possibly pointing towards more awareness for hate speech detection tasks and multivariate utilization of the web service. As

Table 6
Statistical tests for mean scores of use scenarios into the groups of the demographic variables age (df = 4) and education (df = 3).

	Use Scenarios	Age			Education		
		H	p	η^2	H	p	η^2
Main task	The web interface helps with hate speech and sentiment load detection	3.027	0.553		3.508	0.320	
Purpose Scenarios	Check my texts	7.093	0.131		5.467	0.141	
	Check others' texts	8.444	0.077		9.377	0.025[a]	0.022
	Contribute to the project's database	10.536	0.032[a]	0.022	12.042	0.007[a]	0.031
	Retrieve relative content from the project's database	9.375	0.052		6.622	0.085	
	Conduct geo-location of articles	9.431	0.051		9.656	0.022[a]	0.023
Addressing Scenarios	Secondary Education	9.598	0.048[a]	0.019	6.015	0.111	
	Academic Education	4.847	0.303		5.253	0.154	
	Multidisciplinary Research	3.016	0.555		7.258	0.064	
	Vocational Training and Digital Literacy	7.077	0.132		11.465	0.009[a]	0.029
	Events in Social Media	3.630	0.458		8.919	0.030[a]	0.020
	Participatory Infotainment Scenarios	8.502	0.075		9.377	0.025[a]	0.022

[a] Statistically significant difference between groups at a = 0.05 significance level.

Table 7
Statistical tests for mean scores of Use Scenarios into the groups of the demographic variables Gender and Profession.

	Use Scenarios	Gender			Profession	
		U	p	η^2	H	p
Main task	The web interface helps with hate speech and sentiment load detection	7870.5	0.047[a]	0.012	2.163	0.339
Purpose Scenarios	Check my texts	6031.5	<0.001[a]	0.078	0.150	0.928
	Check others' texts	7729.5	0.029[a]	0.015	2.044	0.360
	Contribute to the project's database	8511.5	0.354		1.638	0.441
	Retrieve relative content from the project's database	7810	0.038[a]	0.014	2.257	0.324
	Conduct geo-location of articles	8168.5	0.14		3.839	0.147
Addressing Scenarios	Secondary Education	7133	0.002[a]	0.032	0.169	0.919
	Academic Education	6819.5	<0.001[a]	0.043	0.525	0.769
	Multidisciplinary Research	7260.5	0.003[a]	0.028	1.035	0.596
	Vocational Training and Digital Literacy	7732	0.029[a]	0.015	1.402	0.496
	Events in Social Media	7405.5	0.006[a]	0.024	1.240	0.538
	Participatory Infotainment Scenarios	7102	0.001[a]	0.033	0.780	0.677

[a] Statistically significant difference between groups at a = 0.05 significance level.

Table exhibits the larger effect sizes are recorded in Check my texts orientation ($\eta^2 = 0.078$), while the mean scores appear quite different between male/female participants for the utilization of the web service in Academic Education ($\eta^2 = 0.043$). Finally, the profession of the respondents didn't seem to impact their evaluation regarding the purposing and addressing scenarios.

6. Conclusion and future directions

Overall, the presented interface was positively evaluated by the participants, indicating a well-designed web application for a wide variety of user groups. Moreover, the users' technical background has a small impact on the multivariate utilization of the developed web platform. The interface can be primarily used for checking texts with possible hate speech content and, secondly, for retrieving relevant content from the database. Furthermore, the results denote that the platform is more suitable for Academic Education and Multidisciplinary research, while it can be taken into consideration in users' social media actions/events.

However, the analysis revealed some weaknesses of the platform, and special attention will be given towards overcoming these issues. Most notably, the participants who didn't agree that the web service supports its main task, evaluated the platform lower compared to users that acknowledged the usefulness of the PHARM project. According to the aforementioned observations, platform optimizations are crucial to persuade for its importance and orientation, hence raise assessment ratings. This finding indicates that the improvement of the usability and design aspects of the platform alone cannot guarantee an improved overall experience. It is important to define the different use cases, target groups, and user roles to support the value of the provided functionality in order to engage users in using the interface. This can be addressed through dissemination acts that include new users in the process of reevaluation of the project and the definition of future directions.

The increasing data volumes that are being gathered and annotated using the PHARM Interface are being monitored while users' feedback is constantly evaluated. This feedback, combined with the examination of web analytics, direct the next steps for further refining the design of the interface and improving its usability. According to the chosen human-centred LUCID design, additional evaluation phases will follow towards succeeding in delivering an online tool for text analysis that can be easily adapted and deployed in other than detecting hate speech contexts, as well.

Author contribution statement

Rigas Kotsakis: Lazaros Vrysis: Nikolaos Vryzas: Theodora Saridou: Maria Matsiola: Andreas Veglis: Charalampos Dimoulas: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Data availability statement

Data associated with this study has been deposited at https://pharmproject.usal.es/?lang=el.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

References

[1] A. AlDayel, W. Magdy, Stance detection on social media: state of the art and trends, Inf. Process. Manag. 58 (4) (2021), 102597.

[2] D. Antonakaki, P. Fragopoulou, S. Ioannidis, A survey of Twitter research: data model, graph structure, sentiment analysis and attacks, Expert Syst. Appl. 164 (2021), 114006.

[3] A. Arango, J. Pérez, B. Poblete, Hate speech detection is not as easy as you may think: a closer look at model validation, in: Proceedings of the 42nd Int*ernational Acm Sigir Conference on Research and Development in Information Retrieval*, 2019, July, pp. 45–54.

[4] C. Arcila-Calderón, D. Blanco-Herrero, M. Matsiola, M. Oller-Alonso, T. Saridou, S. Splendore, A. Veglis, Framing migration in southern European media: perceptions of Spanish, Italian, and Greek specialized journalists, J. Pract. (2021) 1–24, https://doi.org/10.1080/17512786.2021.2014347.

[5] C. Arcila-Calderón, J.J. Amores, P. Sánchez-Holgado, L. Vrysis, N. Vryzas, M. Oller Alonso, 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) (2022), 13094.

[6] F.E. Ayo, O. Folorunso, F.T. Ibharalu, I.A. Osinuga, A. Abayomi-Alli, A probabilistic clustering model for hate speech classification in twitter, Expert Syst. Appl. 173 (2021), 114762.

[7] F.E. Ayo, O. Folorunso, F.T. Ibharalu, I.A. Osinuga, Machine learning techniques for hate speech classification of twitter data: state-of-the-art, future challenges and research directions, Computer Science Review 38 (2020), 100311.

[8] M. Barnidge, B. Kim, L.A. Sherrill, Ž. Luknar, J. Zhang, Perceived exposure to and avoidance of hate speech in various communication settings, Telematics Inf. 44 (2019), 101263.

[9] D.R. Beddiar, M.S. Jahan, M. Oussalah, Data expansion using back translation and paraphrasing for hate speech detection, Online Soc. Net. Media 24 (2021), 100153.

[10] P. Berkowitz, The harm in hate speech, Hedgehog Rev. 15 (3) (2013) 100–102.

[11] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[12] S. Boberg, T. Schatto-Eckrodt, L. Frischlich, T. Quandt, The moral gatekeeper? Moderation and deletion of user-generated content in a leading news forum, Media Commun. 6 (4) (2018) 58–69.

[13] N. Chetty, S. Alathur, Hate speech review in the context of online social networks, Aggress. Violent Behav. 40 (2018) 108–118.

[14] N. Chetty, S. Alathur, An architecture for digital hate content reduction with mobile edge computing, Digital Communic. Networks 6 (2) (2020) 217–222.

[15] K. Clemens, Geocoding with openstreetmap data, in: Proceedings of the GEO Processing 2015, Lisbon, Portugal, 22–27 February, 2015, p. 10.

[16] M. Costello, J. Hawdon, T. Ratliff, T. Grantham, Who views online extremism? Individual attributes leading to exposure, Comput. Hum. Behav. 63 (2016) 311–320.

[17] K. Crowston, I. Fagnot, Stages of motivation for contributing user-generated content: a theory and empirical test, Int. J. Hum. Comput. Stud. 109 (2018) 89–101.

[18] W.H. Dutton, B.C. Reisdorf, Cultural divides and digital inequalities: attitudes shaping Internet and social media divides, Inf. Commun. Soc. 22 (1) (2019) 18–38, https://doi.org/10.1080/1369118X.2017.1353640.

[19] S.A. Einwiller, S. Kim, How online content providers moderate user-generated content to prevent harmful online communication: an analysis of policies and their implementation, Pol. Internet 12 (2) (2020) 184–206.

[20] P. Fortuna, J. Soler-Company, L. Wanner, How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? Inf. Process. Manag. 58 (3) (2021), 102524.

[21] E. Gonzalez-Holland, D. Whitmer, L. Moralez, M. Mouloua, Examination of the use of nielsen's 10 usability heuristics & outlooks for the future, Proc. Hum. Factors Ergon. Soc. Annu. Meet. 61 (1) (2017) 1472–1475.

[22] K.M. Grosser, V. Hase, F. Wintterlin, Trustworthy or shady? Journal. Stud. 20 (4) (2019) 500–522, https://doi.org/10.1080/1461670X.2017.1392255.

[23] J. Han, Z. Zhang, N. Cummins, B. Schuller, Adversarial training in affective computing and sentiment analysis: recent advances and perspectives, IEEE Comput. Intell. Mag. 14 (2) (2019) 68–81.

[24] J. Hawdon, A. Oksanen, P. Räsänen, Exposure to online hate in four nations: a cross-national consideration, Deviant Behav. 38 (3) (2017) 254–266.

[25] M. Herz, P. Molnár (Eds.), The Content and Context of Hate Speech: Rethinking Regulation and Responses, Cambridge University Press, 2012.

[26] G. Kalliris, M. Matsiola, C. Dimoulas, A. Veglis, Emotional aspects and quality of experience for multifactor evaluation of audiovisual content, Int. J. Monit. Surveill. Technol. Res. 2 (4) (2014) 40–61.

[27] A. Kalogeropoulos, R.K. Nielsen, Social inequalities in news consumption. https://ssrn.com/abstract=3270975, 2018.

[28] P. Kapil, A. Ekbal, A deep neural network based multi-task learning approach to hate speech detection, Knowl. Base Syst. 210 (2020), 106458.

[29] A.N. Katsaounidou, A. Gardikiotis, N. Tsipas, C.A. Dimoulas, News authentication and tampered images: evaluating the photo-truth impact through image verification algorithms, Heliyon 6 (12) (2020), e05808.

[30] J. Kocoń, A. Figas, M. Gruza, D. Puchalska, T. Kajdanowicz, P. Kazienko, Offensive, aggressive, and hate speech analysis: from data-centric to human-centered approach, Inf. Process. Manag. 58 (5) (2021), 102643.

[31] R. Kotsakis, C. Dimoulas, G. Kalliris, A. Veglis, Emotional prediction and content profile estimation in evaluating audiovisual mediated communication, Int. J. Monit. Surveill. Technol. Res. 2 (4) (2014) 62–80.

[32] A. Kumar, S. Abirami, T.E. Trueman, E. Cambria, Comment toxicity detection via a multichannel convolutional bidirectional gated recurrent unit, Neurocomputing 441 (2021) 272–278.

[33] M. Laurent, Project Hatemeter: helping NGOs and Social Science researchers to analyze and prevent anti-Muslim hate speech on social media, Proc. Comput. Sci. 176 (2020) 2143–2153.

[34] L. Leventhal, J. Barnes, Usability Engineering Process, Products, and Examples, Pearson Prentice Hall, 2008.

[35] I.-F. Liu, M.C. Chen, Y.S. Sun, D. Wible, C.-H. Kuo, Extending the TAM model to explore the factors that affect intention to use an online learning community, Comput. Educ. 54 (2) (2010) 600–610.

[36] S. Modha, P. Majumder, T. Mandl, C. Mandalia, Detecting and visualizing hate speech in social media: a cyber watchdog for surveillance, Expert Syst. Appl. 161 (2020), 113725.

[37] I. Mollas, Z. Chrysopoulou, S. Karlos, G. Tsoumakas, ETHOS: a multi-label hate speech detection dataset, Complex & Intellig. Syst. 8 (6) (2022) 4663–4678.

[38] Z. Mossie, J.H. Wang, Vulnerable community identification using hate speech detection on social media, Inf. Process. Manag. 57 (3) (2020), 102087.

[39] S. Nagar, F.A. Barbhuiya, K. Dey, Towards more robust hate speech detection: using social context and user data, Soc. Network Anal. Mining 13 (1) (2023) 47.

[40] Netsafe, Online hate speech: a survey on personal experiences and exposure among adult New Zealanders. https://www.netsafe.org.nz/wp-content/uploads/2019/11/onlinehatespeechsurvey-2018.pdf, 2018.

[41] A. Obadimu, T. Khaund, E. Mead, T. Marcoux, N. Agarwal, Developing a socio-computational approach to examine toxicity propagation and regulation in COVID-19 discourse on YouTube, Inf. Process. Manag. (2021), 102660.

[42] E.W. Pamungkas, V. Basile, V. Patti, Misogyny detection in twitter: a multilingual and cross-domain study, Inf. Process. Manag. 57 (6) (2020), 102360.

[43] E.W. Pamungkas, V. Basile, V. Patti, A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection, Inf. Process. Manag. 58 (4) (2021), 102544.

[44] K. Panagiotidis, N. Tsipas, T. Saridou, A. Veglis, A participatory journalism management platform: design, implementation and evaluation, Soc. Sci. 9 (2) (2020) 21.

[45] J.M. Pérez, F.M. Luque, D. Zayat, M. Kondratzky, A. Moro, P.S. Serrati, J. Zajac, P. Miguel, N. Debandi, A. Gravano, V. Cotik, Assessing the impact of contextual information in hate speech detection, IEEE Access (2023), https://doi.org/10.1109/ACCESS.2023.3258973.

[46] G. Petz, M. Karpowicz, H. Fürschuß, A. Auinger, V. Stříteský, A. Holzinger, Reprint of: computational approaches for mining user's opinions on the Web 2.0, Inf. Process. Manag. 51 (4) (2015) 510–519.

[47] R. Piryani, D. Madhavi, V.K. Singh, Analytical mapping of opinion mining and sentiment analysis research during 2000–2015, Inf. Process. Manag. 53 (1) (2017) 122–150.

[48] F.M. Plaza-del-Arco, M.D. Molina-González, L.A. Ureña-López, M.T. Martín-Valdivia, Comparing pre-trained language models for Spanish hate speech detection, Expert Syst. Appl. 166 (2021), 114120.

[49] A. Reichelmann, J. Hawdon, M. Costello, J. Ryana, C. Blayac, V. Llorent, A. Oksanen, P. Räsänen, I. Zych, Hate knows no boundaries: online hate in six nations, Deviant Behav. (2020) 1–12, https://doi.org/10.1080/01639625.2020.1722337.

[50] B. Rieder, Y. Skop, The fabrics of machine moderation: studying the technical, normative, and organizational structure of Perspective API, Big Data & Society 8 (2) (2021), 20539517211046181.

[51] J. Risch, R. Krestel, Delete or not delete? Semi-automatic comment moderation for the newsroom, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, ACL, New Mexico, 2018, pp. 166–176.

[52] S.T. Roberts, Behind the Screen, Yale University Press, 2019.

[53] P. Sánchez-Holgado, C. Arcila-Calderón, Supervised sentiment analysis of science topics: developing a training set of tweets in Spanish, J. Inf. Technol. Res. 13 (2020) 80–94.

[54] A. Seffah, M. Donyaee, R.B. Kline, H.K. Padda, Usability measurement and metrics: a consolidated model, Software Qual. J. 14 (2) (2006) 159–178.

[55] L. Shang, Y. Zhang, Y. Zha, Y. Chen, C. Youn, D. Wang, AOMD: an Analogy-Aware Approach to Offensive Meme Detection on Social Media, 2021 *arXiv preprint arXiv:2106.11229*.

[56] R.S. Solomon, P.Y.K.L. Srinivas, A. Das, B. Gamback, T. Chakraborty, Understanding the psycho-sociological facets of homophily in social network communities, IEEE Comput. Intell. Mag. 14 (2) (2019) 28–40.

[57] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, A. Kappas, Sentiment strength detection in short informal text, J. Am. Soc. Inf. Sci. Technol. 61 (2010) 2544–2558.

[58] L.J. Thun, P.L. Teh, C.B. Cheng, CyberAid: are your children safe from cyberbullying? J. King Saud Univ. Comp. Inform. Sci. 34 (7) (2022) 4099–4108.

[59] A. Veglis, Moderation techniques for social media content, in: Proccedings of the International Conference on Social Computing and Social Media (SCSM 2014), vol. 8531, Springer, Cham, 2014, pp. 137–148, https://doi.org/10.1007/978-3-319-07632-4_13.

[60] L. Vrysis, N. Tsipas, C. Dimoulas, G. Papanikolaou, Crowdsourcing audio semantics by means of hybrid bimodal segmentation with hierarchical classification, J. Audio Eng. Soc. 64 (12) (2016) 1042–1054.

[61] L. Vrysis, N. Vryzas, R. Kotsakis, T. Saridou, M. Matsiola, A. Veglis, C. Arcila-Calderón, C. Dimoulas, A web interface for analyzing hate speech, Future Internet 13 (3) (2021) 80.

[62] N. Vryzas, M. Matsiola, R. Kotsakis, C. Dimoulas, G. Kalliris, Subjective evaluation of a speech emotion recognition interaction framework, in: Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion, 2018, pp. 1–7.

[63] S. Wang, Moderating uncivil user comments by humans or machines? The effects of moderation agent on perceptions of bias and credibility in news content, Digital Journalism 9 (1) (2020) 64–83, https://doi.org/10.1080/21670811.2020.1851279.

[64] M. Weber, C. Viehmann, M. Ziegele, C. Schemer, Online hate does not stay online–How implicit and explicit attitudes mediate the effect of civil negativity and hate in user comments on prosocial behavior, Comput. Hum. Behav. 104 (2020), 106192.

[65] S. Weingartner, Digital Omnivores? How Digital Media Reinforce Social Inequalities in Cultural Consumption, New Media & Society, 2020, https://doi.org/10.1177/1461444820957635.

[66] H.C. Yang, H.W. Hsiao, C.H. Lee, Multilingual document mining and navigation using self-organizing maps, Inf. Process. Manag. 47 (5) (2011) 647–666.

[67] L. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: a survey, Wiley Interdisc. Reviews: Data Min. Knowl. Discov. 8 (4) (2018) e1253.

[68] H. Zhao, Z. Liu, X. Yao, Q. Yang, A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach, Inf. Process. Manag. 58 (5) (2021), 102656.

[69] M. Ziegele, P. Jost, M. Bormann, D. Heinbach, Journalistic counter-voices in comment sections: patterns, determinants, and potential consequences of interactive moderation of uncivil user comments, Studies in Communicat. Media 7 (4) (2018) 525–554, https://doi.org/10.5771/2192-4007-2018-4-525.