Complex & Intelligent Systems (2022) 8:4663–4678
https://doi.org/10.1007/s40747-021-00608-2

ORIGINAL ARTICLE

ETHOS: a multi-label hate speech detection dataset

Ioannis Mollas[1] · Zoe Chrysopoulou[1] · Stamatis Karlos[1] · Grigorios Tsoumakas[1]

Received: 1 July 2021 / Accepted: 3 December 2021 / Published online: 4 January 2022

Abstract
Online hate speech is a recent problem in our society that is rising at a steady pace by leveraging the vulnerabilities of the corresponding regimes that characterise most social media platforms. This phenomenon is primarily fostered by offensive comments, either during user interaction or in the form of a posted multimedia context. Nowadays, giant corporations own platforms where millions of users log in every day, and protection from exposure to similar phenomena appears to be necessary to comply with the corresponding legislation and maintain a high level of service quality. A robust and reliable system for detecting and preventing the uploading of relevant content will have a significant impact on our digitally interconnected society. Several aspects of our daily lives are undeniably linked to our social profiles, making us vulnerable to abusive behaviours. As a result, the lack of accurate hate speech detection mechanisms would severely degrade the overall user experience, although its erroneous operation would pose many ethical concerns. In this paper, we present 'ETHOS' (multi-labEl haTe speecH detectiOn dataSet), a textual dataset with two variants: binary and multi-label, based on YouTube and Reddit comments validated using the Figure-Eight crowdsourcing platform. Furthermore, we present the annotation protocol used to create this dataset: an active sampling procedure for balancing our data in relation to the various aspects defined. Our key assumption is that, even gaining a small amount of labelled data from such a time-consuming process, we can guarantee hate speech occurrences in the examined material.

Keywords Hate speech · Dataset · Machine learning · Multi-label · Classification · Active learning

Mathematics Subject Classification I.2.6 · I.2.7 · I.5.4 · H.2.4

Introduction

Hate speech (HS) is a form of insulting public speech directed at specific individuals or groups of people on the basis of characteristics, such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity[1].

This phenomenon is manifested either verbally or physically (e.g., speech, text, and gestures), promoting the emergence of racism and ethnocentrism. Because of the social costs arising out of HS, several countries consider it an illegal act, particularly when violence or hatred is encouraged [10]. Although a fundamental human right, freedom of speech, it is in conflict with laws that protect people from HS. Therefore, almost every country has responded by drawing up corresponding legal frameworks, while research which is related to mechanisms that try to remedy such phenomena has recently been done by the Data Mining and Machine Learning (ML) research communities [21].

Another important issue is that the occurrence of HS phenomena is emerging in the social media ecosystem, distorting their initial ambition of favouring communication between their corresponding members independently of geographical restrictions and enriching similar activities [48]. The anonymity of social media is the primary explanation for the growth of such phenomena, as is the deliberate avoidance of subsequent legislation. As a result, large corporations, such as Google and Facebook, are obligated to remove such violent

[1] https://en.wikipedia.org/wiki/Hate_speech.

(co)winning CrowdFlower's AI for Everyone Challenge for Q4 of 2017: https://prn.to/2KVWubz.

✉ Ioannis Mollas
iamollas@csd.auth.gr

Zoe Chrysopoulou
zoichrys@csd.auth.gr

Stamatis Karlos
stkarlos@csd.auth.gr

Grigorios Tsoumakas
greg@csd.auth.gr

[1] Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

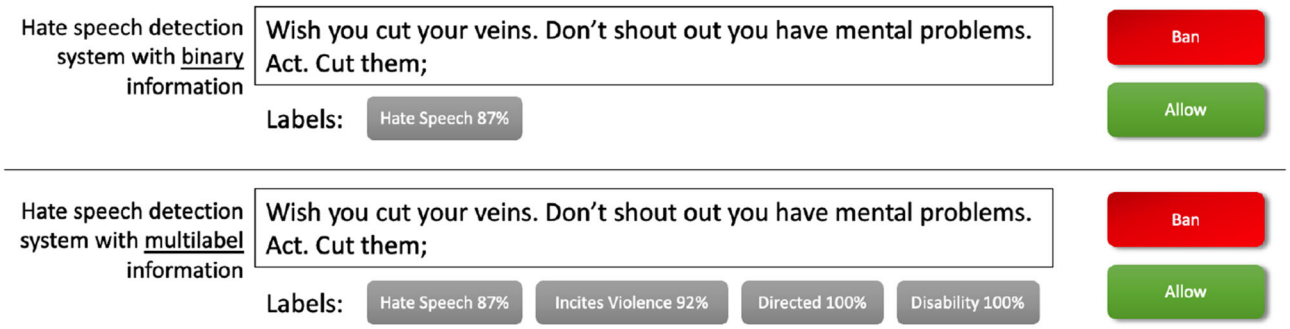

Fig. 1 A realistic example of informing a human reviewer about an investigated comment on binary (top) and multi-label (bottom) level

content off their platforms. Therefore, Artificial Intelligence (AI) methodologies are employed to detect (semi-) automatically HS in real time, or even to prevent users from publishing similar content with appropriate warnings or bans.

The solution of quarantining in an online fashion has recently been demonstrated, trying to smooth the censorship and the possible harmful consequences of HS attacks [53], while learning from short-text segments is blooming in the last years [51]. Two of the most important features accompanying the short-text segments, sparseness, and the presence of noise [50], settle HS detection, a difficult task for the creation of fully automated solutions. Whereas problems of scalability arise when large quantities of data are simply collected without pre-processing or filtering. These points are of primary importance to this work.

To achieve high performance in real-world tasks, AI methodologies require balanced, accurate, and unbiased datasets. This requirement, however, is rarely met without applying proper annotation stages [6]. This is the direction in which our work aims to make a significant contribution, motivated by the HS use case, providing also a generic-based protocol that could be extended to a wide variety of learning tasks. To be more precise, the relevant literature currently contains numerous manually created HS datasets [58,61]. However, since the majority of them were not carefully collected during the corresponding sampling stages, they are essentially large sets of annotated samples on which undesirable phenomena occur frequently. Specifically, highly imbalanced classes or redundant information prevent the subsequent implemented learning models from effectively harnessing the underlying patterns.

Moreover, by sampling the regions of feature space which express only a restricted level of uncertainty when unlabelled data are queried may settle the learning strategy myopic. All these phenomena violate the previously specified desired requirements resulting in solutions with low variance and/or high bias [55]. Furthermore, most of them are concerned with binary or multi-class classification tasks, while overlooking the more practical case of multi-label classification (MLL). Label dependencies and the semantic overlap that occurs on MLL cannot be ignored when protection from hateful comments is the main task. Since an online comment can fit to more than one defined label at the same time, rather than being limited to just one outcome, investigation of the latter scenario appears to be more effective (see Fig. 1). This aspect is also studied here, because the difficulties described previously are enforced under the MLL scenario.

A simple application that uses the MLL schema provided by the proposed HS dataset could be an assistance system for human staff reviewing comments on social media platforms. This would make it easier for the reviewer(s) to decide if the message contains HS content by providing more insights. For example, if a comment is presented as targeting people with disabilities, directed at a person, and encourages violence, it will be more helpful for the reader to conclude and condemn it for containing HS rather than being presented with a single label (e.g., 'may contain HS':{'yes','no'}). In terms of the ethical issues that emerge in the case of HS, it appears that a proper manipulation protocol is required for preventing possible defects. Such protocols have addressed wider or more focused research topics, such as news articles, although similar directions have recently been explored in the field of HS detection [42].

In this paper, we present the process of creating a dataset with two variations, a balanced binary and a multi-label one, with a step-by-step narrative, to avoid the consequences that typically occur in attempts with data that depend on social media platforms, and to increase the likelihood of mining more informative instances. Although the design of the proposed protocol can fit with any target domain indisputably, we are currently focusing on addressing the HS scenario and provide some insightful analysis of this use case. In this attempt, an existing dataset mined from popular social media platforms has been exploited, while a well-known crowdsourcing platform was used for validating the final result. The proposed annotation protocol's effects are discussed in detail and visualised using explanatory methods. Following that, a series of experiments are being conducted to determine the baseline performance of this particular dataset using state-of-the-art (SOTA) techniques. From traditional ML

algorithms and ensemble models to neural networks (NNs) with and without embeddings (*emb*) information, binary and multi-label experiments have been performed, inspired primarily by other similar approaches to presenting research datasets [2,7,30]. Experiments using Transformers (BERT and DistilBERT) were also included, because they appear to be extremely promising in many text-related machine learning tasks, with evidence of outstanding performance in hate speech detection, as well [27]. Despite the limited size of the investigated dataset, its careful design during the active sampling stage and the consistency of the included samples were proven beneficial based on our results.

Our ultimate ambition, by describing the total procedure and providing the corresponding dataset, is to foster any interested researchers/businesses to take into consideration an approach that attempts to transform the existing insulting environment of social media into a non-hate, inclusive online society. Adoption of the proposed annotation protocol into different scientific fields could prove quite beneficial, especially when the knowledge acquired by oracles during annotation may be ambiguous. The assets also gained by examining the HS problem through a multi-label view help us clarify the harasser's actual motivations and lead to more targeted comments when dedicated platforms try to inform the corresponding victims [10]. And, of course, the insights gained through such protocols could enhance the ability of ML learners to generalise when applied to different datasets that contain similar classification categories, despite the limited size of the proposed dataset over which they are trained. The proposed strategy of actively creating a balanced dataset, preserving the informativeness of each class and minimising the redundancy of the included instances, constitutes the key asset of our protocol. Our in-depth experiments support our hypotheses, particularly regarding the most difficult classes to detect.

The rest of this paper is structured as follows: Sect. 2 includes several well-documented attempts to address the HS problem using samples gathered from related sources. The proposed annotation protocol is defined in Sect. 3, followed by some extended single/multi-label classification experiments in Sect. 4, which demonstrate the discriminating ability of several algorithms under consideration. Sect. 5 presents a few studies with a variation of the original dataset and two additional datasets. Finally, Sect. 6 discusses the more crucial assets of the proposed dataset, and the annotation protocol, also regarding the recorded experiments, reporting later some remarkable future points that could be further investigated.

Related datasets

In this section, we present datasets related to HS, along with their formulation, as well as some useful information about their structure and/or the manner under which their composition took place. This section also describes the Hatebusters' data that we utilise as a seed data through the proposed protocol to produce the final structure of data, named *ETHOS* (multi-labEl haTe speecH detectiOn dataSet). Finally, a few literature gaps are presented in the last paragraph of this section.

A collection of 16,914 hate speech tweets was introduced in a study of how different features improve the identification of users that use analogous language online [58]. Out of the total number of messages, 3383, 1972, and 11,559 concerned sexism, racism, and did not include HS, respectively, while were sent by 613, 9, and 614 users. The corpus was generated by a manual tweet search, containing popular slurs and terms related to sexual, religious, gender, and ethnic minorities to include samples that are not offensive regardless of the inclusion of such words. The main drawback here is the access to the text of the tweets only through the public Twitter API.

Another dataset (D1) [7] contains 24,783 tweets, manually classified as HS (1,430), offensive but not HS (19,190), and neither hate nor offensive speech (4163) by Figure-Eight's[2] members. The data were gathered again via the Twitter API, filtering tweets containing HS words submitted to Hatebase.org. The outcome was a sample of 33,548 instances, while 85.4 million tweets were collected from the accounts of all users. A random sample of this collection led to the final dataset. Nevertheless, this dataset lacks diversity in terms of HS content. For example, the gender-based HS tweets are biased towards women, while the greatest number of them contains ethnicity content.

Research focusing on the identification of misogynistic language on Twitter uses a dataset called Automatic Misogyny Identification (AMI) [12] with 4000 annotated comments and binary labels. Apart from this labelling mode, every comment is defined by two extra fields. The first one concerns the type of misogynistic behaviour: stereotype, dominance, derailing, sexual harassment, discredit, or none (if the tweet is not misogynous). The second one concerns the subject of the misogynistic tweet: active, when it attacks a specific target (individual), passive, when it denotes potential receivers (generic), and again none, if there is no misogyny in the tweet.

The largest online community of white nationalists, called Stormfront, was used to form another dataset [8]. The content in this forum revolves around discussions of race, with various degrees of offensiveness, included. The annotation of the samples is at the sentence level, which is a technique that keeps the smallest unit containing hate speech and reduces noise. The dataset contains 10,568 sentences that are classified as HS (1119 comments) or not (8537 comments), as well as two supplementary classes, *relation* for sentences

[2] Formerly Crowdflower and latterly Appen: https://appen.com/figure-eight-is-now-appen/.

that express HS only when related to each other and *skip* for sentences which are not in English or do not contain any information as to be accordingly classified. Furthermore, information like the post-identifier and the sentence's position in the post, a user identifier, and a sub-forum identifier, as well as the number of previous posts the annotator had to read before deciding over the sentence's category are also recorded. The samples were picked randomly from 22 subforums covering diverse topics and nationalities.

A dataset introduced by Fox News [16] consists of 1528 Fox News users' comments (435 hateful), which were acquired from 10 discussion threads of 10 widely read Fox News articles published during August 2016. Context information is considered extremely important, so details such as the screen name of the user, all the comments in the same thread and the original article, are also included.

A recent multi-lingual work (D2) [34], a trilingual (English, French, and Arabic) dataset with tweets, was created attempting to mine similar expressions of 15 common phrases over these languages, focused on different sources of obscene phrases (e.g., more sensitive topic-based discussions based on locality criteria). After tackling some linguistic challenges per separate language, and a strict rule set that was posed to human annotators from the Amazon Mechanical Turk platform to ensure trustworthy feedback, a pilot test set was provided. Having gathered the necessary evaluations, another one reconstruction of the label set was applied, before the final formulation of 5647 English, 4014 French, and 3353 Arabic tweets was reached, annotated over 5 separate tasks. Apart from the binary directness of each tweet that was tackled better by single-task language models, the rest 4 classification tasks, which included at least 5 label gradations, were clearly boosted via multitask single/multi-language or single/multi-lingual models.

The issue of cyberbullying has been recently investigated too, where the skewed distribution of positive and negative comments was tackled by tuning a cost-sensitive linear SVM learner over various combinations of joined feature spaces and obtaining similar performance on both English and Dutch corpus [54]. Additionally, an investigation of recognising the role of each participant during such phenomena took place, while a qualitative analysis raised the difficulty of reducing misclassification scores when irony exists in offensive comments.

A small collection of 454 YouTube comments annotated as HS (120) or not (334) was introduced by the creators of the Hatebusters platform [3], which aims to establish an online inclusive community of volunteers actively reporting illegal HS content on YouTube. This dataset, through semi-supervised learning, was evolving in the Hatebusters platform, improving the predictivity of the ML models. However, this unpremeditated expansion of the dataset led to a more redundant variant of its original form. We use the initial collection of Hatebusters' data as a seed to the protocol that we propose in the following section.

There is clearly a lot of work and numerous publicly available datasets for hate speech identification. Nevertheless, the majority of these works fail to address a few critical issues. The first is related to data imbalance, and by that we mean not just numerical imbalance across classes, but also semantic imbalance between samples of the same class. Another issue that many works overlook is the heterogeneity of data sources. While most works focus on and gather data from platforms such as Twitter and Facebook, comments from platforms such as YouTube and Reddit, which include a significant amount of hate speech content, remain underrepresented.

ETHOS dataset creation

To overcome the key weaknesses of the existing collections of HS instances, we introduce a small, yet fairly, informative dataset, ETHOS, that does not suffer from issues such as imbalanced or biased labels (e.g., gender), produced appropriately following the proposed protocol. Considering the aforementioned popular approaches of mining similar datasets for tackling with HS problem, we assume that an appropriate pre-process of initially collected data could improve in general their overall utilisation under ML or AI products, improving the total fitness of data quality, blending data mining techniques related with the field of Active Learning [44], such as query strategy and crowdsourcing platforms. The overview of the proposed annotation protocol is visualised through a flowchart in Fig. 2. The finally obtained dataset is the outcome of a three-stage process, which we describe shortly in the current section.

Initial dataset creation and manual annotation

The first three procedures, mentioned as "Platform Selection and Data Collection", "Data Prediction", and "Manual Data Annotation", could be seen as the initial stage (Stage 1) which is executed until a stopping criterion is satisfied regarding the cardinality of the collected instances, based on the original available HS dataset which operates as the input. This stage works like a "stream", specifically for groups of comments that we have already collected, annotating their weak labels' predictions through a predefined ML classifier, before an active selection and manually annotation takes place over some unlabelled (U) mined examples.

Platform selection and data collection

To create this dataset (D), initially $D = \emptyset$, a data collection protocol has been designed. We chose the platforms

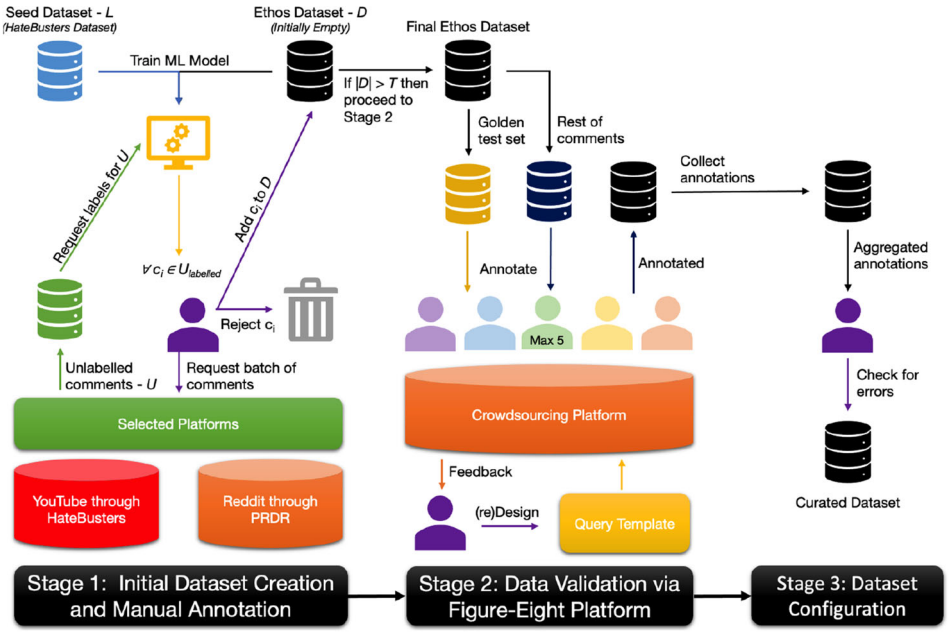

Fig. 2 Dataset creation stages flowchart

of Hatebusters[3] and Reddit through the Public Reddit Data Repository[4] to collect our data. Hatebusters platform collects new data daily via the YouTube Data v3 API. After these new data have been collected, the Hatebusters platform performs the classification process. The locally retained pre-trained ML model predicts the class of each comment, exporting a 'hate' score. Currently, this model is a Support Vector Machine (SVM) [56] model with a linear kernel embedded with the well-known vectorization technique of the term frequency-inverse document frequency (TF-IDF). Instead of transforming the output of the SVM learner to a confidence score, we kept its inherent property to compute the distance from the decision boundary. Through this, lower time overheads and more faithful decisions are drawn.

After granting access to Hatebusters' SQL database, based on the input data, this first part was to query the Hatebusters' database for comments already annotated by the corresponding users, without spending any monetisation resources. These comments were deemed to be accurate, and they were the first group of comments to be manually annotated. The second part concerns the enrichment of the gathered comments, by querying Hatebusters' database with a specific frequency (e.g., daily) for a time period—in our case, this was equal to 2 months—with various queries. Based on the data obtained each previous day, the applied query strategy had been updated concerning only them. For example, when we received a sufficient amount for all categories of HS, except for one category, the queries in the Hatebusters' database were updated to make comments specific to the residual category. Later on, we will show the categories and the number of comments we have received.

Regarding the Reddit platform, the data collection process was based on a public Reddit data archive, which provides batches of files regarding Reddit comments on a monthly basis. The files of this directory were processed through a JSON crawler for selecting comments from specific subreddits for particular time periods. The discovery of subreddits incorporating different HS contents has been investigated[5,6], we distinguished the next entities:

- **Incels**, this subreddit became known as a place where men blamed women for their unintended celibacy, often promoting rape or other abuse. Those posts had a misogynistic and sometimes racist content.
- **TheRedPill**, which is devoted to the rights of men, containing misogynous material.
- **The_Donald**, a subreddit where the participants create discussions and memes supportive of U.S. President Donald Trump. This channel has been described as hosting conspiracy theories and racist, misogynous, Islamophobic, and antisemitic content.
- **RoastMe**, in this subreddit, Reddit users can ask their followers to 'roast' (insult) them.

While some of these subreddits were suspended and shut down by Reddit at the end of 2017 due to their context, it was possible to access comments from these subreddits by selecting files from the archive for October 2017 and earlier.

[3] https://hatebusters.org.

[4] https://files.pushshift.io/reddit/comments/.

[5] https://en.wikipedia.org/wiki/R/The_Donald.

[6] https://en.wikipedia.org/wiki/Incel.

Following the finalisation of the platforms' selection, data are being collected from both of them in batches. As a result, in each iteration of Stage 1, a user-defined number of unlabelled comments U^{current}, or U^c are extracted equally from those sources.

Data prediction

The next process of Stage 1 is the "Data Prediction". For each batch of comments extracted from the first part, the assignment of some useful labels to the available unlabelled set (U^c) is triggered through an ML model trained on an expanded version ($L \cup D$) of the Hatebusters' dataset (L) and the new data manually annotated on the following step (D). Per each iteration of the previous part, we were performing a grid search among a bunch of classification methods in the currently expanded dataset, obtaining the best algorithm through a typical tenfold-cross-validation process to be set as the annotator of the (U^c). Following training, the best algorithm assigns a probability between [0, 1], where 0 denotes strong confidence about the non-existence of hate speech, while 1 stands for the opposite, for each comment of U^c, resulting to $U^c_{labelled} = U^c_l = [(c_1, p_1), (c_2, p_2), \ldots, (c_n, p_n)]$, where each set of (c_i, p_i) represented the comment (c_i) and its prediction (p_i).

The selected bunch consisted of various ML models: SVMs, Random Forests (RF) , Logistic Regression (LR) , as well as simple or more complex architectures of Neural Networks (NNs). In addition to the classifier tuning, some TF-IDF vectorization techniques—with word or char n-grams (n from 1 to 13)—were also examined in this search.

Manual data annotation

By the end of the "Data Prediction" phase, the "Data Annotation" process is initiated. In the sense of active learning concept, a hybrid combination of query strategy has been employed to pick informative comments for manual annotation. The mentioned query strategy combines appropriately both concepts of Uncertainty Sampling and Maximum Relevance with predefined ranges of accepted confidence values based on the expected labels of the classifier we had trained [39].

More specifically, as depicted in Algorithm 1, we were annotating the comments within the [.4, .6] probability range, while we were examining few comments in the ranges [.0, .1]∪[.9, 1.0] to detect any major misclassification. Then, we examined if there were any comments in D that were similar to each comment (c_i) and had the same labels (l_i). If there was, the corresponding c_i comments were rejected. We did this to avoid creating a dataset with several similar sentences for each label. The latter asset stems directly from the existence of the human factor, since the class probabilities that

Algorithm 1: Annotation and selection process of comments by annotator

Input: U^c_l - Automatically (by ML) labelled canditate comments, D—new annotated dataset
Output: A - Finally accepted comments

$A \leftarrow \emptyset$
for $p_i, c_i \in U^c_l$ **do**
  $r \leftarrow$ random 0 or 1
  if $p_i \in [0.4, 0.6]$ **then**
    $l_i \leftarrow annotated(c_i, p_i)$
    if *c_i is not similar to other instance of D with l_i labels* **then**
      $A \leftarrow A \cup [c_i, l_i]$
    end
  else if *($p_i \in [0, 0.1]$ or $p_i \in [0.9, 1]$) and $r = 1$* **then**
    $l_i \leftarrow annotated(c_i, p_i)$
    if *c_i is not similar to other instance with l_i labels* **then**
      $A \leftarrow A \cup [c_i, l_i]$
    end
  end
end
return A

are produced by any ML classifier just express its confidence independently of the underlying content. This kind of filtering is adequately addressed here by the human factor.

For example, the comments "I hate white people" and "I hate whites" are nearly identical, and only one would be added in D. This assisted us in developing labels with semantic balancing. This phenomenon of similar comments was especially noticeable in the hate speech categories related to gender and sexual orientation, where without this similarity criterion, comments about women would constitute the vast majority of instances of the gender label, while comments about gay people would dominate the sexual orientation label. Eventually, only comments with specific labels and content were added to the new dataset (D), preserving both the *balance of the labels* and the *diversity of the comments per label*.

At the end of this process, if the number of comments collected is not more than a targeted threshold (T)—in our case $T = 1000$—we update the D, and Stage 1 will be repeated to request new unlabelled comments. Otherwise, Stage 2 will be triggered. Despite the limited cardinality of the exported dataset, the adopted actively sampling process eliminates defects of redundancy, maintaining the informativeness of each label, and reducing at the same time overfitting phenomena. The issue of obtaining a myopic strategy is also eliminated, since different regions of uncertainty are explored [25]. The efficacy of such methods has been highly declared in the literature [26]. Therefore, an in-depth evaluation stage regarding several learning models has been conducted in Sect. 4. The use of Query-by-Committee, another popular active learning strategy, might introduce practical difficulties in practice, and thus was not investigated

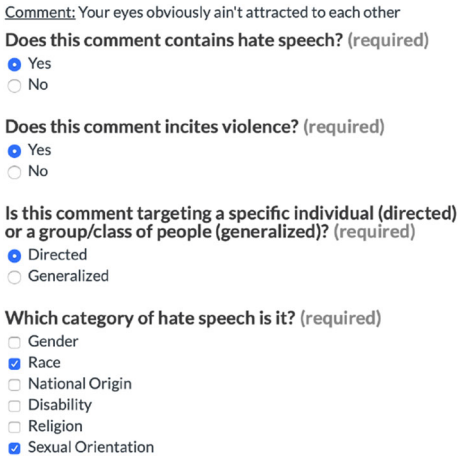

Fig. 3 Example of Figure-Eight's platform design

in that analysis. The reason for this choice is twofold: independent classifiers are needed for properly formatting such a committee, which constitutes a hard task under the shortage of large amounts of data, while the corresponding stage of hyperparameter tuning would induce more computational overhead.

Data validation via figure-eight platform

The second stage will begin when T—in our case 1000—comments have been collected. Moreover, Hatebusters' dataset is discarded, since it does not further contribute to our protocol. After a number of different experiments on the Figure-Eight platform, we settled on the next process. First, given a specific comment, we ask the contributors to identify whether that comment *contains HS or not*. In a positive scenario, we raise 3 more questions: whether the comment *incites violence*, defining violence as "the use of physical force to injure, abuse, damage, or destroy", and whether the comment includes *directed* or *generalised* HS. The case of targeting a single person or a small group of people is defined as directed HS, whereas the case of targeting a class/large group of people is described as generalised HS. Finally, we ask the contributors to pick *one* or *more* from the following *HS categories*, which, according to their opinion, better reflect(s) the content of the comments. The categories of HS concern gender, race, national origin, disability, religion, and sexual orientation.

Through Fig. 3, an example of the designed query template on the crowdsourcing platform is presented, depicting the types of questions and potential responses. With respect to the proposed template, we introduced a set of 40 questions for validating the trustworthiness of human annotators, as well as a test run of 100 unlabelled comments, with 5 different contributors determining the labels for each one of these. The annotators were frequently reminded that there was the possibility of explicit material in the comments through the crowdsourcing platform. The feedback we collected helped us to finalise some design and instruction issues. Most importantly, we have fixed some naive and few serious mistakes on the validation questions, such as the following:

[naive mistake]: For the comment "J**s are lower class pigs." we had 'Yes' on the question 'Does this comment incite violence?', while we gave as an explanation "It spreads hatred, but not violence".

[serious mistake]: We misclassified the comment "Q****s are an abomination and need to be helped to go straight to H**l!" to the wrong category. By receiving the feedback, we fixed it.

Then, we executed the task for the whole D, collecting in total 5360 judgements. Almost every comment was therefore annotated by five different annotators. The level of expertise of the annotators was the third, on a scale of three levels. "The 3rd level annotators are the smallest group of the most experienced, most accurate, contributors" according to the Figure-Eight System. We also computed the Fleiss' kappa, a statistical measure for assessing the reliability of agreement of annotators, and we present the results in Table 1. A kappa value greater than 0.75 implies good agreement, while kappa values greater than 0.90 indicate perfect agreement [20].

Dataset configuration

The final stage regards dataset configuration. Taking as input the results from the Stage 2, the dataset takes its final form. First, the annotations of every comment are aggregated as described in the second paragraph of Sect. 3.4. Examining the aggregated annotated data one last time manually, we checked for any misclassification. A few errors occurred on some of the most disambiguous examples, assuring us about the quality of the annotators that participated. Although the Figure-Eight platform provides several attributes for informing suitably the human annotators, even stricter measures

Table 1 Reliability of annotators agreement per label

	Contains hate Speech	Violence	Directed vs generalised	Gender	Race	National origin	Disability	Sexual orientation	Religion
Fleiss' Kappa	0.814	0.865	0.854	0.904	0.931	0.917	0.977	0.954	0.963

should be taken into consideration when large-scale datasets are aimed to be obtained [18].

The use of representative test questions that follow a more realistic label distribution than the uniform could be useful to the overall process. This might be improved further by incorporating an interactive procedure that alerts annotators to mislabelled samples and/or allows them to provide feedback when they disagree. Despite the inherent uncertainties introduced by the human factor, crowdsourcing is the sole viable technique for gathering the required information regarding the label space. This is true not only for large-scale datasets, but also for smaller cases [33].

Furthermore, given the semantic overlap of label space encountered during HS detection, the assumption of obtaining cheap labels is violated. Given the idiomatic expressions and highly unstructured nature of the comments posted on social media platforms, this becomes especially clear when examined in a multi-label fashion. To address this, additional human supervision, as stated at this stage, is required, while the active sampling process, which aims to create a balanced dataset, is clearly justified.

ETHOS dataset overview

Two datasets[7] were the product of the above operation. "Ethos_Binary.csv", the first one, includes 998 comments and a label on the presence or absence of hate speech content ('*isHate*'). The second file, called "Ethos_Multi_Label.csv", includes 433 hate speech messages along with the following 8 labels: ('*violence*', '*directed_vs _generalised*', '*gender*', '*race*', '*national_origin*', '*disability*', '*sexual_orientation*', '*religion*').

For every comment c_i, a number of annotators, N_i, voted for the labels that we set. The label '*isHate*' was the result of summing up the positive votes $P_{1,i}$ of the contributors, divided by N_i, so its values are within the range of [0, 1]. We measured the '*violence*' label by summarising the positive votes of the contributors $P_{2,i}$ to the question: "Does this comment incite violence?", which was divided by $P_{1,i}$ to be normalised to [0, 1]. Likewise, the value of the label '*directed_vs_generalised*' was determined by summarising the annotators replied 'directed' $P_{3,i}$ to the question, "Is this comment targeting a specific individual (directed) or a group/class of people (generalised)?", divided by $P_{1,i}$. Finally, we accumulated the votes of the N_i contributors for each of the six hate speech categories, and dividing them by $P_{1,i}$, we obtained six independent labels.

This dataset achieves to create balanced labels. In particular, it maintains balance between the two classes of 'isHate' label (55.61% comments without hate speech and 44.39% comments with hate speech content), almost perfect balance between the 6 labels of hate speech categories, with 19.41% for gender, 17.16% for race, 16.70% for national origin, 11.96% for disability, 18.28% for religion and 16.48% for sexual orientation. Additionally, our dataset keeps a fair ratio between the rest of the labels, 32.28% and 67.72% for violent and non-violent comments, respectively, and 31.83% and 68.71% for direct and generalised comments, respectively. All this information is also visible in Fig. 4. In Table 2, the balance between hate speech categories (last column) and their correlation with violence and directed/generalised labels is further portrayed.

[7] https://git.io/JwFh6.

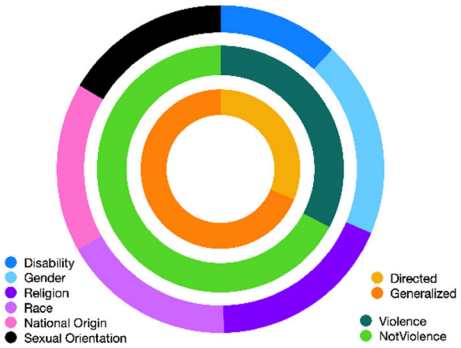

Fig. 4 Ratio of labels

Table 2 Correlation of HS categories with (not) violence (nV-V) and directed/generalised (D-G) labels

	V–D	nV–D	V–G	nV–G	
Gender	14	22	13	37	86
Race	4	13	12	47	76
National origin	5	11	18	40	74
Disability	12	15	8	18	53
Religion	11	8	24	38	81
Sexual orientation	11	15	11	36	73
	57	84	86	216	443

Dataset baseline evaluation

To evaluate ETHOS, after pre-processing the data, we used a variety of algorithms in binary/multi-label scope to present the baseline performance in this dataset. For the purpose of providing the unbiased performance of each algorithm, we performed nested-cross-validation [57] evaluation, using a variety of parameter setups, for every algorithm except NNs, where we applied tenfold cross-validation [17]. In addition, we binarise the values of each label, which are initially discrete in a range of [0,1], to the {0,1} classes using the rule "*If value* $\geq 0.5 \rightarrow 1$ *Else value* $\rightarrow 0$". More in-depth details follow next.

Table 3 Performance of selected models on binary HS classification

	F_1 Score	F_1 Hate	Accuracy	Precision	Sensitivity	Recall	Recall hate	Specificity
MultinomialNB	63.78	59.14	64.73	64.06	58.82	63.96	59.45	69.2
BernoulliNB	47.78	44.52	48.3	48.23	47.81	48.16	41.65	48.51
Logistic regression	66.5	64.35	66.94	66.94	68.78	67.07	60.46	65.36
SVM	66.07	63.77	66.43	66.47	68.08	66.7	59.96	65.32
Random forests	64.41	60.07	65.04	64.69	60.61	64.68	59.54	68.75
Gradient boosting	63.55	59.21	64.33	64.34	59.67	64.2	58.76	68.73
CNN+Attention + FT + GV	75.76	71.76	76.56	76.86	68.64	75.66	75.18	82.68
LSTM + FT + GV	75.24	72.24	75.95	76.57	72.11	75.53	72.36	78.95
FF + LSTM + CNN + FT + GV	75.49	72.08	76.15	76.29	70.88	75.52	73.28	80.16
BiLSTM + FT + GV	77.84	75.40	78.16	78.05	77.15	78.04	73.73	78.94
BERT	79.60	77.13	79.96	79.89	**77.87**	79.73	76.4	81.59
DistilBERT	**79.92**	**77.16**	**80.36**	**80.28**	76.47	**79.91**	**77.87**	**83.36**

The best performance is denoted in bold

Data preparation

The pre-processing methodology used in our case begins with lowercasing transformation, contraction transformations (available into the zip file), removal of punctuation marks, and stemming, and lemmatization via snowball stemmer [38] and WordNet lemmatizer [32].

Before we proceed to the experiments, we transform the pre-processed textual data into word vectors using TF-IDF and Text-to-Sequences processes. Particularly, for the former, several parameter tuples of (n_gram, max _features, stopwords existence) were examined, while on the latter, the corresponding number of maximum features was set at 50 k. Moreover, three pre-trained models that concern computation of *emb* were included: FastText (FT) [22], GloVe (GV) [35], Bert Language Model (BERT) [9], and the distilled version of BERT (DistilBERT) [43]. We should mention that the steps of stemming and lemmatization were skipped in the Text-to-Sequence experiments.

Binary classification

A lot of applications are investigating the problem of HS detection through a binary scope. It is therefore necessary to present the performance of SOTA algorithms on such a version of this dataset.

Thus, we used the following algorithms for our experiments in this stage: Multinomial and Bernoulli variations of Naive Bayes (MNB and BNB, respectively) [31], LR, SVMs, RF, and Gradient Boosting (Grad) [13]. Moreover, we used four different NN architectures, as other similar works attempt [36]. The first one utilises convolutional NNs (CNNs) [15] with an attention [4] layer. A single LSTM-based NN constitutes the second architecture. The third model is an NN with multiple parallel layers, which contain CNNs, LSTMs, and FeedForward layers (FFs). The last architecture consists of Bidirectional LSTMs (BiLSTMs). We combined these NNs with FT and GV. Finally, we used BERT and DistilBERT, which were fine-tuned in our classification task. Such architectures have met great acceptance in the related ML community [37,59].

We chose accuracy and precision, recall, and F_1 score with macro-indication, and the confusion matrix as metrics. Furthermore, we calculate specificity TN/N and sensitivity TP/P. However, in applications like HS monitoring where human interference is essential to ensure that users' rights are not abused on the grounds of incorrect HS charges, we must rely on metrics such as high recall and precision of HS category that we can guarantee to not overwhelm the human effort of checking redundant content. However, in such applications as HS reporting and handling, where human intervention is required to ensure that users' rights are not violated by false HS accusations, we should focus on metrics like high recall and F_1 score of the HS category, which ensure that human personnel checking redundant content are not overburdened.

The handling of textual data is a thoroughly researched task and has a dedicated category, NLP, which stands for natural language processing. We used common and widely accepted techniques to process them, as described previously. In Table 3, we are showcasing the results of the selected evaluation processes per each classifier. The best performance per metric is highlighted in bold format. The NNs seem to outperform the conventional ML techniques. It is worth mentioning that Bayesian learners had the lowest performance in terms of almost every metric, while tree-ensembles achieved similar performance between them, but lower compared to the SVMs and LR.

Between the examined NNs, those who achieved the highest performance using *emb* were the architectures using BiLSTMs. BiLSTMs + FT + GV achieved the highest recall on hate category, as well as high accuracy. Finally, BERT and DistilBERT outperformed every other model in any metric, using fine-tuning on the data, with DistilBERT performing slightly better than BERT, validating its superior performance on similar tasks [40].

Multi-label classification

Providing a dataset with multi-label information about HS, we are able to uncover new insights. HS is indeed an ML task that cannot be studied thoroughly just through the binary aspect. Indeed, it is a multidimensional task.

The algorithms handling MLL can be either problem transformation or adaptation techniques [52]. MLkNN [62] and MLARAM [5], as well as Binary Relevance (BR) and Classifier Chains (CC) [41] with base learners like LR, SVMs, and RF are utilised. We used FT *emb* for our NNs and designed models inspired by classic MLL systems, such as BR and CC. Specifically, NNBR is an NN containing BiLSTMs, an attention layer, two FFs, and an output layer with 8 outputs in a BR fashion. NNCC is inspired by the CC technique, but during its output, each label is given as input for the next label prediction.

In the evaluation of MLL systems, a very common measure is the Hamming loss (symmetric difference between the ground truth labels and the predicted ones). Furthermore, subset accuracy (symmetric similarity), as well as precision, recall, and F_1 score, are contained here (instance-based metrics). Moreover, some label-based metrics like B-macro and B-micro, where $B \in \{F_1, \text{Precision}, \text{Recall}\}$ were computed. We present our results in Table 4. The superior performance of neural-based approaches compared to classical ML models is observed. Specifically, NNBR achieves the highest score in 12 out of 13 metrics.

Dataset experimentation

After setting the baseline performance of ETHOS in multiple ML algorithms, in both binary and multi-label scope, this section aims at highlighting some interesting views and aspects of its usefulness over other learning tasks. First, we fulfil our experimental soundness by setting a fair comparison between a balanced subset and a random subset of ETHOS capturing useful insights under a 1-vs-1 evaluation stage. Second, we examine how the ETHOS dataset can generalise over separate HS datasets when it is applicable. Thus, we transfer its discriminative ability obtained by the proposed underlying representation through training proper ML models. These experiments have been conducted for two well-known datasets on binary (D1) [7] (2017) and multi-label (D2) [34] (2019) level, as described briefly in Sect. 2, commenting the produced results regarding the aspects that we had initially posed and providing accurate explanations about any mismatches over this attempt.

Balanced vs random comparison

Initially, we are going to experiment with the proposed dataset using just a few variations in the binary level. More precisely, we create two versions of ETHOS, one of which collects 75% of data at random (DRa), while the other collects 75% of data preserving the class balance (DBa), from a pool of 87.5%. The remainder of the data (DRe), which is 12.5%, will be used as test data. Two SVM models are then trained on DRa and DBa using a TF-IDF vectorizer and evaluated on the DRe. We are running this experiment ten times, shuffling our data appropriately. In addition, the two SVM models are evaluated on the D1 dataset, as well. Under this scenario, we are further investigating the learning capacity of the constructed ETHOS dataset comparing two different variants: a strictly balanced and a random one, while our evaluation protocol is consistent with maintaining the balancing property of the generated sub-optimal subsamples. The application of the trained learners into separate datasets may also confirm

Table 4 Performance of selected models on MLL HS

	Example			Macro			Micro			AP		Subset accuracy	Hamming loss
	F_1	P	R	F_1	P	R	F_1	P	R	Macro	Micro		
MLkNN	48.01	55.27	46.28	53.04	71.29	45.04	53.74	69.95	43.98	46.63	42.79	26.53	0.1566
MLARAM	18.47	21.44	17.69	6.06	3.78	16.25	18.71	21.44	18.27	20.79	21.55	7.15	0.2948
BR	48.59	57.69	45.30	52.49	79.74	42	56.76	79.37	44.37	47.66	47.04	26.28	0.1395
CC	56.51	62.49	56.54	59.24	69.08	56.22	58.23	63.44	53.99	49.74	44.07	31.4	0.1606
NNBR	**75.05**	**81.02**	**74.33**	**76.23**	**83.21**	**73.04**	**74.87**	79.27	**71.29**	**67.33**	**62.64**	**48.39**	**0.0993**
NNCC	47.66	57.34	44.06	51.25	73.36	42.40	55.47	**84.27**	41.70	50.02	47.36	26.61	0.1378

The best performance is denoted in bold
P Precision; *R* Recall, *AP* Average precision

Table 5 Comparison of SVM performance (metric ± std) trained on random and balanced subsets of ETHOS and tested on unknown data from the same source (DRe) and a different one (D1)

	DRe	D1	
Train on DRa	63.15 ± 3.93	**50.62** ± 1.10	Accuracy
Train on DBa	**67.99** ± 2.17	43.61 ± 12.39	
Train on DRa	64.19 ± 4.89	36.15 ± 1.05	F_1 weighted
Train on DBa	**69.06** ± 2.29	37.21 ± 8.25	

The best performance is denoted in bold

Table 6 Performance of SVM (metric ± std) on D1 per label

	D1	
Train on DRa	**66.53** ± 1.01	F_1 Non-HS
Train on DBa	54.48 ± 16.32	
Train on DRa	5.77 ± 19.94	F_1 HS
Train on DBa	**19.94** ± 3.34	

The best performance is denoted in bold

our assumptions about the efficacy of our strategy: the active selection of multi-label samples for constructing a balanced HS dataset.

The results are shown in Table 5, verifying that the performance of the SVM on the test set is higher when the dataset maintains a balance between classes. However, in terms of accuracy, a higher score is obtained by random datasets. We cannot conclude for the F_1 weighted performance of DRa and DBa on D1, as the wide standard deviation of the DBa makes it difficult. This result comes of course with an explanation: a defining characteristic of the D1 dataset concerns its imbalanced nature. This indicates that the SVM trained on random data is more biased towards the majority class. To investigate this, the weighted F_1 score per label is shown in Table 6.

As we previously assumed, the SVM model trained on DRa has a bias towards the majority class (No Hate) obtaining a better score than the SVM model trained on DBa. However, this is not the case for the minority class, which seems to be best predicted by the SVM trained on the DBa. In tasks such as hate speech identification, it would be more valuable to identify comments of hate speech more precisely. Consequently, a balanced dataset despite its limited cardinality may play a crucial role in tackling this phenomenon, verifying the assets of the proposed protocol.

Generalising on binary level

In an attempt to prove that a small but carefully collected dataset is of higher quality and more useful than larger datasets collected under unknown conditions, we will compare ETHOS to D1, a dataset 24 times larger. In this

Table 7 SVM model trained on ETHOS and predicting D1

	ETHOS	D1
Balanced accuracy	58.03	54.03
F_1 weighted	56.41	87.32
F_1 Non-HS	74.03	91.88
F_1 HS	33.21	12.85

Table 8 SVM model trained on D1 and predicting ETHOS

	D1	ETHOS
Balanced accuracy	50.90	53.33
F_1 weighted	42.67	92.31
F_1 Non-HS	72.66	97.10
F_1 HS	3.53	12.38

cross-validation experiment, we train an SVM model (with default parameters) on the ETHOS dataset and predict the D1 dataset, and vice versa. We have also computed the performance of SVMs on the D1 through nested cross-validation, resulting in 66.18% balanced accuracy, 68.77% F_1 weighted score, 96.97% F_1 on non-HS tweets, and 42.09% on HS tweets, revealing thus its optimal performance which also did not manage to get improved regarding the predictiveness of HS instances.

The results of each cross-validation training are shown in Tables 7 and 8. It is visible that both SVMs perform equally in both metrics. It could be expected that the SVM trained on D1, a larger dataset, would perform better than a smaller dataset, but the more sophisticated manner of collecting and annotating data in the case of ETHOS overcomes its limited cardinality offering similar predictive ability with a quite larger collection of instances.

It is peculiar that the two models do not predict the other's hate speech instances. Digging into that further, we can see that there are few problematic instances in D1. For example, the following sentence: *"realdonaldtrump he looks like reg memphis tn trash we got them everywhere"* does not contain hate speech content, rather than offensive. Moreover, the distribution of the hate instances to hate categories in D1 is non-uniform, favouring three categories: race (dark-skinned people), sexual orientation (homosexual people), and gender (women). The aforementioned conclusion was the product of applying the ETHOS Multi-labelled dataset, predicting 326—race, 257—sexuality, and 230—gender instances out of the 1430 hate speech tweets, as well as the product of a simple word frequency calculation, suggesting that there are 378—race (words: 'n****r', 'n***a', 'n****h'), 417—sexuality (words: 'f****t', 'f*g', 'g*y', 'q***r'), and 352—gender (words: 'b***h', 'c**t', 'h*e') instances.

Finally, it would be interesting to investigate the overall performance of an SVM model trained on a combination dataset of those two. After a tenfold cross-validation training the combined dataset achieved 55.27% balanced accuracy, 90.88% F_1 weighted score, 96.48% F_1 on Non-Hate Speech, and 18.84% F_1 on Hate Speech. The overall performance of the model increased, implying that combining datasets with different dynamics can lead to better models. To this aspect, one of the posed ambitions of our work seems to be satisfied, since its integration with the D1 dataset leads to improved learning behaviour.

Generalising on multi-label level

The dataset of ETHOS has two variants, a binary and a multi-labelled dataset. After experimenting with the binary version of it, we use the D2 dataset in this section to show the usefulness of ETHOS. D2 is a multi-lingual and multi-aspect hate speech dataset containing information for tweets such as hostility type, directness, target attribute, and category, as well as annotator's sentiment. However, there is no one-to-one mapping between these attributes and the attributes of ETHOS. For example, the type of hostility defines the sentiment of a tweet as abusive, hateful, offensive, disrespectful, fearful, and normal. We assign instances described as abusive, hateful, or fearful as violent, while others are described as non-violent. The mapping of the hostility directness to the ETHOS directed_vs_generalised label is straightforward. Finally, the mapping between the hate categories and the target attributes is almost the same, while the 'race' category is absent. However, by extracting information from the target group attribute, we assign tweets to the 'race' category when the target group is either 'African descent' or 'Asian'.

Training a neural network with BiLSTM layers using ETHOS multi-labelled dataset, we are predicting the labels of D2. In Table 9, the performance of the model on the D2 dataset per label is showcased. The model achieved to predict perfectly the 'sexual_orientation' label, decently the 'disability', 'national_origin' and 'gender' labels, but poorly the 'directed_vs_generalised', 'violence', 'race' and 'religion' labels. Specifically, on the 'religion' label, the model can identify if a tweet does not contain hate speech towards religion by 97.82%, but its performance is downgraded on the opposite case, achieving 27.31%. About the 'violence' label, the model fails to predict when a sentence incites violence, with 29.09%. The worst predicted label by the model is the 'directed_vs_generalised'. This means that the model cannot generalise well when a tweet is targeting a specific individual.

As it regards the 'race', due to the lack of information in the D2 about this label, it was expected to counter such a low performance. To be more convenient with this aspect, we depict some of the instances which had as groups 'Asian' or 'African descent', and our model did not categorise as race the following four:

> "well my parents like carikla ching chong guy in your college"
> "yay kelas ching chong today"
> "okay ching chong"
> "remember it was some ching chong hoe on here that was flexin on him years ago found out they was fuckin smh"

It seems the BiLSTM model has not encountered such examples. Indeed, ETHOS dataset does not contain any example with the phrase 'ching chong'. However, we should investigate the reversed situation as well, namely, the instances that did not have the race label, but the BiLSTM model assigned it erroneously. This misclassification occurred to 35 instances, while 26 of them contain hate speech targeting 'race'. We present here the most representative of them:

> "see the type of n****r you are hmph"
> "die n****r" and 20 similar
> "now yes this politically motivated terrorist is white and leftist" and 3 similar

Such issues are quite possible to occur because of mismatching between the separate collections of data. Enrichment of the source dataset, in our case the ETHOS, by a careful selection of instances that describe such cases could help our attempt. Therefore, the adoption of metric learning mechanisms may help us alleviate the hubness phenomenon which puts obstacles on recovering distinct classes [24].

Discussion

The provision of a new well-designed dataset to the public on a specific subject is always considered a significant contribution [19,47]. In this sense, our HS dataset, called ETHOS, collected from social media platforms, could be reused by the ML and AI communities. Alleviating redundant information through balancing the proposed dataset between fine-grained classes through a fine-tuned learner and an Active Learning scheme benefited us both from the aspect of less human-laborious effort and, of course, by scoring good learning rates despite the limited cardinality of our collected instances. Redundancy reduction has been shown to be quite beneficial for a variety of learning tasks. More specifically, the proposed protocol offers us a balanced dataset with a rich quality of included instances for both binary and multi-label HS problems. At the same time, our experimental procedure revealed that a proper balance has been achieved between

Table 9 The performance of the model trained on ETHOS predicting the labels of D2

	Violence	Directed vs generalised	Gender	Race	National origin	Disability	Religion	Sexual orientation
Accuracy	50.86	55.28	70.34	75.97	67.88	69.64	71.65	89.83
F_1 weighted	59.48	55.39	87.71	92.78	68.97	83.81	97.65	94.21
F_1 (negative)	72.50	59.36	92.94	94.61	74.89	91.06	98.51	96.50
F_1 (positive)	29.09	19.98	46.59	24.06	61.23	53.44	27.31	71.29

F_1 (negative): The label is not appearing in the instance; F_1 (positive): The label is appearing in the instance

the discriminative ability of the learners, both traditional and neural networks, and the computational resources consumed.

The issue of imbalanced data collection has also affected the performance of similar works, where the need for proper manipulation is clearly stated [18,33]. The solution of proactive learning has been applied in the latter approach, trying to match the expertise of each human annotator with the most appropriate unlabelled instances. Based on this, the negative effect of harmful annotations can be seriously avoided. This asset should be carefully explored and adopted by our side before enlargement of the current dataset takes place or new data collection attempts get started. We must emphasise once more that, despite the relatively small size of the ETHOS dataset, the human resources invested in adequate labelling cannot be overlooked (2 consecutive months of daily querying of the targeted databases, human annotation in 2 stages, input by a crowdsourcing process). Thus, besides the need for high-quality annotators, mining informative instances that retain the ability to discriminate between hate speech examples, both in binary and multi-label classification tasks, is of high importance. The conducted experiments verify our assumptions following our straightforward protocol, since the learning performance of various models is satisfactory, especially these based on embeddings. Simultaneously, a proof-of-concept of how to exploit the ETHOS dataset's learning capacity was provided, serving as a seed dataset for generalising to similar hate speech detection datasets.

Some promising directions of our work are mentioned here, trying to take advantage of its assets and the baselines that were posed. The main issue, the shortage of collected data, is a fact that depends on the limitations that occur during exploiting crowdsourcing platforms (e.g., restricted budget, users' traffic) and the further costs that are induced by the human-intensive stage of actively selecting instances that keep a balanced profile of the target dataset on a daily basis. Investigating the related literature, we have mined some clever ideas that tackle this limitation. Another constraint of the annotation protocol is the requirement for a single annotator in the first stage. This restriction derives from the fact that we intended the annotator to remember the comments approved in the dataset, not let extremely similar comments in, and aggregate semantically diverse comments for each category. However, in future work, we want to overcome this constraint using deep learning techniques to assess the similarity of incoming comments with approved comments and enable multiple annotations by the first stage of the protocol.

We record here the case where an annotation process has been designed using a game-based approach, motivating the human oracles to contribute to assigning sentiment labels to a variety of Twitter instances, surpassing the monetisation incentive [14]. Further enrichment of this dataset could also be carried out, integrating either multi-lingual resources for capturing even more hate speech occurrences, or applying data augmentation techniques [45]. From the perspective of the ML models that we used, pre-processing stages—such as feature selection mechanisms [49] or methods for creation of semantic features [46]—which are established in the realm of short-text input data, could improve the obtained results, and retain interpretability properties in specific cases.

In addition, the ETHOS can be combined with various similar HS datasets—as we stated here initially with two different data collections—for evaluation reasons. The development of hybrid weakly supervised HS detection models, merging semi-supervised and active learning strategies under common frameworks, alleviating human intervention based on decisions over the gathered unlabelled instances that come solely from the side of a robust learner [23,60], constitutes another very promising ambition. Online HS detection and prevention tools, such as Hatebusters among others, are highly favoured by such approaches. The impact of such detection tools could have been very beneficial in terms of enforcing social awareness and addressing effective ethical issues [1,10]. Furthermore, it would be interesting to investigate how our annotation protocol can be used to collect a multi-modal hate speech dataset [28], as well as how our collected balanced dataset can enhance tasks like multi-lingual hate speech detection [29].

Finally, the fact of examining ETHOS under the spectrum of multi-labelled nature appears favouring to reviewers on social media platforms, facilitating informative suggestions for HS comments regarding the level of violence, the target of comments, and the categories of HS that are present. However, this is not a multi-purpose HS detection dataset, as the mined comments are based on social media. This means that the corpus contains relatively small sentences. Thus, models

trained on this dataset may fail to detect HS in documents on a larger scale without segmentation. On the other hand, the general structure of the proposed protocol could be applied to a variety of learning tasks, especially on large databases, towards better predictions and less intensive annotation [11]. Last but not least, examination of alternative query sampling strategies that support inherent MLL could have proven quite beneficial regarding both the reduction of human effort and the enrichment of attempts like the proposed one [26].

Funding The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H. F. R. I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant" (Project Number: 514).

Code availability Experiments' code is available in the GitHub repository: https://git.io/JwFh6.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Alharthi DN, Regan AC (2020) Social engineering defense mechanisms: a taxonomy and a survey of employees' awareness level. In: Arai K, Kapoor S, Bhatia R (eds) Intelligent computing - proceedings of the 2020 computing conference, volume 1, SAI, London, UK, 16–17 July 2020, *Advances in Intelligent Systems and Computing*, vol. 1228, pp. 521–541. Springer (2020). https://doi.org/10.1007/978-3-030-52249-0_35
2. Almeida T, Hidalgo JMG, Silva TP (2013) Towards sms spam filtering: results under a new dataset. Int J Inform Secur Sci 2(1):1–18
3. Anagnostou A, Mollas I, Tsoumakas G (2018) Hatebusters: a web application for actively reporting youtube hate speech. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pp. 5796–5798. International Joint Conferences on Artificial Intelligence Organization, Stockholm, Sweden. https://doi.org/10.24963/ijcai.2018/841
4. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015, May 7-9, 2015, Conference Track Proceedings. San Diego, California, USA
5. Benites F, Sapozhnikova E (2015) Haram: a hierarchical aram neural network for large-scale text classification. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 847–854. IEEE Computer Society, USA. https://doi.org/10.1109/ICDMW.2015.14
6. Chen J, Mao J, Liu Y, Zhang M, Ma S (2019) Tiangong-st: a new dataset with large-scale refined real-world web search sessions. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, November 3-7, 2019 pp. 2485–2488. ACM, Beijing, China. https://doi.org/10.1145/3357384.3358158
7. Davidson T, Warmsley D, Macy M, Weber I (2017) Automated hate speech detection and the problem of offensive language. In: Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM '17, pp. 512–515. AAAI Press, Montreal, Canada
8. de Gibert O, Perez N, García-Pablos A, Cuadros M (2018) Hate speech dataset from a white supremacy forum. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). https://doi.org/10.18653/v1/w18-5102
9. Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (1), pp. 4171–4186. Association for Computational Linguistics
10. Dinakar K, Picard RW, Lieberman H (2015) Common sense reasoning for detection, prevention, and mitigation of cyberbullying (extended abstract). In: Yang Q, Wooldridge MJ (eds) Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25–31, 2015, pp. 4168–4172. AAAI Press. http://ijcai.org/Abstract/15/589
11. Dramé K, Mougin F, Diallo G (2016) Large scale biomedical texts classification: a knn and an esa-based approaches. J Biomed Semant 7:40. https://doi.org/10.1186/s13326-016-0073-1
12. Fersini E, Rosso P, Anzovino M (2018) Overview of the task on automatic misogyny identification at ibereval 2018. In: IberEval@SEPLN, pp. 214–228
13. Friedman J (1999) Stochastic gradient boosting. department of statistics. Tech. rep., Stanford University, Technical Report, San Francisco, CA
14. Furini M, Montangero M (2018) Sentiment analysis and twitter: a game proposal. Pers. Ubiquitous Comput. 22(4):771–785. https://doi.org/10.1007/s00779-018-1142-5
15. Gambäck B, Sikdar UK (2017) Using convolutional neural networks to classify hate-speech. In: Waseem Z, Chung WHK, Hovy D, Tetreault JR (eds) Proceedings of the First Workshop on Abusive Language Online, ALW@ACL 2017, Vancouver, BC, Canada, August 4, 2017, pp. 85–90. Association for Computational Linguistics. https://doi.org/10.18653/v1/w17-3013
16. Gao L, Huang R (2017) Detecting online hate speech using context aware models. In: RANLP
17. Geisser S (1993) Predictive inference, vol 55. CRC Press, Boca Raton
18. Haagsma H, Bos J, Nissim M (2020) MAGPIE: a large corpus of potentially idiomatic expressions. In: Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, Goggi S, Isahara HH, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S (eds) Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, pp. 279–287. European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.35/
19. Hoang T, Vo KD, Nejdl W (2018) W2E: a worldwide-event benchmark dataset for topic detection and tracking. In: Proceedings of the 27th ACM International Conference on Information and

Knowledge Management, CIKM 2018, Torino, Italy, October 22–26, 2018, pp. 1847–1850. ACM. https://doi.org/10.1145/3269206.3269309
20. Inc., M.: Kappa statistics for attribute agreement analysis. Available at https://support.minitab.com/en-us/minitab/18/help-and-how-to/quality-and-process-improvement/measurement-system-analysis/how-to/attribute-agreement-analysis/attribute-agreement-analysis/interpret-the-results/all-statistics-and-graphs/kappa-statistics/ (2021/04/17)
21. Jirotka M, Stahl BC (2020) The need for responsible technology. J Respons Technol 1: 100002. https://doi.org/10.1016/j.jrt.2020.100002. http://www.sciencedirect.com/science/article/pii/S2666659620300020
22. Joulin A, Grave E, Bojanowski P, Douze M, Jégou H, Mikolov T (2016) Fasttext.zip: compressing text classification models
23. Karlos S, Kanas VG, Aridas CK, Fazakis N, Kotsiantis S (2019) Combining active learning with self-train algorithm for classification of multimodal problems. In: IISA 2019, Patras, Greece, July 15-17, 2019, pp. 1–8. https://doi.org/10.1109/IISA.2019.8900724
24. Kim S, Kim D, Cho M, Kwak S (2020) Proxy anchor loss for deep metric learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pp. 3235–3244. IEEE. https://doi.org/10.1109/CVPR42600.2020.00330
25. Krempl G, Kottke D, Lemaire V (2015) Optimised probabilistic active learning (OPAL) - for fast, non-myopic, cost-sensitive active classification. Mach Learn 100(2–3):449–476. https://doi.org/10.1007/s10994-015-5504-1
26. Kumar P, Gupta A (2020) Active learning query strategies for classification, regression, and clustering: a survey. J Comput Sci Technol 35(4):913–945. https://doi.org/10.1007/s11390-020-9487-4
27. Kumari K, Singh JP (2020) Ai_ml_nit_patna @hasoc 2020: BERT models for hate speech identification in indo-european languages. In: Mehta P, Mandl T, Majumder P, Mitra M (eds) Working notes of FIRE 2020—forum for information retrieval evaluation, Hyderabad, India, December 16–20, 2020, *CEUR Workshop Proceedings*, vol. 2826, pp. 319–324. CEUR-WS.org. http://ceur-ws.org/Vol-2826/T2-29.pdf
28. Kumari K, Singh JP (2021) Identification of cyberbullying on multi-modal social media posts using genetic algorithm. Trans Emerg Telecommun Technol **32**(2). https://doi.org/10.1002/ett.3907
29. Kumari K, Singh JP (May 2020) Ai_ml_nit_patna @ TRAC - 2: Deep learning approach for multi-lingual aggression identification. In: Kumar R, Ojha AK, Lahiri B, Zampieri M, Malmasi S, Murdock V, Kadar D (eds) Proceedings of the second workshop on trolling, aggression and cyberbullying, TRAC@LREC 2020, Marseille, France, pp. 113–119. European Language Resources Association (ELRA) (2020). https://aclanthology.org/2020.trac-1.18/
30. Ljubešić N, Erjavec T, Fišer D (2018) Datasets of slovene and croatian moderated news comments. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pp. 124–131. Association for Computational Linguistics, Brussels, Belgium. https://doi.org/10.18653/v1/W18-5116. https://www.aclweb.org/anthology/W18-5116
31. McCallum A, Nigam K, et al. (1998) A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization, vol. 752, pp. 41–48. Citeseer
32. Miller GA (1995) Wordnet: a lexical database for english. Commun ACM 38(11):39–41
33. Nghiem M, Baylis P, Ananiadou S (2021) Paladin: an annotation tool based on active and proactive learning. In: Gkatzia D, Seddah D (eds) Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, EACL 2021, Online, April 19–23, 2021, pp. 238–243. Association for Computational Linguistics. https://www.aclweb.org/anthology/2021.eacl-demos.28/
34. Ousidhoum N, Lin Z, Zhang H, Song Y, Yeung D (2019) Multilingual and multi-aspect hate speech analysis. In: EMNLP-IJCNLP 2019, November 3–7, 2019, pp. 4674–4683. Association for Computational Linguistics, Hong Kong, China. https://doi.org/10.18653/v1/D19-1474
35. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. Doha, Qatar. http://www.aclweb.org/anthology/D14-1162
36. Pitenis Z, Zampieri M, Ranasinghe T (2020) Offensive language identification in greek. In: LREC, pp. 5113–5119. European Language Resources Association
37. Polignano M, Basile P, de Gemmis M, Semeraro G, Basile V (2019) Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In: Bernardi R, Navigli R, Semeraro G (eds) Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13–15, 2019, *CEUR Workshop Proceedings*, vol. 2481. CEUR-WS.org. http://ceur-ws.org/Vol-2481/paper57.pdf
38. Porter MF (2001) Snowball: A language for stemming algorithms. Published online. http://snowball.tartarus.org/texts/introduction.html. Accessed 11.03.2008, 15.00h
39. Pupo OGR, Altalhi AH, Ventura S (2018) Statistical comparisons of active learning strategies over multiple datasets. Knowl Based Syst 145:274–288. https://doi.org/10.1016/j.knosys.2018.01.033
40. Ranasinghe T, Zampieri M, Hettiarachchi H (2019) BRUMS at HASOC 2019: Deep learning models for multilingual hate speech and offensive language identification. In: Working Notes of FIRE 2019, December 12-15, 2019, *CEUR Workshop Proceedings*, vol. 2517, pp. 199–207. CEUR-WS.org, Kolkata, India. http://ceur-ws.org/Vol-2517/T3-3.pdf
41. Read J, Pfahringer B, Holmes G, Frank E (2009) Classifier chains for multi-label classification. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 254–269. Springer, Springer, Bled, Slovenia
42. Rosenthal S, Atanasova P, Karadzhov G, Zampieri M, Nakov, P (2021) SOLID: A large-scale semi-supervised dataset for offensive language identification. In: ACL/IJCNLP (Findings), *Findings of ACL*, vol. ACL/IJCNLP 2021, pp. 915–928. Association for Computational Linguistics
43. Sanh V, Debut L, Chaumond J, Wolf T (2019) Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In: NeurIPS EMC^2 Workshop
44. Sharma M, Zhuang D, Bilgic M (2015) Active learning with rationales for text classification. In: Mihalcea R, Chai JY, Sarkar A (eds) NAACL HLT 2015, Denver, Colorado, USA, May 31 - June 5, 2015, pp. 441–451. The Association for Computational Linguistics. https://doi.org/10.3115/v1/n15-1047
45. Shim H, Luca S, Lowet D, Vanrumste B (2020) Data augmentation and semi-supervised learning for deep neural networks-based text classifier. In: Hung C, Cerný T, Shin D, Bechini A (eds) SAC '20: The 35th ACM/SIGAPP Symposium on Applied Computing, online event, [Brno, Czech Republic], March 30 - April 3, 2020, pp. 1119–1126. ACM. https://doi.org/10.1145/3341105.3373992
46. Skrlj B, Martinc M, Kralj J, Lavrac N, Pollak S (2021) tax2vec: constructing interpretable features from taxonomies for short text classification. Comput Speech Lang 65:101–104. https://doi.org/10.1016/j.csl.2020.101104
47. Sun C, Asudeh A, Jagadish HV, Howe B, Stoyanovich J (2019) Mithralabel: flexible dataset nutritional labels for responsible data science. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019,

Beijing, China, November 3–7, pp. 2893–2896. ACM, Beijing, China (2019). https://doi.org/10.1145/3357384.3357853
48. Tang MJ, Chan ET (2020) Social media: influences and impacts on culture. In: Arai K, Kapoor S, Bhatia R (eds) Intelligent computing—proceedings of the 2020 computing conference, Volume 1, SAI 2020, London, UK, 16–17 July 2020, *Advances in Intelligent Systems and Computing*, vol. 1228, pp. 491–501. Springer. https://doi.org/10.1007/978-3-030-52249-0_33
49. Tommasel A, Godoy D (2018) A social-aware online short-text feature selection technique for social media. Inf Fus. 40:1–17. https://doi.org/10.1016/j.inffus.2017.05.003
50. Tommasel A, Godoy D (2019) Short-text learning in social media: a review. Knowl Eng Rev 34:e7. https://doi.org/10.1017/S0269888919000018
51. Tommasel A, Godoy D (2018) A social-aware online short-text feature selection technique for social media. Inform Fus 40:1–17 https://doi.org/10.1016/j.inffus.2017.05.003. http://www.sciencedirect.com/science/article/pii/S1566253516302354
52. Tsoumakas G, Katakis I (2007) Multi-label classification: an overview. Int J Data Warehous Min (IJDWM) 3(3):1–13
53. Ullmann S, Tomalin M (2020) Quarantining online hate speech: technical and ethical perspectives. Ethics Inf Technol 22(1):69–80. https://doi.org/10.1007/s10676-019-09516-z
54. Van Hee C, Jacobs G, Emmery C, Desmet B, Lefever E, Verhoeven B, De Pauw G, Daelemans W, Hoste V (2018) Automatic detection of cyberbullying in social media text. PLOS One 13(10). https://doi.org/10.1371/journal.pone.0203794
55. van Rosendaal J, Caselli T, Nissim M (2020) Lower bias, higher density abusive language datasets: a recipe. In: Monti J, Basile V, di Buono MP, Manna R, Pascucci A, Tonelli S (eds) Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language, ResTUP@LREC 2020, Marseille, France, May 2020, pp. 14–19. European Language Resources Association (ELRA). https://www.aclweb.org/anthology/2020.restup-1.4/
56. Vapnik VN (2000) The nature of statistical learning theory, Second Edition. Statistics for Engineering and Information Science. Springer
57. Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. BMC Bioinform 7(1):91
58. Waseem Z, Hovy D (2016) Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of the NAACL Student Research Workshop, pp. 88–93. Association for Computational Linguistics, San Diego, California. http://www.aclweb.org/anthology/N16-2013
59. Yang F, Peng X, Ghosh G, Shilon R, Ma H, Moore E, Predovic G (2019) Exploring deep multimodal fusion of text and photo for hate speech classification. In: Proceedings of the Third Workshop on Abusive Language Online, pp. 11–18. Association for Computational Linguistics, Florence, Italy. https://doi.org/10.18653/v1/W19-3502. https://www.aclweb.org/anthology/W19-3502
60. Yu D, Fu B, Xu G, Qin A (2019) Constrained nonnegative matrix factorization-based semi-supervised multilabel learning. Int J Mach Learn Cyber 10(5):1093–1100. https://doi.org/10.1007/s13042-018-0787-8
61. Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R (2019) Predicting the type and target of offensive posts in social media. In: NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers), pp. 1415–1420. https://doi.org/10.18653/v1/n19-1144
62. Zhang ML, Zhou ZH (2007) Ml-knn: a lazy learning approach to multi-label learning. Pattern Recogn 40(7):2038–2048