A Survey on Automatic Detection of Hate Speech in Text

PAULA FORTUNA, INESC TEC
SÉRGIO NUNES, INESC TEC and Faculty of Engineering, University of Porto

The scientific study of hate speech, from a computer science point of view, is recent. This survey organizes and describes the current state of the field, providing a structured overview of previous approaches, including core algorithms, methods, and main features used. This work also discusses the complexity of the concept of hate speech, defined in many platforms and contexts, and provides a unifying definition. This area has an unquestionable potential for societal impact, particularly in online communities and digital media platforms. The development and systematization of shared resources, such as guidelines, annotated datasets in multiple languages, and algorithms, is a crucial step in advancing the automatic detection of hate speech.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; *Information extraction*; **Information systems**; Sentiment analysis;

Additional Key Words and Phrases: Hate speech, literature review, text mining, opinion mining, natural language processing

ACM Reference format:
Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.* 51, 4, Article 85 (July 2018), 30 pages.
https://doi.org/10.1145/3232676

1 INTRODUCTION

Hate speech is a crime that has been growing in recent years [29], not only in face-to-face interactions but also in online communication. Several factors contribute to this. On one hand, on the internet, and social networks in particular, people are more likely to adopt an aggressive behavior because of the anonymity provided by these environments [8]. On the other hand, people have an increased willingness to express their opinions online [78], thus contributing to the propagation of hate speech as well. Since this type of prejudiced communication can be extremely harmful to society, governments and social network platforms can benefit from detection and prevention tools. Through this survey, we contribute to a solution to this problem by providing a systematic overview of research conducted in the field. We frame the problem, its definition, and identify methods and resources. We adopt a systematic approach, that critically analyses not only theoretical aspects but also practical resources, such as datasets and other projects.

This work is partially supported by FourEyes, a research line within project "TEC4Growth—Pervasive Intelligence, Enhancers and Proofs of Concept with Industrial Impact/NORTE-01- 0145-FEDER-000020" financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).
Authors' address: P. Fortuna and S. Nunes, INESC TEC, Campus da FEUP, Rua Dr. Roberto Frias, 4200 - 465 Porto, Portugal.

https://doi.org/10.1145/3232676

After this introductory section, in Section 2, we analyse a previous survey in the same topic. Next, in Section 3, we bring attention to the motivations for conducting research in this area. In Section 4, we analyse the theoretical aspects of studying this topic; we distinguish amongst different definitions of the concept, analyse particular cases of hate speech, relate hate speech with other close concepts, see how hate speech online has been evolving, and who are the main targets of it. To identify what has been done so far, in Section 5, we conduct a systematic literature review, with detailed methods and results. We also present a summary, including both quantitative data (e.g., evolution of the number of publications by year) and qualitative data (e.g., features explored in previous works). In Section 6, we present related datasets and open-source projects. In Section 7, we summarize the main challenges in this field and highlight some research opportunities. Finally, Section 8 concludes this survey with a summary of the main contributions and future perspectives.

2 RELATED WORK

The scientific study of hate speech, from a computer science point of view, is recent, and the number of studies in the field is low. We found only one survey article during the process of literature review. In this survey [66], the authors provide a short, comprehensive, structured, and critical overview of the field of automatic hate speech detection in natural language processing. The study is divided into several sections. First, they present the terminology necessary for studying hate speech. Then, in more depth, they analyse the features used in this problem. Later, they focus on research about bullying. Additionally, there is a section with some applications, namely, anticipating alarming societal changes. They also present a section dedicated to classification methods, challenges, and another about data.

Our approach is complementary to the referred study and at the same time our survey has specificities that we present here. First, we provide more detailed definitions: we compare hate speech with other related concepts, subtypes of hate, and enumerate rules that are helpful in the task of hate speech classification. Moreover, we utilize a systematic method and analyse not only documents focusing on algorithms but also focusing on descriptive statistics about hate speech detection. Complementarily, we also give an overview of the evolution of the area in recent years. Regarding the procedure of feature extraction, we use two categories, the "generic text mining features" and the "specific hate speech detection features." In our approach, this distinction is relevant, because the second category of features is focused on the specificities of this problem. We also enumerate existing data collections for this task in a more exhaustive way when compared to the previous study. We summarize open-source projects and conferences, as one of our goals is to enumerate useful resources in the field. Finally, we present motivations for considering this problem and difficulties found, as well.

3 WHY STUDY HATE SPEECH AUTOMATIC DETECTION?

Hate speech has become a popular topic in recent years. This is reflected not only by the increased media coverage of this problem but also by the growing political attention. There are several reasons to focus on hate speech automatic detection, which we discuss in the following list:

- **European Union Commission directives.** In recent years, the European Union Commission has been conducting different initiatives for decreasing hate speech. Several programs are being founded toward the fighting of hate speech (e.g., No Hate Speech Movement by the Council of Europe [38]). Another strategy used by the European Union to tackle this problem is through legislation. Recently, the European Union Commission pressured Facebook, YouTube, Twitter, and Microsoft to sign an EU hate speech code [40]. This includes the requirement to review the majority of valid notifications for removal of illegal hate speech

in less than 24h [40]. Also, European regulators accused Twitter of not being good enough at removing hate speech from its platform [46].

- **Automatic techniques not available**. Automated techniques aim to programmatically classify text as hate speech, making its detection easier and faster for the ones that have the responsibility to protect the public [9, 65]. These techniques can give a response in less than 24h, as presented in the previous point. Some studies have been conducted about the automatic detection of hate speech, but the tools provided are scarce.
- **Lack of data about hate speech.** There is a general lack of systematic monitoring, documentation, and data collection of hate and violence, namely, against LGBTI (lesbian, gay, bisexual, transgender, and intersex) people [42]. Nevertheless, detecting hate speech is a very important task, because it is connected with actual hate crimes [42, 65, 77], and automatic hate speech detection in text can also provide data about this phenomenon.
- **Hate speech removal.** Some companies and platforms might be interested in hate speech detection and removal [78]. For instance, online media publishers and online platforms, in general, need to attract advertisers and therefore cannot risk becoming known as platforms for hate speech [41]. Additionally, users can be interested in blocking discourse with hate speech to avoid being exposed to it.
- **Quality of service.** Social media companies provide a service [59]: They ease the communication between its users. They profit from this service and, therefore, assume public obligations with respect to the contents transmitted. In this case, quality of service regarding hate speech involves taking steps to discourage online hate and remove hate speech within a reasonable time. Both can be measured and compared to a standard imposed through legislation.

After outlining the motivations for studying this topic, in the next section, we define the important concepts in the field.

4 WHAT IS HATE SPEECH?

Deciding if a portion of text contains hate speech is not simple, even for humans. Hate speech is a complex phenomenon, intrinsically associated to relationships between groups, and also relying in language nuances. This is notorious in the low agreement found between annotators in the process of building new collections [65, 75]. Therefore, it is crucial to clearly define hate speech to make the task of its automatic identification easier [65].

4.1 Definitions from Several Sources

In this section, we collect different definitions of hate speech and compare perspectives from diverse sources (Tables 1 and 2). Concerning the sources of the definitions (Table 1), we decided for a wide range of origins, and we present here the motivations for that:

- European Union Commission (source Code of Conduct on Table 1) regulates other institutions.
- International minorities associations (ILGA) aim to protect people that are usually target of hate speech.
- Scientific papers, to include also a perspective from the scientific community. We provide only one example due to the similarity between the definitions adopted by the scientific community.
- Social networks conditions and terms (Facebook, YouTube, and Twitter), because in these platforms hate speech occurs regularly.

Table 1. Hate Speech Definitions

Source	Definition
Code of Conduct, between EU and companies	"All conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic" [79]
ILGA	"Hate speech is public expressions which spread, incite, promote or justify hatred, discrimination or hostility toward a specific group. They contribute to a general climate of intolerance which in turn makes attacks more probable against those given groups." [42]
Nobata et al.	"Language which attacks or demeans a group based on race, ethnic origin, religion, disability, gender, age, disability, or sexual orientation/gender identity." [58]
Facebook	"Content that attacks people based on their actual or perceived race, ethnicity, national origin, religion, sex, gender or gender identity, sexual orientation, disability or disease is not allowed. We do, however, allow clear attempts at humor or satire that might otherwise be considered a possible threat or attack. This includes content that many people may find to be in bad taste (ex: jokes, stand-up comedy, popular song lyrics, etc.)." [28]
YouTube	"Hate speech refers to content that promotes violence or hatred against individuals or groups based on certain attributes, such as race or ethnic origin, religion, disability, gender, age, veteran status and sexual orientation/gender identity. There is a fine line between what is and what is not considered to be hate speech. For instance, it is generally okay to criticize a nation-state, but not okay to post malicious hateful comments about a group of people solely based on their ethnicity." [82]
Twitter	"Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease." [72]

Table 2. Content Analysis of Hate Speech Definitions

Source	Hate speech is to incite violence or hate	Hate speech is to attack or diminish	Hate speech has specific targets	Humour has a specific status
EU Code of conduct	Yes	No	Yes	No
ILGA	Yes	No	Yes	No
Scientific paper	No	Yes	Yes	No
Facebook	No	Yes	Yes	Yes
YouTube	Yes	No	Yes	No
Twitter	Yes	Yes	Yes	No

To better understand the definitions found in Table 2, we consider distinct aspects: the source (shared with Table 1); and four dimensions in which the definitions can be compared ("Hate speech has specific targets," "Hate speech is to incite violence or hate," "Hate speech is to attack or diminish," "Humour has a specific status"). These columns are the result of a manual analysis of the definitions, employing a method based on content analysis [48]. We expand and analyse these four dimensions in the next paragraphs.

- **Hate speech has specific targets.** All the quoted definitions point out that hate speech has specific targets and it is based on specific characteristics of groups, like ethnic origin, religion, or other.
- **Hate speech is to incite violence or hate.** The several definitions use slightly different terms to describe when hate speech occurs. The majority of the definitions point out that hate speech is to incite violence or hate toward a minority (definitions from Code of Conduct, ILGA, YouTube, Twitter).
- **Hate speech is to attack or diminish.** Additionally, some other definitions state that hate speech is to use language that attacks or diminishes these groups (definitions from Facebook, YouTube, Twitter).
- **Humour has a specific status.** However, Facebook points out that some offensive and humorous content is allowed (definition from Facebook). The exceptional status of humour makes the boundaries about what is forbidden in the platform more difficult to understand.

Despite the similarities between the definitions, we conclude that there are some nuances that distinguish them (e.g., the usage of humour). Moreover, the content analysis presented allows us to suggest a definition for hate speech.

4.2 Our Definition of Hate Speech

In the previous section, we used four dimensions to compare some definitions of hate speech. This analysis led us to propose a definition of hate speech:

> Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used.

We should also complement this definition. If violence can occur physically and explicitly, then it can also be subtle. This is the case when stereotypes are reinforced, giving a justification to discrimination and negative bias toward these groups. Consequently, we consider that all subtle forms of discrimination, even jokes, must be marked as hate speech. This is the case, because this type of joke indicates relations between the groups of the jokers and the groups targeted by the jokes, racial relations, and stereotypes [49]. Moreover, repeating these jokes can become a way of reinforcing racist attitudes [45] and, although they are considered harmless, they also have negative psychological effects for some people [26]. In the next subsection, we go deeper in the notion of hate speech by discussing particular cases and examples.

4.3 Particular Cases and Examples of Hate Speech

In this section, we start by analysing how one particular social network has been tackling the problem of hate speech detection. Some definitions and cases that Facebook uses to train its workers on this task are revealed. According to Facebook directives [47], a message contains hate speech when two conditions are met:

Table 3. Text Messages Classified by Facebook (Table from Krause and Grassegger [47])

Message	Evaluation
Don't trust boys!	Violating—delete
Refugees should face the figuring squad!	Violating—delete
Fucking Muslims!	Violating—delete
Fucking migrants!	Non-violating—Ignore
Migrants are filthy cockroaches that will infect our country	Violating—delete
I'm such a faggot people call me diva!	Non-violating—Ignore
The French are alcoholics	Violating—delete
All English people are dirty!	Violating—delete
Don't try to explain—Irish Catholics are just idiots	Violating—delete
Migrants are scum!	Violating—delete
People should stop to use the word nigger.	Non-violating—Ignore
I hate migrants!	Non-violating—Ignore
Don't trust boys who say they love you!	Non-violating—Ignore
Tall girls are just freaks!	Non-violating—Ignore
American shitheads!	Violating—delete
Migrants are so filthy!	Non-violating—Ignore
Refugees! More like rape-fugees!	Violating—delete
Asylum seekers out!	Violating—delete
Group for blacks only!	Non-violating—Ignore

- a verbal attack occurs.
- the target of the attack is from a "protected category" (religious affiliation, national origin, etc.).

Some rules for hate speech classification are [47]:

- members of religious groups are protected, religion itself is not.
- speaking badly about countries (e.g., France or Germany) is allowed, in general; however, condemning people on the basis of their nationality is not.
- a protected category combined with another protected category results in yet another protected category (e.g., if someone writes "Irish women are dumb," they would be breaking the rules and their post would be deleted, because "national origins" and "sex" categories apply).
- combining a protected category with an unprotected category, however, results in an unprotected category. For this reason, the sentence "Irish teenagers are dumb" does not need to be deleted, because the term teenager does not enjoy special protection.
- saying "fucking Muslims" is not allowed, as religious affiliation is a protected category [47].
- however, the sentence "fucking migrants" is allowed, as migrants are only a "quasi-protected category"—a special form that was introduced after complaints were made. This rule states that promoting hate against migrants is allowed under certain circumstances: Statements such as "migrants are dirty" are allowed, while "migrants are dirt" is not [47].

In addition, some sentences are used to exemplify what should be marked as hate speech (Table 3). The examples marked as "violating" should be deleted by the workers, whereas the examples marked as "non-violating" should be ignored.

The rules presented so far can be discussed. From our point of view, there is no reason to restrain hate speech to specific "protected categories." First, in the case that new targets of hate speech appear, those are undetectable unless the "protected categories" are redefined. Besides, prejudice can occur even when protected categories are not specifically implied. For instance, boys and men receive at an early age confining and stereotypical messages. Those come from family, peers, or media, instructing them how to behave, feel, relate to each other, to girls, and to women. Some of these messages are harmful and have short- and long-term consequences for the boys but also for women, their families, their community, and society as a whole [51].

Despite the critique in the previous paragraph, the approach described in the article is systematic, and some of the rules presented are indeed useful for hate speech identification. Additionally, other scientific papers also provide rules that can be regarded as a guide for the classification of hate speech. We propose a list of the main rules for hate speech identification. We are in the presence of hate speech when people:

- call attention to the fact that an individual belongs to a group and invoke a well known and disparaging stereotype about that group [74].
- make generalized negative statements about minority groups as in "the refugees will live off our money," due to the incitation of a negative bias toward the group. However, some authors [65] were unsure about this example as being hate speech.
- use disparaging terms and racial epithets with the intent to harm.
- use sexist or racial slurs [65].
- use those words to show pride, even when the speaker belongs to the targeted group. If there is no contextual clue about the group membership, then such terms are categorized as hateful [74].
- endorse organizations that promote hate speech. Despite that this is not a direct verbal attack on other group, this must be marked as hate speech. In this aspect, we oppose to the perspective of some other authors [74].
- speak badly about countries or religions (e.g., France, Portugal, Catholicism, Islam) is allowed in general, but discrimination is not allowed based on these categories.
- make statements about the superiority of the in-group.

It is also important to point out that in some particular cases, we are not in the presence of hate speech. This is the case when people:

- discuss offensive words (e.g., explanation of the meaning). Such expressions might be acceptable in those contexts [74].
- refer to an organization associated with hate crimes. For instance the name "Ku Klux Klan" is not hateful, as it may appear in historical articles or other legitimate communication [74].
- use words like "black," "white," "filthy," or other. This is marked as hate speech only in some circumstances. Outside of context, these words bear no racial undertones of their own [50].

The presented rules, and our critical view on them, pointed out that we have a more inclusive and general definition about hate speech than some other perspectives found in the literature. This is the case, because we propose that subtle forms of discrimination found on the internet and online social networks should also be spotted.

4.4 Hate Speech and Other Related Concepts

In the previous sections, we analysed different definitions of hate speech and presented some examples. Another way of better understanding this complex phenomenon is by comparison with other related concepts. Several of those concepts found in literature were hate [68], cyberbullying

Table 4. Comparison between Hate Speech Definition and Related Concepts

Concept	Definition of the concept	Distinction from hate speech
Hate	Expression of hostility without any stated explanation for it [68].	Hate speech is hate focused on stereotypes, and not so general.
Cyberbullying	Aggressive and intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time, against a victim who can not easily defend him or herself [10].	Hate speech is more general and not necessarily focused on a specific person.
Discrimination	Process through which a difference is identified and then used as the basis of unfair treatment [69].	Hate speech is a form of discrimination, through verbal means.
Flaming	Flaming are hostile, profane and intimidating comments that can disrupt participation in a community [35]	Hate speech can occur in any context, whereas flaming is aimed toward a participant in the specific context of a discussion.
Abusive language	The term abusive language was used to refer to hurtful language and includes hate speech, derogatory language and also profanity [58].	Hate speech is a type of abusive language.
Profanity	Offensive or obscene word or phrase [23].	Hate speech can use profanity, but not necessarily.
Toxic language or comment	Toxic comments are rude, disrespectful or unreasonable messages that are likely to make a person to leave a discussion [43].	Not all toxic comments contain hate speech. Also some hate speech can make people discuss more.
Extremism	Ideology associated with extremists or hate groups, promoting violence, often aiming to segment populations and reclaiming status, where outgroups are presented both as perpetrators or inferior populations. [55].	Extremist discourses use frequently hate speech. However, these discourses focus other topics as well [55], such as new members recruitment, government and social media demonization of the in-group and persuasion [62].
Radicalization	Online radicalization is similar to the extremism concept and has been studied on multiple topics and domains, such as terrorism, anti-black communities, or nationalism [2].	Radical discourses, like extremism, can use hate speech. However in radical discourses topics like war, religion and negative emotions [2] are common while hate speech can be more subtle and grounded in stereotypes.

[10], abusive language [58], discrimination [69], profanity [23], toxicity [43], flaming [35], extremism [55, 62], and radicalization [2]. In Table 4, we distinguish between these concepts and hate speech. In addition to the concepts already presented, it is also important to identify each type of hate speech that we found in literature (Table 5).

If, on one hand, all the concepts presented in Table 4 are slightly distinct from hate speech, then, on the other hand, they are related to it. Therefore, literature and empirical studies focusing on them can give insight about how to automatically detect hate speech.

5 WHAT HAS BEEN DONE SO FAR IN AUTOMATIC HATE SPEECH DETECTION?

With the goal of understanding the work already developed in this field, we conducted a systematic literature review. In this section, we describe the method adopted and the achieved results in detail. In this context, we use the term document as a synonym for paper, thesis, or any other sort of text manuscript.

Table 5. Types of Hate Speech and Examples
(Table from Silva et al. [67])

Categories	Example of possible targets
Race	nigga, black people, white people
Behavior	insecure people, sensitive people
Physical	obese people, beautiful people
Sexual orientation	gay people, straight people
Class	ghetto people, rich people
Gender	pregnant people, cunt, sexist people
Ethnicity	chinese people, indian people, paki
Disability	retard, bipolar people
Religion	religious people, jewish people
Other	drunk people, shallow people

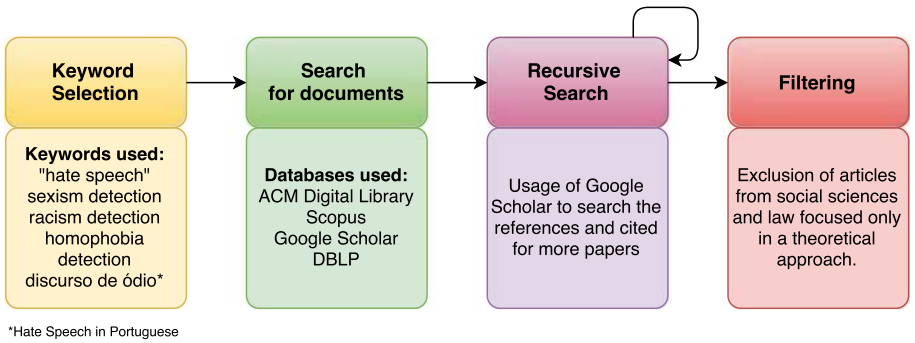

Fig. 1. Methodology for document collection.

5.1 Systematic Literature Review

We conducted a systematic literature review with the goal of collecting the largest possible number of documents in the area, and for that, we developed a method that is described in the next section.

5.1.1 Method Description. The method is structured in four phases, that are presented and summarized in Figure 1.

We describe here the different phases in more detail:

- **Keyword selection.** The first phase conducted was the keywords selection. We bear in mind that hate speech is a concept that became more popular recently. Therefore, some other related concepts could have been used in the past by the scientific community. We considered terms referring to particular types of hate speech (sexism, racism, and homophobia). Besides, we also considered the search for "hate speech" in other languages (Portuguese and Spanish).
- **Search for documents.** We searched in different databases and services (ACM Digital Library, Scopus, Google Scholar, and DBLP), aiming to gather the largest possible number of documents in the areas of computer science and engineering. Databases from other scientific areas were not considered.

- **Recursive search.** We used Google Scholar to get both the references and documents that cite the original work. We check on these two sets and search for the expression "hate speech" on the titles of the candidate documents. Recursively, we repeated the search with the new documents found.
- **Filtering.** An initial step of filtering was conducted. Documents from social sciences and law were excluded from our collection.

5.1.2 Documents Collection and Annotation. The process of collecting documents was conducted from September 1, 2016 to May 18, 2017. We ended up with a total of 128 documents that we described using the following metrics:

- Name.
- Area of knowledge (we created the categories: "Law and Social Sciences" and "Computer Science and Engineering").
- Conference or journal name.
- Keywords in the document.
- Particular hate (while some articles focus generally on hate speech, others focus on particular types, such as racism).
- Social network (refers to the platform used as the source of data).
- Number of instances used (refers to the size of the dataset used in the work).
- Algorithms used.
- Type of document (we created the categories: "algorithms about hate speech," "algorithms but not about hate speech," "descriptive statistics about hate speech," "descriptives statistics but not about hate speech," and "theoretical").
- Year of the document.

In the next sections, we present the main results from our systematic literature review.

5.1.3 Area of Knowledge. We classified each document as "Law and Social Sciences" or "Computer Science and Engineering." We concluded that the majority of the works we found is from the first category (N = 76), whereas a fewer number of articles is from "Computer Science and Engineering" (N = 51). In the scope of this work, we are only interested in analysing the papers from the set "Computer Science and Engineering." In the following sections, we only focus on this group.

5.1.4 Year of the Document. As we can see in Figure 2, before 2014 the number of documents related to hate speech, from the type "Computer Science and Engineering," was very low. However, since 2014 this number has been increasing. Regarding the smaller value in 2017, we should bear in mind that the collection of new documents stopped in May 2017.

5.1.5 Publication Venue. From the total of 51 documents in the set of "Computer Science and Engineering," we found 37 different venues. The publication contexts with more than one occurrence in our collection are presented in Table 6. The more common platform for publication of hate speech documents was ArXiv, an open-access repository of electronic preprints. This can partially be explained by the fact that hate speech detection is a recent area with a significant number of autonomous and exploratory work being conducted. Additionally, the results about the publication of documents point out that the venues found are not specific for hate speech. However, we try to find if such platforms exist (e.g., journals or conferences). We discovered some conferences more related with hate speech automatic detection (Table 7), that seem to be at an early stage.

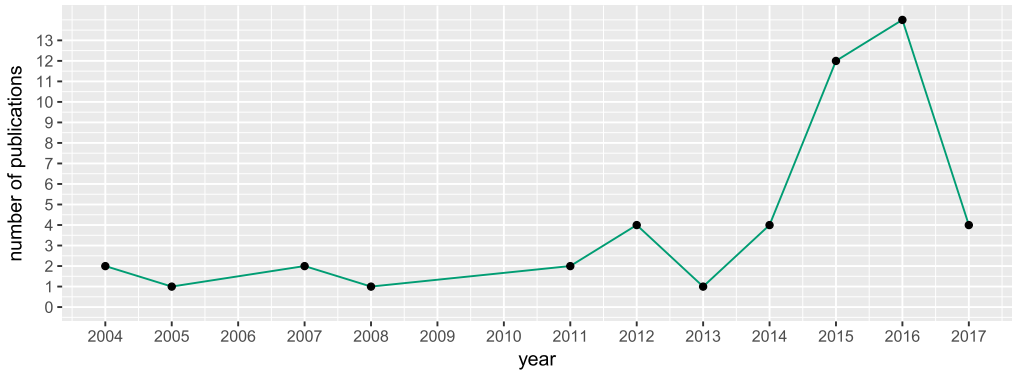

Fig. 2. Number of publications per year from the "Computer Science and Engineering" set (N = 51).

Table 6. Most Used Platforms for Publication of Documents from "Computer Science and Engineering"

Platform for publication	n
ArXiv	6
International Conference on World Wide Web	2
Master Thesis	2

Table 7. Conferences Related to Hate Speech Detection, Respective Area of Conference and Reference

Conferences related to hate speech detection	Area	Ref
ALW1: 1st Workshop on Abusive Language Online (2017)	Computer science	[1]
Workshop on Online Harassment (2017)	Computer science	[11]
Text Analytics for Cybersecurity and Online Safety (2016)	Computer science	[21]
Hate speech Conference (2017)	Social Sciences	[37]
UNAOC #SpreadNoHate (2017)	Social Sciences	[60]
Interdisciplinary conference on Hate Speech (2017)	Humanities	[13]

5.1.6 Number of Citations. We collected the number of citations for each document in Google Scholar and concluded that the majority of works are cited less than four times (Figure 3). The top five papers with more citation in our sample are presented in Table 8.

5.1.7 Keywords in the Document. All the keywords referred in the documents from the area "Computer Science and Engineering" were grouped and analysed for absolute frequencies (Table 9). We can infer that these documents study hate speech when it is related with:

- **"related concepts"** (cyberbullying, cyber hate, sectarianism, and freedom of speech).
- **"machine learning"** (classification, sentiment analysis, filtering systems, and machine learning).
- **"social media"** (internet, social media, social network, social networking, and hashtag).

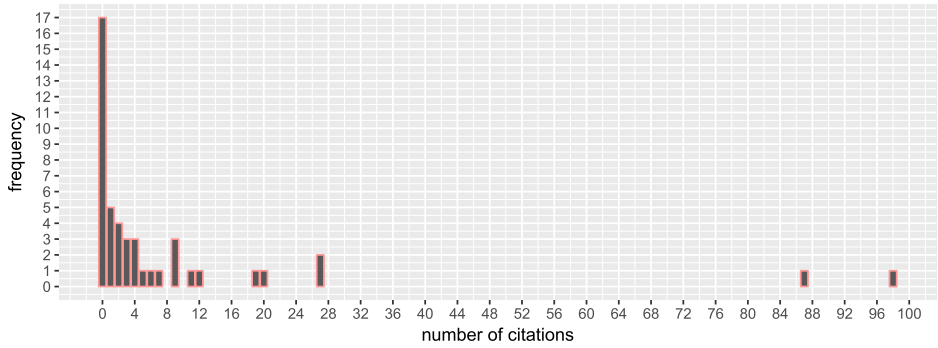

Fig. 3. Number of citations of the papers from "Computer Science and Engineering."

Table 8. Most Cited Papers from the "Computer Science and Engineering" Set

Paper	**Citations**
Modelling the Detection of Textual Cyberbullying (2011) [24]	87
Perverts and sodomites: Homophobia as hate speech in Africa (2002) [64]	63
Classifying racist texts using a support vector machine (2004) [33]	27
Improved Cyberbullying Detection Using Gender Information (2012) [16]	27
Detecting Hate Speech on the World Wide Web (2012) [74]	20

Table 9. Keywords of the Papers from "Computer Science and Engineering"

Keyword	Frequency
Cyberbullying	5
Social media	5
Classification	4
Internet	4
Freedom of speech	3
Hate speech	3
Machine learning	3
NLP	3
Sentiment analysis	3
Social network	3
Social networking (online)	3
Cyber hate	2
Filtering systems	2
Hashtag	2
Sectarianism	2

5.1.8 Social Networks. The found documents analyse datasets with messages that were collected from social networks. Twitter is the most commonly used source, followed by general sites, YouTube, and Yahoo! (Table 10).

5.1.9 Number of Used Instances. Generally, in machine learning, an instance is an example object that can be used in different steps of the learning process, either for training, validating, or

Table 10. Social Networks Used in the Papers from "Computer Science and Engineering"

Social network	Frequency
Twitter	16
Sites	5
YouTube	3
Yahoo! finance	2
American Jewish Congress (AJC) sites	1
Ask.fm	1
Blogs	1
Documents	1
Facebook	1
formspring.me	1
myspace.com	1
Tumblr	1
Whisper	1
White supremacist forums	1
Yahoo news	1
Yahoo!	1

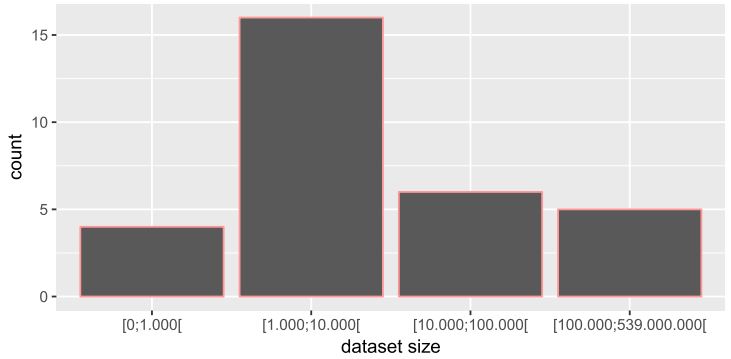

Fig. 4. Dataset sizes used in the papers from "Computer Science and Engineering."

testing. In the context of hate speech automatic detection, an instance is a text message, with a classification label. Regarding the number of instances per dataset, this number has a wide range of magnitudes (Figure 4). Nevertheless, we can conclude that the majority of papers use between 1,000 and 10,000 instances.

5.1.10 General or Particular Hate Speech. We analyse if the found documents focus on general hate speech or on more particular types of hate. The majority (N = 26) considers general hate speech (Table 11), however, there is a large number of papers (N = 18) that focus particularly on racism.

5.1.11 Algorithms Used. The most common approach found in our systematic literature review consists of building a Machine Learning model for hate speech classification. We also found that the most common algorithms used are SVM, Random Forests, and Decision Trees (Table 12).

Table 11. Type of Hate Speech Analysed in the Papers from "Computer Science and Engineering"

Hate type	Frequency
General hate speech	26
Racism	18
Sexism	6
Religion	4
Anti-semitism	1
Nationality	1
Other	1
Physical/mental handicap	1
Politics	1
Sectarianism	1
Socio-economical status	1

Table 12. Algorithms Used in the Papers from "Computer Science and Engineering"

Algorithms	Frequencies
SVM	10
Random forests	5
Decision trees	4
Logistic regression	4
Naive bayes	3
Deep learning	1
DNN	1
Ensemble	1
GBDT	1
LSTM	1
Non-supervised	1
One-class classifiers	1
Skip-bigram model	1

5.1.12 Type of Approach in the Document. In our study, we want to identify the documents that focus on algorithms for hate speech detection. For that we analyse in depth the "Computer Science and Engineering" articles. We classified the approach of the documents in one of the following categories: "algorithms for hate speech," "algorithms but not for hate speech," "descriptive statistics about hate speech," "descriptives statistics but not about hate speech," and "theoretical." In Figure 5, we can see that the most common types are "algorithms for hate speech" and "algorithms but not for hate speech." However, the category "descriptives statistics but not about hate speech" only has one paper about hashtags usage as a way of monitoring discourse [30].

In the next sections, we focus on the "descriptives statistics about hate speech" (N = 9) and "algorithms for hate speech" (N = 17) documents, because those are the most relevant for the particular field of automatic hate speech detection.

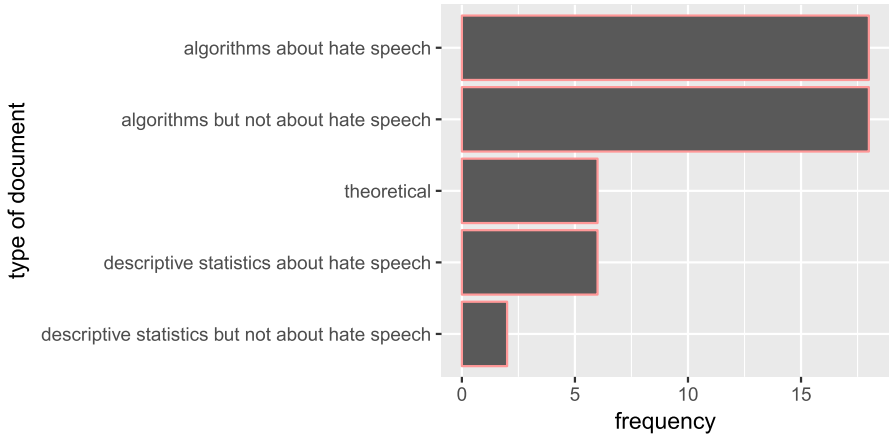

Fig. 5. Type of papers from "Computer Science and Engineering."

5.2 Documents Focusing on Descriptives Statistics about Hate Speech Detection

In the previous sections, we saw that, according to the definitions of hate speech, its targets are groups or individuals based on their specific attributes, such as ethnic origin, religion, disability, gender identity, age, veteran status, sexual orientation or other. Studies have been conducted with the goal of describing online hate speech and which groups are more threatened. This section presents the main conclusions found on the articles that we labeled as having a more descriptive approach to the problem of hate speech detection. We found descriptive articles about Racism, Sexism, Prejudice toward refugees, Homophobia, and General hate speech.

Racism. In one study [50], the authors tried to understand when hate speech occurs and why messages in social networks are cataloged as racist. They concluded that in the majority of the cases (86%) this is because of the "presence of offensive words." Other motives are "references to painful historical contexts" and "presence of stereotypes or threatening." The authors of another study [84] describe racism across the United States and tried to understand the geographic distribution of racist tweets. They used information gathered in Twitter to describe the frequencies of tweets in the several states, using the geographic reference of the messages.

Sexism. In a study about sexism [41], a very simplistic approach was conducted. Tweets using offensive words toward woman were collected using the Twitter search API. Approximately 5,500 tweets were gathered and coded by one researcher, using a simple binary model. Despite the limitations of the study (e.g., many of the tweets were repeating the title or lyrics from popular songs that included the searched offensive words), it was still relevant for understanding that offensive communication toward woman is a reality in Twitter. A second study also describes misogynistic language on Twitter [6]. The main conclusions were that 100,000 instances of the word rape used in UK-based Twitter accounts were found, from which around 12% appeared to be threatening. Moreover, approximately 29% of the rape tweets appeared to use the term in a casual or metaphorical way. However, this study also points out that women are as almost as likely as men to use offensive terms against women on Twitter.

Prejudice Toward Refugees. Another study was focused on the annotation of a dataset in German for hate speech against refugees [65]. The main goal of this study was to point out the difficulties and challenges when annotating a dataset.

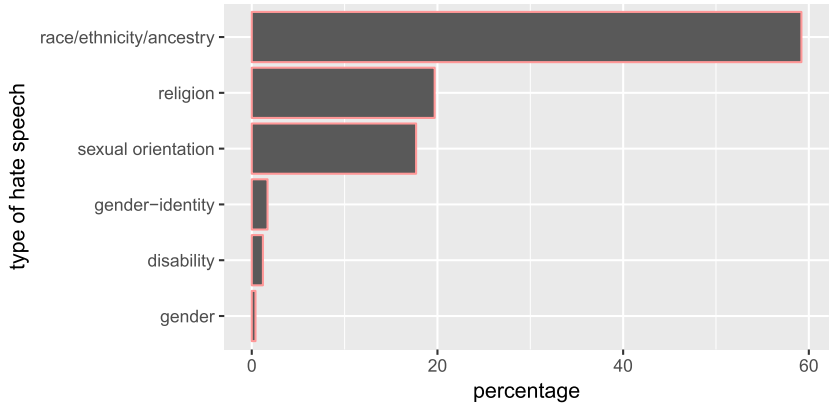

Fig. 6. Percentage for each type over all hate crimes in the USA (Source: FBI [29]).

Homophobia. Some other study [64], using an ethnographic methodology, was conducted in Africa. Data was collected from several sources (e.g., newspapers, sites) to conclude that homophobic discourses were using arguments related with Abnormality, Xenophobia, Racism, Barbarism, Immorality, Unpatriotism, Heterosexism, AntiChristianity, UnAfrican, Animalistic behaviour, Inhumane, Criminality, Pathology, and Satanism.

General Hate Speech. Finally, other studies take into consideration several types of hate speech at the same time. In one particular case [67], two social networks (Twitter and Whisper) were crawled with expressions that follow a rigid pattern: $I <$ *intensity* $><$ *userintent* $><$ *hatetarget* $>$. One message following this pattern would be "I really hate people." After collecting the messages, the researchers tried to infer the target of hate in the tweets. With this method, they concluded that "race," "behavior," and "physical" were the most hated categories. Finally, an analysis of data recorded by the FBI in 2015 [29] for victims in the USA of single-bias hate crime incidents, showed that the offender's bias was toward different targets in different proportions (Figure 6).

5.3 Documents Focusing on Algorithms for Hate Speech Detection

Regarding the documents focusing on "algorithms for hate speech detection," in what concerns to the methodology, the researchers used machine learning for hate speech classification. Additionally, we found that in the majority of the cases, the research was conducted in English. However, there were some exceptions. In these cases, the considered languages were Dutch [71], German [65], and Italian [22]. In the next sections, we present details on how these studies obtain datasets and compare the performances of the different approaches.

5.3.1 Datasets Used in the Papers. In the majority of the 17 papers focusing on "algorithms for hate speech," new different data was collected and annotated. However, only in a few studies data is made available for other researchers (label "own, available"), and only in one case an already published dataset is used ("published dataset") (Figure 7). The reduced number of datasets that are publicly shared is a relevant aspect in this area, making more difficult the comparison between different approaches. The datasets available in the area are described in Section 6.

5.3.2 Achieved Performances. In the collected papers, several metrics were computed to estimate the performance of the models. Precision, Recall, and F-measure were the most common metrics and in some other studies, Accuracy and AUC (Area Under Curve) were also considered.

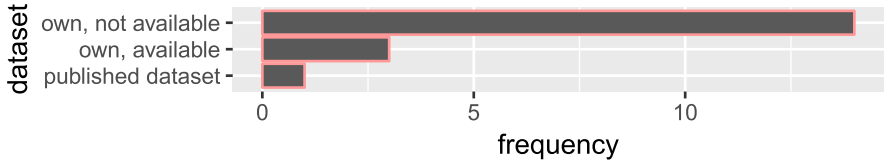

Fig. 7. Dataset availability in the documents with "algorithms about hate speech."

In Table 13, the results of the studies are presented in descending order of the F-measure value. These results should be analysed with some caution, because different configurations, datasets, and definitions are being compared. We try to summarize the best results for each paper. We conclude that it is not clear which approaches perform better. On one hand, the best results were achieved when deep learning was used. On the other hand, this was not a consistent result. Comparative studies could help to understand this question.

5.4 Text Mining Approaches in Automatic Hate Speech Detection

In this section, we analyse features described in the papers focusing on algorithms for hate speech detection, and also other studies focusing on related concepts (e.g., Cyberbullying). Finding the right features for a classification problem can be one of the more demanding tasks when using machine learning. Therefore, we allocate this specific section to describe the features already used by other authors. We divide the features into two categories: general features used in text mining, which are common in other text mining fields; and the specific hate speech detection features, which we found in hate speech detection documents and are intrinsically related to the characteristics of this problem. We present our analysis in this section.

5.4.1 General Features Used in Text Mining. The majority of the papers we found try to adapt strategies already known in text mining to the specific problem of automatic detection of hate speech. We define general features as the features commonly used in text mining. We start by the most simplistic approaches that use dictionaries and lexicons.

Dictionaries. One strategy in text mining is the use of dictionaries. This approach consists in making a list of words (the dictionary) that are searched and counted in the text. These frequencies can be used directly as features or to compute scores. In the case of hate speech detection, this has been conducted using:

- Content words (such as insults and swear words, reaction words, personal pronouns) collected from www.noswearing.com [53].
- Number of profane words in the text, with a dictionary that consists of 414 words, including acronyms and abbreviations, where the majority are adjectives and nouns [16].
- Label Specific Features that consisted in using frequently used forms of verbal abuse as well as widely used stereotypical utterances [24].
- Ortony Lexicon was also used for negative affect detection; the Ortony lexicon contains a list of words denoting a negative connotation and can be useful, because not every rude comment necessarily contains profanity and can be equally harmful [24].

This methodology can be used with an additional step of normalization, by considering the total number of words in each comment [16]. Besides, it is also possible to use this kind of approach with regular expressions [54].

Table 13. Results Evaluation of the Papers in the Metrics Accuracy (Acc), Precision (P), Recall (R), F-measure (F) and AUC, Respective Features, and Algorithms Used

Year	Acc	P	R	F	AUC	Features	Algorithms	Paper
2017	—	0.93	0.93	0.93	—	—	Logistic Regression, Random Forest, SVM, GBDT, DNN, CNN	[4]
2004	—	~0.90	0.9	0.9	0.9	BOW, N-grams, POS	SVM	[32]
2017	—	0.91	0.9	0.9	—	TF-IDF, POS, sentiment, hashtags, mentions, retweets, URLs, number of characters, words, and syllables	Logistic Regression, SVM	[19]
2017	—	0.833	0.872	0.851	—	POS, sentiment analysis, word2vec, CBOW, N-grams, text features	SVM, LSTM	[22]
2016	—	0.83	0.83	0.83	—	N-grams, length, punctuation, POS	Skip-bigram Model	[58]
2014	—	0.89	0.69	0.77	—	N-gram, typed dependencies	Random Forest Decision Tree, SVM	[7]
2015	—	0.89	0.69	0.77	—	N-gram, typed dependencies	Random Forest Decision Tree, SVM, Bayesian Logistic Regression, Ensemble	[8]
2016	—	0.72	0.77	0.73	—	User features	Logistic Regression	[77]
2016	—	0.79	0.59	0.68	—	BOW, dictionary, typed dependencies	SVM, Random Forest, Decision Tree	[9]
2015	—	0.65	0.64	0.65	—	Rule-based approach, sentiment analysis, typed dependencies	Non-supervised	[31]
2012	—	0.68	0.6	0.63	—	Template-based strategies, word sense, disambiguation,	SVM	[74]
2016	—	0.49	0.43	0.46	0.63	Dictionaries	SVM	[71]
2015	—	—	—	—	0.8	paragraph2vec	Logistic Regression	[25]
2016	0.91	—	—	—	—	word2vec	Deep Learning	[83]
2013	0.76	—	—	—	—	N-grams	Naive Bayes	[50]
2016	—	0.73	0.86	—	—	Topic modelling, sentiment analysis, tone analysis, semantic analysis, contextual metadata	One-class Classifiers, Random Forest, Naive Bayes, Decision Trees	[3]
2004	—	0.93	0.87	—	—	BOW, N-grams, POS	SVM	[33]
2014	—	0.97	0.82	—	—	TF-IDF, N-grams, topic similarity, sentiment analysis	Naive Bayes	[52]

Distance Metric. Some studies have pointed out that in text messages it is possible that the offensive words are obscured with an intentional misspelling, often a single character substitution [74]. Examples of these terms are "@ss," "sh1t" [58], "nagger," or homophones, such as "joo" [74]. The Levenshtein distance, i.e., the minimum number of edits necessary to transform one string into another, can be used for this purpose [57]. The distance metric can be used to complement dictionary-based approaches.

Bag-of-words (BOW). Another model similar to dictionaries is the bag-of-words [9, 33, 50]. In this case, a corpus is created based on the words that are in the training data, instead of a predefined set of words, as in the dictionaries. After collecting all the words, the frequency of each one is used as a feature for training a classifier. The disadvantages of this kind of approaches are that the word sequence is ignored, and also its syntactic and semantic content. Therefore, it can lead to misclassification if the words are used in different contexts. To overcome this limitation N-grams can be adopted.

N-grams. N-grams are one of the most used techniques in hate speech automatic detection and related tasks [4, 9, 19, 33, 52, 58, 77]. The most common N-grams approach consists in combining sequential words into lists with size *N*. In this case, the goal is to enumerate all the expressions of size *N* and count all occurrences. This allows improving classifiers' performance, because it incorporates at some degree the context of each word. Instead of using words it is also possible to use N-grams with characters or syllables. This approach is not so susceptible to spelling variations as for when words are used. Character N-gram features proved to be more predictive than token N-gram features, for the specific problem of abusive language detection [56].

However, using N-grams also have disadvantages. One disadvantage is that related words can have a high distance in a sentence [9] and a solution for this problem, such as increasing the *N* value, slows down the processing speed [10]. Also, studies point out that higher *N* values (5) perform better than lower values (unigrams and trigrams) [52]. In a survey [66], researchers report that N-grams features are often reported to be highly predictive in the problem of hate speech automatic detection, but perform better when combined with others.

Profanity Windows. Profanity windows are a mixture of a dictionary approach and N-grams. The goal is to check if a second person pronoun is followed by a profane word within the size of a window and then create a boolean feature with this information [16].

TF-IDF. The TF-IDF (term frequency-inverse document frequency) was also used in this kind of classification problems [24]. TF-IDF is a measure of the importance of a word in a document within a corpus and increases in proportion to the number of times that a word appears in the document. However, it is distinct from a bag of words, or N-grams, because the frequency of the term is off-setted by the frequency of the word in the corpus, which compensates the fact that some words appear more frequently in general (e.g., stop words).

Part-of-speech. Part-of-speech (POS) approaches make it possible to improve the importance of the context and detect the role of the word in the context of a sentence. These approaches consist in detecting the category of the word, for instance, personal pronoun (PRP), Verb non-3rd person singular present form (VBP), Adjectives (JJ), Determiners (DT), Verb base forms (VB). Part-of-speech has also been used in hate speech detection problem [33]. With these features, it was possible to identify frequent bigram pairs, namely PRP_VBP, JJ_DT and VB_PRP, which would map as "you are" [24]. It was also used to detect sentences such as "send them home," "get them out," or "should be hung" [7]. However, POS proved to cause confusion in the classes identification, when used as features.

Lexical Syntactic Feature-based (LSF). In a study [10], the natural language processing parser, proposed by Stanford Natural Language Processing Group [34], was used to capture the grammatical dependencies within a sentence. The features obtained are pairs of words in the form "(governor, dependent)", where the dependent is an appositional of the governor (e.g., "You, by any means, an idiot." means that "idiot," the dependent, is a modifier of the pronoun "you," the governor). These features are also being used in hate speech detection [10].

Rule Based Approaches. Some rule-based approaches have been used in the context of text mining. A class association rule-based approach, more than frequencies, is enriched by linguistic knowledge. Rule-based methods do not involve learning and typically rely on a pre-compiled list or dictionary of subjectivity clues [36]. For instance, rule-based approaches were used to classify antagonistic and tense content on Twitter using associational terms as features. They also included accusational and attributional terms targeted at only one or several persons following a socially disruptive event as features, in an effort to capture the context of the terms used.

Participant-vocabulary Consistency (PVC). In a study about cyberbullying [63], this method is used to characterize the tendency of each user to harass or to be harassed, and the tendency of a key phrase to be indicative of harassment. For applying this method it is necessary a set of messages from the same user. In this problem, for each user, it is assigned a bully score (b) and a victim score (v). For each feature (e.g., N-grams) a feature-indicator score (w) is used. It represents how much the feature is an indicator of a bullying interaction. Learning is then an optimization problem over parameters b, v, and w.

Template Based Strategy. The basic idea of this strategy is to build a corpus of words, and for each word in the corpus, collect K words that occur around [61]. This information can be used as context. This strategy has been used for feature extraction in the problem of hate speech detection as well [74]. In this case, a corpus of words and a template for each word was listed, as in "W-1:go W+0:back W+1:to." This is an example of a template for a two word window on the word "back."

Word Sense Disambiguation Techniques. This problem consists in identifying the sense of a word in the context of a sentence when it can have multiple meanings [81]. In a study, the stereotyped sense of the words was considered, to understand if the text is anti-semitic or not [74].

Typed Dependencies. Typed dependencies were also used in hate speech related studies. First, to understand the type of features that we can obtain with this, the Stanford typed dependencies representation provides a description of the grammatical relationships in a sentence, that can be used by people without linguistic expertise [20]. This was used for extracting Theme-based Grammatical Patterns [31] and also for detecting hate speech specific othering language [7, 8], that we will present within the specific hate speech detection features. Some studies report significant performance improvements in hate speech automatic detection based on this feature [9, 31].

Topic Classification. With these features, the aim is to discover the abstract topic that occurs in a document. In a particular study [3], topic modelling linguistic features were used to identify posts belonging to a defined topic (Race or Religion).

Sentiment. Bearing in mind that hate speech has a negative polarity, authors have been computing the sentiment as a feature for hate speech detection [3, 19, 22, 31, 52, 53]. Different approaches have been considered (e.g., multi-step, single-step) [66]. Authors usually use this feature in combination with others that proved to improve results [52].

Word Embeddings. Some authors [25] use a paragraph2vec approach to classify language on user comments as abusive or clean and also to predict the central word in the message [25]. FastText is also being used [4]. A problem that is referred in hate speech detection is that sentences must be classified and not words [66]. Averaging the vectors of all words in a sentence can be a solution, however, this method has limited effectiveness [58]. Alternatively, other authors propose comment embeddings to solve this problem [25].

Deep Learning. Deep learning techniques are also recently being used in text classification and sentiment analysis, with high accuracy [83].

Other Features. Other features used in this classification task were based in techniques such as **Named Entity Recognition** (NER) [14], **Topic Extraction** [52], Word Sense Disambiguation Techniques to check **Polarity** [31, 58], frequencies of **personal pronouns** in the first and second person, the presence of **emoticons** [16, 22] and **capital letters** [16]. Before the feature extraction process, some studies have also used **stemming** and removed **stop-words** [7, 19, 52]. Characteristics of the message were also considered such as hashtags, mentions, retweets, URLs, number of tags, terms used in the tags, number of notes (reblog and like count), and link to multimedia content, such as image, video, or audio attached to the post [3].

5.4.2 Specific Hate Speech Detection Features. Complementary to the approaches commonly used in text mining analysis, several specific features are being used to tackle the problem of hate speech automatic detection. We briefly present the approaches found.

Othering Language. Othering has been used as a construct surrounding hate speech [9] and consists in analysing the contrast between different groups by looking at "Us versus Them." It describes "Our" characteristics as superior to "Theirs," which are inferior, undeserving, and incompatible [17]. Expressions like "send them home" show this cognitive process. Othering terms and language were identified using an implementation of the Stanford Lexical Parser, along with a context-free lexical parsing model, to extract typed dependencies [9]. Typed dependencies provide a representation of syntactic grammatical relationships in a sentence. For instance [7], in the tweet "Totally fed up with the way this country has turned into a heaven for terrorists. Send them all back Home," one resultant typed dependency is nsubj(home-5, the-2). This identifies the relationship nsubj, which is an abbreviation of the nominal subject between the fifth word "home" and the second word "them." The association between both words, is an example of "othering" phrase, because the opposition between "them" from "us," through the relational action of removing "them" to their "home" [7].

Perpetrator Characteristics. Some other studies also consider features more related with the social network graph. In this particular case [77], this study was linking the available messages from the same user and focusing on the user characteristics like gender and geographic localization.

Objectivity-Subjectivity of the Language. On one hand, some authors [31] argue that hate speech is related with more subjective communication. In this study, a rule-based approach is used to separate objective sentences from subjective ones and, after this step, erase the objective sentences from the analysis. On the other hand, there are authors [74] pointing out that in some cases prejudiced and hateful communication can be conducted recurring to scientifically worded essays. In this case, in some sites, they found that the anti-semitic speech was not presenting explicitly pejorative terms. Instead, it presented extremely anti-semitic ideologies and conclusions in a scientific manner. The differences found in both studies cited in this subsection point out that hate speech detection can occur in several forms. Therefore, it is important to understand what is contributing for its different expressions and how to include the plurality of the concept and several nuances in the developed model.

Declarations of Superiority of the Ingroup. In addition to the question of the objectivity and subjectivity of the language, declarations of the superiority of the ingroup can also be considered hate speech. In this case, hate speech can also be present when there are only defensive statements or declarations of pride, rather than attacks directed toward a specific group [74].

Focus on Particular Stereotypes. In some studies [74], authors hypothesize that hate speech often employs well-known stereotypes and therefore they subdivide such speech according to the

stereotypes. This approach can be useful, because each stereotype has a specific language: words, phrases, metaphors, and concepts. For instance, the anti-Hispanic speech might make reference to border crossing; anti-African American speech often references unemployment or single parent upbringing; and anti-Semitic language often refers to money, banking, and media [74]. Given this, creating a language model for each stereotype is a necessary prerequisite for building a general model for all hate speech [74]. In some other studies [67], authors also point out the different hate speech categories. They combine Hatebase [39] information along with the categories reported by FBI for hate crimes and they ended up with nine categories: Race, Behavior, Physical, Sexual Orientation, Class, Gender, Ethnicity, Disability, Religion, and "Other."

Intersectionism of Oppression. Intersectionality is a concept that points out the connection between several particular types of hate speech (e.g., burka prohibition can be analysed either as an Islamophobic or as a sexist behavior, since this symbol is used by Muslims, but just by women). Intersectionism of several kinds of oppressions presents a particular challenge for the automated identification of hate speech and it has been considered in the literature. In a particular study [9], the intersection of subtypes of hate is considered only in the evaluation of the model, where more than one class was regarded at the same time.

5.4.3 Summary from the Text Mining Approaches. In this section, we tried to understand which specific features have been used in hate speech detection and related concepts. The different studies used several features, and in some cases, the conclusions seem contradictory. The results of the categorization conducted are summarized in Figures 8 and 9.

5.5 Main Conclusions from the Systematic Literature Review

We conducted a systematic literature review to understand the state of the art and opportunities in the field of automatic hate speech detection. This proved to be a challenging task, mostly because this topic has been widely discussed in other fields, such as social sciences and law, and therefore we found a large number of documents that would require more resources to process. For solving this problem, we focused only on the documents from computer science and engineering, and we concluded that the number of articles has been increasing in the last years. However, at the same time, it is possible to notice that this area remains in an early phase. The existing papers are published in a wide range of venues, not specific for hate speech, and the few conferences toward this topic that exist are having now its first editions. Besides, the majority of the papers found also have a low number of citations.

Regarding the practical work conducted, hate speech is being analysed in connection with other related concepts, specifically social media and machine learning. From the possible approaches from machine learning, automatic identification of hate speech is being tackled as a classification task. The wide majority of the studies considers this a binary classification problem (hate speech messages vs. not hate speech messages). However, a few have also used a multiclass approach, where racism is one of the classes more regarded. In the majority of the works, researchers collect new datasets. Twitter is the preferred social network, and English the most common language. We concluded that authors do not use public datasets and do not publish the new ones they collect. This makes very difficult to compare results and conclusions. Comparative studies and surveys are also scarce in the area. Finally, regarding the features used, we observed that the majority of the studies consider general approaches of text mining and do not use particular features for hate speech.

6 RESOURCES FOR HATE SPEECH CLASSIFICATION

In the conducted literature review, some resources were found. In this section, we present the datasets and open source projects.

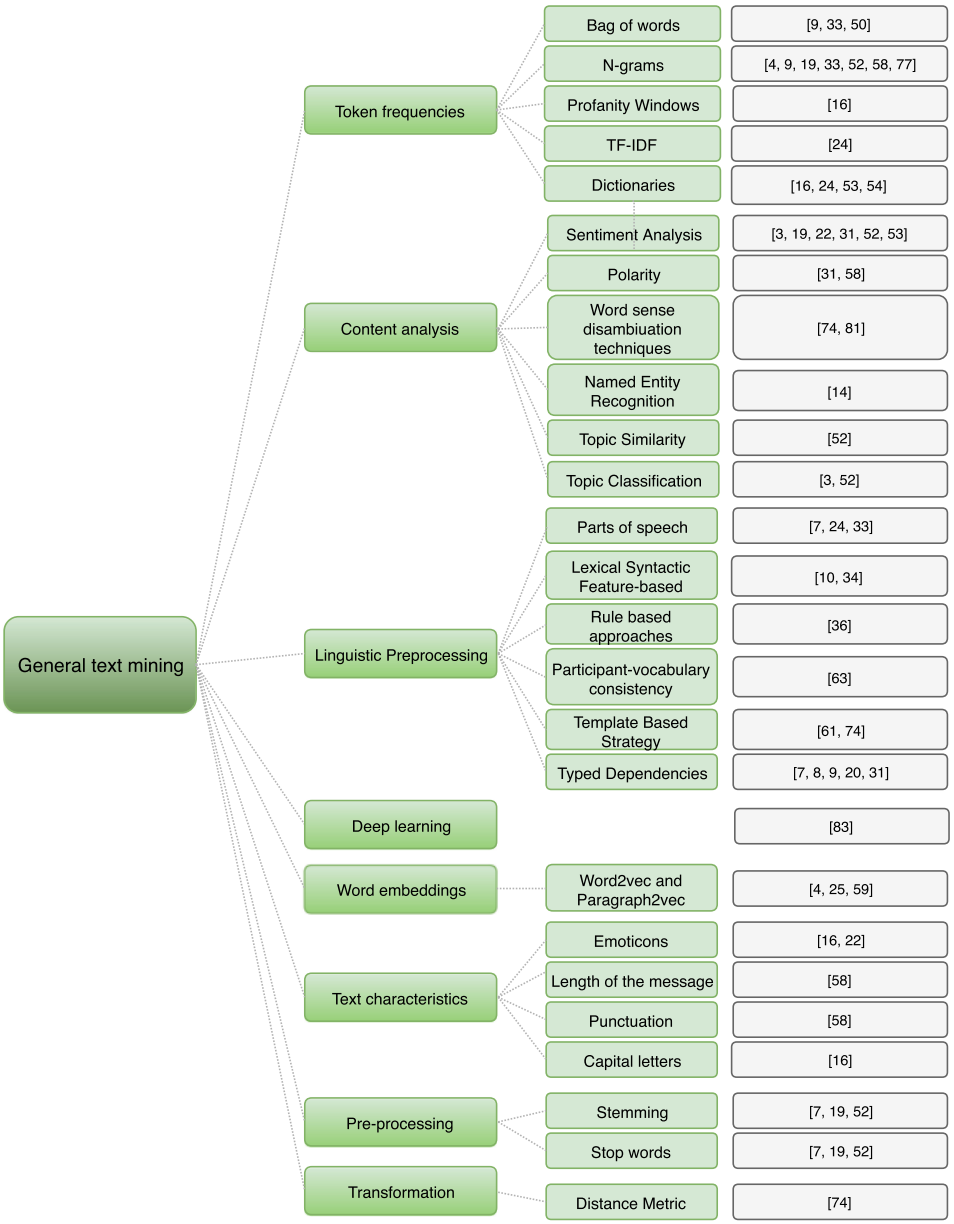

Fig. 8. Papers using generic text mining features.

6.1 Datasets for Hate Speech Classification

Regarding the datasets and corpus found, we summarize the main information in Table 14. Despite the fact that some datasets and corpus for hate speech already exist, there are no established ones.

6.2 Open Source Projects for Hate Speech Automatic Detection

We had the goal to check if there are any projects available for hate speech automatic detection that can be used as examples or sources for annotated data. For this we inspected GitHub using

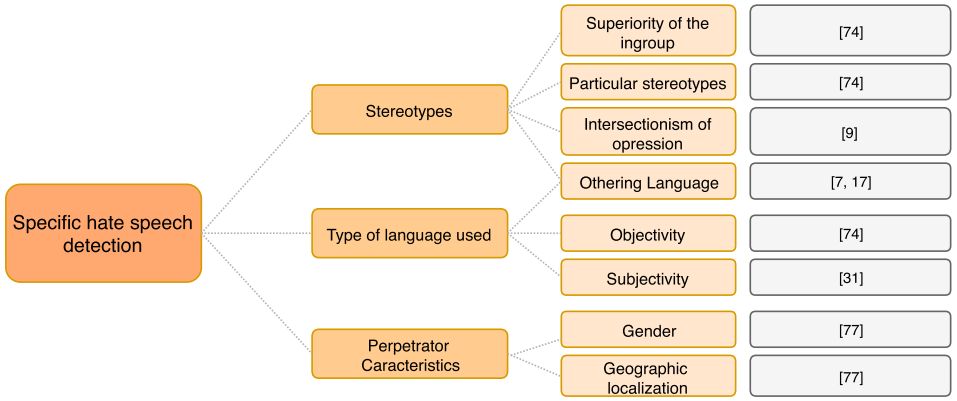

Fig. 9. Papers using specific hate speech detection features.

Table 14. Datasets and Corpus for Hate Speech Detection

Name	Distribution	Year	Type	Number of instances	Classes Used	Language	Ref.
Hate Speech Twitter annotations	GitHub repository	2016	Dataset	16,914	Sexist, racist	English	[76]
Hate Speech identification	Available for the community	2015	Dataset	14,510	Offensive with hate speech, offensive with no hate speech, not offensive	English	[15]
Abusive language dataset	Not available	2016	Dataset	2,000	Hate speech, not offensive	English	[80]
German Hatespeech Refugees	Creative Commons Attribution-ShareAlike 3.0 Unported License	2016	Dataset	470	Hate speech, not offensive	German	[73]
Hatebase	Available for the community	2017	Corpus	-	-	Universal	[44]
Hades	Available for the community	2016	Corpus	-	-	Dutch	[12]
Hate Speech and offensive language	Available for the community	2017	Corpus	-	-	English	[18]

the expression "hate speech" in the available search engine. The search for projects in GitHub occurred in May 2017. We found 25 repositories with some content. We describe here the main conclusions from this search.

6.2.1 The Type of Approach. We manually classified the type of approach followed in the projects (Figure 10). In this figure, we can see that we found projects that aimed to classify text excerpts as containing hate speech (Classification), to crawl messages in social networks (Crawling), analyse sentiment level in messages (Sentiment evaluation), conduct latent semantic analysis

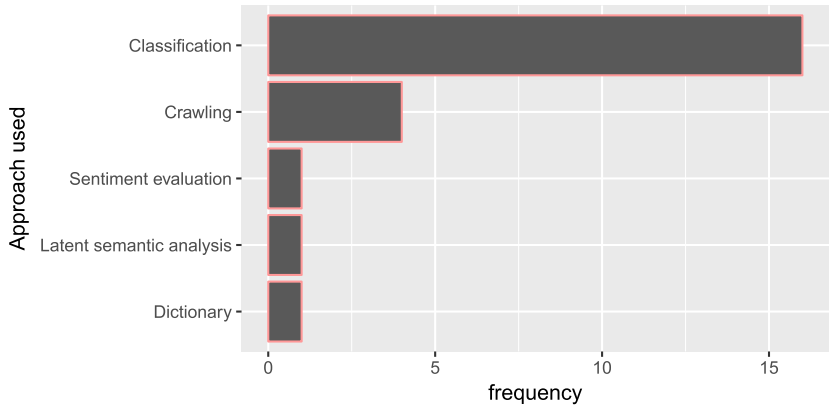

Fig. 10. Approaches used in open source projects about hate speech.

(Latent semantic analysis), and build dictionaries (Dictionary). We conclude that the majority of the projects are concerned with building models and classifying messages as hate speech. In what concerns to the programming languages, all projects were developed in Python, except three developed with JavaScript and Java.

6.2.2 Datasets Used in the GitHub Projects. The datasets used in the GitHub projects are analysed regarding its source, language, and availability. In what concerns to the source, the majority of projects works with Twitter data, and all the projects use messages in English, except two that use Dutch [71] or Finnish. These findings are congruent with our systematic literature review. Regarding availability, we were interested in access to new datasets or messages already annotated. We saw that two repositories provide some more datasets [27, 70]. However, we concluded that the majority of the projects does not provide any new data and there is also a few projects that use datasets already described in the Section 6.

7 RESEARCH CHALLENGES AND OPPORTUNITIES

Hate speech is a complex phenomenon and its detection problematic. Some challenges and difficulties were highlighted by the authors of the surveyed papers:

- Low agreement in hate speech classification by humans, indicating that this classification would be harder for machines [50].
- The task requires expertise about culture and social structure [63].
- The evolution of social phenomena and language makes it difficult to track all racial and minority insults [58].
- Language evolves quickly, in particular among young populations that communicate frequently in social networks [63].
- Despite the offensive nature of hate speech, abusive language may be very fluent and grammatically correct, can cross sentence boundaries, and the use of sarcasm in it is also common [58].
- Finally, hate speech detection is more than simple keyword spotting [58].

We find it relevant to present those difficulties, so that we bear in mind the kind of challenges that researchers face in their work.

It is also important to point out that the systematic literature review conducted allowed us to identify opportunities in this field.

Open Source Platforms or Algorithms. In our systematic literature review, we found that the documents describe methods, features extracted, and algorithms used. However, it is rare to find works with open source code. More sharing of code, algorithms, processes for feature extraction and platforms can help the area to evolve more quickly.

Definition of a Main Dataset. In this field, there are no commonly adopted datasets, even for research in English. That would be an important step in making it easier the comparison between the different studies.

Comparative Studies. Also, comparative studies, where the different approaches are used, are missing, making hard to understand which processes for feature extraction and algorithms are more efficient in tackling the problem of hate speech detection.

Multilingual Research. As we described previously, the majority of the studies uses datasets in English. Besides, only isolated studies were conducted in German, Dutch and Italian. In this case, research in other languages commonly used on the internet is also needed (e.g., French, Mandarin, Portuguese, Spanish).

As a result, our research has manifested the above mentioned opportunities that will strengthen the development of hate speech detection.

8 CRITICAL DISCUSSION AND CONCLUSIONS

In this survey, we presented a critical overview on how the automatic detection of hate speech in text has evolved over the past years. First, we analysed the concept of hate speech in different contexts, from social networks platforms to other organizations. Based on our analysis, we proposed a unified and clearer definition of this concept that can help to build a model for automatic detection of hate speech. Additionally, we presented examples and rules for classification found in the literature, together with the arguments in favor or against those rules. Our critical view pointed out that we have a more inclusive and general definition about hate speech than other perspectives found in the literature. This is the case, because we propose that subtle forms of discrimination on the internet and online social networks should also be spotted. With our analysis, we also concluded that it would be important to compare hate speech with cyberbullying, abusive language, discrimination, toxicity, flaming, extremism and radicalization. Our comparison showed how hate speech is distinct from these related concepts and helped us to understand the limits and nuances of its definition.

Through a systematic literature review, we concluded that there are not many studies and papers published in automatic hate speech detection from a computer science and informatics perspective. In general, the existing works regard the problem as a machine learning classification task. In this field, researchers tend to start by collecting and annotating new messages, and often these datasets remain private. This slows down the progress of the research, because less data is available, making it more difficult to compare results from different studies. Nevertheless, we found three available datasets, in English and German. Additionally, we compared the diverse studies using algorithms for hate speech detection, and we rank them in terms of performance. Our goal was to reach conclusions about which approaches are being more successful. However, and in part due to the lack of standard datasets, we find that there is no particular approach proving to reach better results among the several articles.

Regarding the features used in these studies, we classified them in terms of general text mining approaches and specific approaches for hate speech. For the first, those are mainly N-grams, POS,

rule-based approaches, sentiment analysis, and deep learning. For the specific hate speech detection features, we found mainly othering language, the superiority of the in-group, and focus on stereotypes. Besides, we observed that the majority of the studies only considers generic features and do not use particular features for hate speech. This can be problematic, because hate speech is a complex social phenomenon in constant evolution and supported in language nuances.

Finally, we identified challenges and opportunities in this field, namely the scarcity of open source code and platforms that automatically classify hate speech; the lack of comparative studies that evaluate the existing approaches; and the absence of studies in languages other than English. With our work, we summarized and established the current state of the automatic hate speech detection field. Undoubtedly, this is an area of profound societal impact and with many open research challenges.

REFERENCES

[1] ACL. 2017. ALW1: 1st workshop on abusive language online. Retrieved from https://sites.google.com/site/abusivelanguageworkshop2017/home.

[2] Swati Agarwal and Ashish Sureka. 2015. Using KNN and SVM based one-class classifier for detecting online radicalization on Twitter. In *Proceedings of the International Conference on Distributed Computing and Internet Technology*. Springer, 431–442.

[3] Swati Agarwal and Ashish Sureka. 2017. Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website. *arXiv Preprint arXiv:1701.04931* (2017).

[4] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 759–760.

[5] Tanvi Banerjee, Amir H. Yazdavar, Andrew Hampton, Hemant Purohit, Valerie L. Shalin, and Amit P. Sheth. Identifying pragmatic functions in social media indicative of gender-based violence beliefs. *Manuscript Submitted for Publication*.

[6] Jamie Bartlett, Richard Norrie, Sofia Patel, Rebekka Rumpel, and Simon Wibberley. 2014. *Misogyny on Twitter*. Technical Report. Demos.

[7] Peter Burnap and Matthew L. Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on Twitter: Interpretation and communication for policy decision making. In *Proceedings of the Conference on the Internet, Policy & Politics*. 1–18.

[8] Pete Burnap and Matthew L. Williams. 2015. Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy Internet* 7, 2 (2015), 223–242.

[9] Pete Burnap and Matthew L. Williams. 2016. Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Sci.* 5, 1 (2016), 11.

[10] Ying Chen. 2011. *Detecting Offensive Language in Social Medias for Protection of Adolescent Online Safety*. Ph.D. Dissertation. The Pennsylvania State University.

[11] CHI2017. 2017. 2017 Workshop on online harassment. Retrieved from http://social.umd.edu/woh/.

[12] CLiPS. 2016. HADES. Retrieved from https://github.com/clips/hades.

[13] CONTACT. 2017. Interdisciplinary conference on Hate speech. Definitions, Interpretations and Practices. Retrieved from https://sites.google.com/site/abusivelanguageworkshop2017/home.

[14] Keith Cortis and Siegfried Handschuh. 2015. Analysis of cyberbullying tweets in trending world events. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*. ACM, 7.

[15] CrowdFlower. 2017. Data for everyone. Retrieved from https://www.crowdflower.com/data-for-everyone/.

[16] Maral Dadvar, Franciska de Jong, Roeland Ordelman, and Dolf Trieschnigg. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the 12th Dutch-Belgian Information Retrieval Workshop*. University of Ghent, 23–25.

[17] Ali A. Dashti, Ali A. Al-Kandari, and Hamed H. Al-Abdullah. 2015. The influence of sectarian and tribal discourse in newspapers readers' online comments about freedom of expression, censorship and national unity in Kuwait. *Telemat. Informat.* 32, 2 (2015), 245–253.

[18] Thomas Davidson. 2017. Automated hate speech detection and the problem of offensive language. Retrieved from https://github.com/t-davidson/hate-speech-and-offensive-language.

[19] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *arXiv Preprint arXiv:1703.04009* (2017).

[20] Marie-Catherine De Marneffe and Christopher D. Manning. 2008. *Stanford Typed Dependencies Manual.* Technical report, Stanford University.
[21] Guy De Pauw, Ben Verhoeven, Bart Desmet, and Els Lefever. 2016. First workshop on text analytics for cybersecurity and online safety (TA-COS 2016). In *Proceedings of the 1st Workshop on Text Analytics for Cybersecurity and Online Safety (TACOS'16), collocated with the 10th International Conference on Language Resources and Evaluation (LREC'16).* European Language Resources Association.
[22] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on Facebook. In *Proceedings of the 1st Italian Conference on Cybersecurity.* 86–95.
[23] Cambridge Dictionary. 2017. Profanity. Retrieved from https://dictionary.cambridge.org/dictionary/english/profanity.
[24] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. *Soc. Mobile Web* 11, 02 (2011).
[25] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web.* ACM, 29–30.
[26] Sara Douglass, Sheena Mirpuri, Devin English, and Tiffany Yip. 2016. They were just making jokes: Ethnic/racial teasing and discrimination among adolescents. *Cultur. Divers. Ethnic Minor. Psychol.* 22, 1 (2016), 69.
[27] Eyspahn. 2016. Online hate speech modeling using Python and reddit comment data. Retrieved from https://github.com/eyspahn/OnlineHateSpeech_PyLadiesSea.
[28] Facebook. 2013. What does Facebook consider to be hate speech? Retrieved from https://www.facebook.com/help/135402139904490.
[29] FBI. 2015. 2015 hate crime statistics. Retrieved from https://ucr.fbi.gov/hate-crime/.
[30] Fabio Giblietto and Yenn Lee. 2015. To be or not to be Charlie: Twitter hashtags as a discourse and counter-discourse in the aftermath of the 2015 Charlie Hebdo shooting in France. In *Proceedings of the 4th Workshop on Making Sense of Microposts (#Microposts'14).*
[31] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *Int. J. Multimedia Ubiq. Eng.* 10, 4 (2015), 215–230.
[32] Edel Greevy. 2004. *Automatic Text Categorisation of Racist Webpages.* Ph.D. Dissertation. Dublin City University.
[33] Edel Greevy and Alan F. Smeaton. 2004. Classifying racist texts using a support vector machine. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 468–469.
[34] Stanford NLP Group. 2017. The Stanford NLP Group. Retrieved from http://nlp.stanford.edu/.
[35] Radhouane Guermazi, Mohamed Hammami, and Abdelmajid Ben Hamadou. 2007. Using a semi-automatic keyword dictionary for improving violent Web site filtering. In *Proceedings of the 3rd International IEEE Conference on Signal-Image Technologies and Internet-Based System (SITIS'07).* IEEE, 337–344.
[36] Yannis Haralambous and Philippe Lenca. 2014. Text classification using association rules, dependency pruning and hyperonymization. *arXiv Preprint arXiv:1407.7357* (2014).
[37] Reporting Hate. 2017. Hate speech conference I.H.D.I.P. Retrieved from http://reportinghate.eu/contact2017/.
[38] No hate speech movement. 2017. No hate speech movement. Retrieved from https://www.nohatespeechmovement.org/.
[39] Hatebase. 2017. Hatebase. Retrieved from https://www.hatebase.org/.
[40] Alex Hern. 2016. Facebook, YouTube, Twitter, and Microsoft sign EU hate speech code. Retrieved from https://www.theguardian.com/technology/2016/may/31/facebook-youtube-twitter-microsoft-eu-hate-speech-code.
[41] Sarah Hewitt, Thanassis Tiropanis, and Christian Bokhove. 2016. The problem of identifying misogynist language on Twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science.* ACM, 333–335.
[42] ILGA. 2016. Hate crime and hate speech. Retrieved from http://www.ilga-europe.org/what-we-do/our-advocacy-work/hate-crime-hate-speech.
[43] Jigsaw. 2017. Perspective API. Retrieved from https://www.perspectiveapi.com/.
[44] Kaggle. 2013. Detecting insults in social commentary. Retrieved from https://www.kaggle.com/c/detecting-insults-in-social-commentary/data.
[45] Panos Kompatsiaris. 2016. Whitewashing the nation: Racist jokes and the construction of the african "other" in Greek popular cinema. *Soc. Ident.* 23, 3 (2016), 360–375.
[46] Ivana Kottasová. 2017. Europe says Twitter is failing to remove hate speech. Retrieved from http://money.cnn.com/2017/06/01/technology/twitter-facebook-hate-speech-europe/index.html.
[47] Till Krause and Hannes Grassegger. 2016. Facebook's secret rules of deletion. Retrieved from http://international.sueddeutsche.de/post/154543271930/facebooks-secret-rules-of-deletion.
[48] Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology.* Sage.

[49] Giselinde Kuipers and Barbara van der Ent. 2016. The seriousness of ethnic jokes: Ethnic humor and social change in The Netherlands, 1995–2012. *Humor* 29, 4 (2016), 605–633.
[50] Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the Association for the Advancement of Artificial Intelligence*.
[51] Anti-Defamation League. 2015. The trap of masculinity: how sexism impacts boys and men. Retrieved from https://www.adl.org/sites/default/files/documents/assets/pdf/education-outreach/trap-of-masculinity.pdf.
[52] Shuhua Liu and Thomas Forss. 2014. Combining N-gram based similarity analysis with sentiment analysis in web content classification. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. 530–537.
[53] Shuhua Liu and Thomas Forss. 2015. New classification models for detecting hate and violence web content. In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K'15)*, Vol. 1. IEEE, 487–495.
[54] Wilson Jeffrey Maloba. 2014. *Use of Regular Expressions for Multi-lingual Detection of Hate Speech in Kenya*. Ph.D. Dissertation. iLabAfrica.
[55] Lacy G. McNamee, Brittany L. Peterson, and Jorge Peña. 2010. A call to educate, participate, invoke. and indict: Understanding the communication of online hate groups. *Commun. Monogr.* 77, 2 (2010), 257–280.
[56] Yashar Mehdad and Joel Tetreault. 2016. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial'16)*. 299–303.
[57] B. Nandhini and J. I. Sheeba. 2015. Cyberbullying detection and classification using information retrieval algorithm. In *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET'15)*. ACM, 20.
[58] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 145–153.
[59] Andre Oboler and Karen Connelly. 2014. Hate speech: A quality of service challenge. In *Proceedings of the IEEE Conference on e-Learning, e-Management, and e-Services (IC3e'14)*. IEEE, 117–121.
[60] United Nations Alliance of Civilizations (UNAOC). 2017. #SpreadNoHate: A global dialogue on hate speech against migrants and refugees in the media. Retrieved from https://www.unaoc.org/what-we-do/projects/hate-speech/.
[61] David Martin Powers. 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Int. J. Mach. Learn. Technol.* 2, 1 (2011), 37–63.
[62] Sheryl Prentice, Paul J. Taylor, Paul Rayson, Andrew Hoskins, and Ben O'Loughlin. 2011. Analyzing the semantic content and persuasive composition of extremist media: A case study of texts produced during the Gaza conflict. *Info. Syst. Front.* 13, 1 (2011), 61–73.
[63] Elaheh Raisi and Bert Huang. 2016. Cyberbullying identification using participant-vocabulary consistency. *arXiv Preprint arXiv:1606.08084* (2016).
[64] Vasu Reddy. 2002. Perverts and sodomites: Homophobia as hate speech in Africa. *South. African Linguist. Appl. Lang. Studies* 20, 3 (2002), 163–175.
[65] Bjorn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv Preprint arXiv:1701.08118* (2017).
[66] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Workshop on Natural Language Processing for Social Media (SocialNLP'17)*. 1.
[67] Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. *arXiv Preprint arXiv:1603.07709* (2016).
[68] Natalya Tarasova. 2016. *Classification of Hate Tweets and Their Reasons using SVM*. Master's thesis. Uppsala Universitet.
[69] Neil Thompson. 2016. *Anti-discriminatory Practice: Equality, Diversity and Social Justice*. Palgrave Macmillan.
[70] Annie Thorburn. 2016. Hate Speech ML. Retrieved from https://github.com/anniethorburn/Hate-Speech-ML.
[71] Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A dictionary-based approach to racism detection in dutch social media. *arXiv Preprint arXiv:1608.08738* (2016).
[72] Twitter. 2017. The Twitter Rules. Retrieved from https://support.twitter.com/articles/.
[73] UCSM. 2016. IWG hatespeech public. Retrieved from https://github.com/UCSM-DUE/.
[74] William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, 19–26.
[75] Zeerak Waseem. 2016. Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the 1st Workshop on Natural Language Processing and Computational Social Science*. 138–142.

[76] Zeerak Waseem. 2016. Hate speech Twitter annotations. Retrieved from https://github.com/ZeerakW/hatespeech.
[77] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'16)*. 88–93.
[78] Mike Wendling. 2015. 2015: The year that angry won the internet. Retrieved from http://www.bbc.com/news/blogs-trending-35111707.
[79] Christian Wigand and Melanie Voin. 2017. Speech by Commissioner Jourová—10 years of the EU Fundamental Rights Agency: A call to action in defence of fundamental rights, democracy and the rule of law. Retrieved from http://europa.eu/rapid/press-release_SPEECH-17-403_en.htm.
[80] Yahoo! 2017. Webscope datasets. Retrieved from https://webscope.sandbox.yahoo.com/.
[81] David Yarowsky. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 88–95.
[82] Youtube. 2017. Hate speech. Retrieved from https://support.google.com/youtube/answer/2801939?hl=en.
[83] Shuhan Yuan, Xintao Wu, and Yang Xiang. 2016. A two phase deep learning model for identifying discrimination from tweets. In *Proceedings of the International Conference on Extending Database Technology*. 696–697.
[84] Matthew Zook. 2012. Mapping racist tweets in response to President Obama's re-election. Retrieved from https://www.theguardian.com/news/datablog/2012/nov/09/mapping-racist-tweets-president-obama-reelection.

Received October 2017; revised May 2018; accepted June 2018