Building a formal model for hate detection in French corpora

Delphine Battistelli, Cyril Bruneau, Valentina Dragos

▶ **To cite this version:**

Delphine Battistelli, Cyril Bruneau, Valentina Dragos. Building a formal model for hate detection in French corpora. Procedia Computer Science, 2020, Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020, 176, pp.2358 - 2365. 10.1016/j.procs.2020.09.299 . hal-03184124

HAL Id: hal-03184124
https://hal.science/hal-03184124

Submitted on 29 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 176 (2020) 2358–2365

Procedia Computer Science

www.elsevier.com/locate/procedia

24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Building a formal model for hate detection in French corpora

Delphine Battistelli[a], Cyril Bruneau[a], Valentina Dragos[b,*]

[a]*MODYCO, CNRS-Paris Nanterre University, 92001 Nanterre, France*
[b]*ONERA, The French Aerospace Lab, 91123, Palaiseau, France*

Abstract

This paper investigates the development of a formal model in order to analyse online hate in French corpora. Relevant concepts are identified by exploiting several sources: the cognitive foundations of the appraisal theory, according to which people's emotional response are based on their own evaluative judgments or appraisals of situations, events or objects; a linguistic model of how different kinds of modalities applied to predicative contents are expressed in textual data; several definitions of a hate speech. Based on those inputs, a formal model was developed to describe online hate speech. The model highlights different categories of hate targets and actions, and emphasizes the importance of context for online hate detection.

© 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the KES International.

Keywords: online hate speech ; ontology ; semantics

1. Introduction

The Internet's speed and reach makes it a perfect place for online hate proliferation.Online hate is not a new paradigm [4] and over the past years, interest in online hate detection has continuously increased. From a security perspectives, intelligence analysts are still struggling to understand emergent online phenomena, including online hate but also propaganda and extremist ideas, which are grounded on citizen's feelings towards today's stories and events. Detection of hate is also part of risk prevention strategies for security and defense practitioners. For those domains, there is a practical need to develop semi-automatic approaches for online hate investigation.

Different methods have been applied to reduce hatred messages, including counter speech [29] or mandatory registration [13]. The development of automatic procedures was also largely considered, given the sheer volumes of online data and the inherent difficulties of manual moderation of thousands of comments. Methods developed for online hate detection can be roughly divided into lexicon-based methods, exploiting lexicons or more complex ontologies, and machine-learning methods based on manually annotated corpora [10]. In spite of those research efforts, online hate

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000.
E-mail address: valentina.dragos@onera.fr

1877-0509 © 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the KES International.
10.1016/j.procs.2020.09.299

remains a highly topical research problem with major societal importance. One reason is that detection of online hate requires techniques able to understand the types and targets of hate speech, although effective methods for those tasks are still needed.

This paper reports ongoing research conducted within the frame of a project aiming at investigating extremist contents in French blogosphera. The research presented in this paper is intended to develop a methodology to build an ontology to formalize online hate in textual content in French and is motivated by the lack of resources for French corpora analysis and the interconnection of haltered and extremists contents. In the context of this work, online hate is propagated through hate speech. The paper focuses on modeling development aspects, while practical use cases for annotation of texts or training of classifiers are not addressed. It should be noted that there are very few corpora and semantic resources for French in the more general area of social data analysis [1], [27], [20] contrary to what was done for other languages [23] , [8]. Another contribution of this paper is to develop resources to analyse online contents in French.

The rest of the paper is structured as follows: section 2 discusses related approaches and contextualises the work presented in the paper; section 3 introduces the methodology and resources used to build the model, which is then described in section 4. Critical cases and limitations of hate detection with this model are illustrated in section 5. Conclusion and directions for future work are discussed in section 6.

2. Related work

Research efforts conducted for hate detection are divers and work on a variety of nonstandardized datasets, so that often the authors collect and eventually annotate their own corpora. The detection of hate speech has been tackled in three main directions: building databases and resources, detecting the binary distinction of hate versus non hate content and identifying more specifically subcategories of hate according to the target.

Some studies aimed at building repositories of hatred messages to feed technological blocks. Among the resources available online, Hurtlex [1] is a multilingual computational lexicon of hate words and Hatebase [2] is an online repository of structured, multilingual and usage-oriented hate speech. The base is structured thanks to a vocabulary composed of eight categories, namely : archaic, class, disability, ethnicity, gender, nationality, religion, and sexual orientation. Another collection available online is the Kaggle's Toxic Comment Classification Challenge dataset [3], consisting of around 150k Wikipedia comments annotated for toxic behaviour. More specifically, a date set of hateful comments from a white supremacy forum is described in [12]. Hatebase and its vocabulary was used by Serra and colleagues to build classifiers for hate speech [25]. A variety of techniques were used for this binary classification of hateful non hateful content, including machine learning [17], classification with a mix of features such as images and emojis [19]. Several approaches also tackled the classification of hate content according to finer categories, such as abusive [11] or aggressive [14]. language, racism [21] and extremist contents[2].

From a different perspective, Silva and colleagues investigated the detection of hate targets in messages gleaned on Whisper and Twitter by exploiting the structure of sentences [26] and a characterization of directed and generalized types of hate with respect to the target is described in [9].

Most current methods and techniques developed for online hate selection rely strongly upon linguistic inputs, making use of blacklists, regular expressions, structures of sentences, annotated corpora or lexicons. Those techniques have specific performances and drawbacks: while keyword can be powerful indicators for hateful comments, alone they are not enough to detect all variants of hate speech and often dictionary-based approaches identify accurately the target but fail to distinguish hateful sentences from clean ones. Moreover the annotation of hate speech can be sometimes fairly vague [24] or affected by bias [5]. To overcome some limitations, Salminem and colleagues developed a taxonomy of different types and targets of online hate and then used this taxonomy to train a robust classifier for hate detection [22].

This paper tackles development of a formal model for hate detection in French corpora. The methodology adopted is guided by several definitions of hate categories and exploits two existing ontological resources in order to identify

[1] https://vocabularyserver.com/hurtlex/
[2] https://www.hatebase.org/
[3] https://www.kaggle.com/c/jigsaw-toxic-comment- classification-challenge/data

the French terms related to finer categories of hate discourse and hate targets. Two distinct corpora are used for refinements and validation purposes. Our purpose in modeling this ontology is to highlight the determinants of online hate speech and to depict the interconnectedness of those determinants. The ontology is grounded on the cognitive foundations of the appraisal theory [28], augmented with a set of concepts modeled to capture the context of statements and the commitment of authors towards their own assertions.

3. Resources and methodology

The construction methodology consists of three main phases: acquisition of knowledge, construction of a conceptual model and its consolidation, see Fig. 3. During the acquisition phase, three definitions for hate speech were

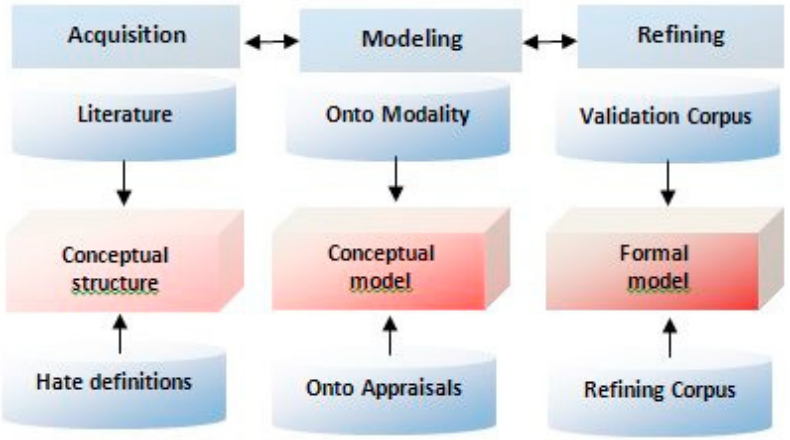

Fig. 1. Methodology and resources

selected in order to gain better understanding of hate semantics in different communities.

The first definition is proposed by Nockleby [18], who defines hate speech as an *abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender or sexual orientation*. The second definition is given by Mondal et al. who define hate speech as *an offensive post, motivated, in whole or in a part, by the writer's bias against an aspect of a group of people* [16]. The third definition selected for this work is developed jointly by Facebook, Twitter and Google; an online hate speech is defined as *a type of speech that takes place online, generally social media or the internet, with the purpose of attacking a person or a group on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender.*

In the light of definitions above, we consider hate speech as a combination of four main elements : an enunciator - or author - of a hateful speech, a hateful action, a target and the linguistic context of validation of the hateful action. The result of this acquisition phase is a set of main concepts called conceptual structure. During the representation phase, this initial structure is enriched by exploiting two existing ontologies.

The ontology of modalities [3] was developed in order to assess the validity of the knowledge extracted from textual content in the scientific literature but is not specific to this kind of textual data. The ontology describes *the linguistic context of validation* of an item of information, see 2. More specifically, the model represents the relations between three linguistic categories involved in the characterization of validity conditions of information: enunciative conditions, aspect and temporal conditions and also modal (or rhetorical) conditions.

This ontology of modalities models several concepts that are relevant for the purposes of this work. First, the root concept of this ontology, ContextOValidation and his two sub concepts, the LinguisticContextOfValidation and the ExtraLinguisticContextOfValidation, respectively. Moreover, the LinguisticContextofValidation deals with the knowledge that is relative to the context of validation as expressed by the author of the text, whereas the ExtraLinguistic-ContextofValidation looks at external knowledge or resources.

Taking into account the enunciative aspect enables us to identify whether the author is fully commitment to items in the sentences, or whether he just provides items attributed to another enunciator, with a marker of agreement or

Fig. 2. Ontology of modalities

disagreement. The epistemic and temporal aspects captured by the ontology of modalities, provide additional attributes of the validation context, for example by capturing whether actions are temporally in the past time or in the present time, or by capturing degrees of certainty or uncertainty of actions from the point of view of the enunciator.

The aspects captured by the ontology of modalities are relevant for any domain.Thus this ontology is independent of any domain ontology and reusable for other applications. The model has been also used for analysing newswire texts for example [6]. The resource was built by focusing on the explicit linguistic markers of different kinds of epistemic qualification and it captures temporal, modal and enunciative and aspect -related features which are indicative of authorial commitment to the information conveyed by sentences.

The appraisal ontology [7] describes the evaluation language according to the principles of the appraisal theory. This cognitive framework introduces three systems called attitude, engagement and graduation, see fig. 3, to explain the construction of evaluative sentences.

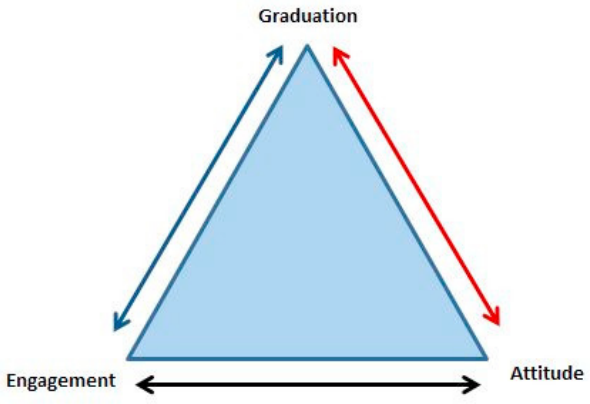

Fig. 3. Appraisal categories

The attitude system is related to linguistic expressions conveying the attitude of authors at the time they write a text. This system covers three main subcategories: affect, appreciation and judgement, discussed hereafter.

Affect is related to linguistic expressions of author's feelings such as happiness, joy, sadness, grief, etc.

Judgment highlights linguistic expressions conveying characterization of persons and behaviors by the author. Generally it conveys opinions and personal tastes about objects, such as nice, ugly, beautiful, shy but also about interactions and behaviors in the social context: heroic, brave, open-minded.

Appreciation is related to assessment and evaluations of entities, objects, events and scenes.

The *engagement system* gathers linguistic expression specifying the author's position with respect to his own statements. When reporting, writers often embed clues as to how strongly they support the content being conveyed and may indicate confidence, doubt, skepticism, conviction, etc., about the information reported. The engagement system is closely related to the notions of trust, confidence, probability or possibility. Categories under this system encompass aspects related to denial, concession, confirmation, endorsement, acknowledgement and distance.

The ontology of appraisals captures how humans interpret events -positive, negative- as well as their position, support and engagement with respect to their own interpretation and report – confidence, support, agreement, disagreement. From a linguistic standpoint, the model highlits how authors use linguistic expressions to communicate their emotional states and engagement.

The ontology of modalities and the appraisal ontology were jointly used during the modeling phase to build a novel ontology named 'ontology of hate'. The intuition behind the joint utilization of those resources is that, while the appraisal ontology can provide a set of concepts related to affect or judgement that are of interest fort hate modeling, the modalities ontology can facilitate the analysis of their linguistic context of validation, thanks mainly to concepts capturing modality and enunciativity features. Our assumption is that the analysis of this linguistic context helps recognizing reported speech, counter speech, negation, linguistic reappropriation, as well as several modal proccesses a locutor could use to distanciate himself from the statement, or limitate its strength.

From these two ontologies, only the relevant concepts for hate speech detection have been kept. At the end of this modeling phase, concepts are added to initial structure and a formal model is built.

The last phase refines the formal model by using two distinct corpora collected on social media. A first corpus was used to manually refine the concepts of the ontology, and this corpora was provide by external contributors. For the validation step, a corpus of tweets was collected using the Twitter API with a combination of relevant keywords. This corpus was collected based on query terms, without manual processing or tagging. The data set is composed of around 300 tweets in French that have been manually selected for their compatibility with hate speech definitions.

Data sets collected for this work are relatively small, especially as a result of the selection procedures undertaken to keep only content which is relevant for hate analysis.

4. Development and description of the model

The ontology was created from scratch, starting with concepts selected from the modalities and appraisal ontologies, and adding additional concepts to characterize online hate. The ontology of hate has four main concepts, see 4: Action (Action), Target (Cible), Context (Contexte) and Orientation (Orientation). The enunciator is part of the *Context* concept.

Action is the core concept to detect hateful content. It is further divided into instigation to violence (AppelALaViolence) or deprecatory judgement (JugementDévalorisant). This last concept is then composed of several sub-concepts extracted from the appraisal ontology. Affect (Affect) captures emotions of the author while Judgement (Jugement) describes assessments of persons and objects and is divided into: social esteem (EstimeSociale), which encompasses capacities (Capacité) e.g stupid, tenacity (Ténacité) e.g unreliable and normality (Normalité) e.g unpopular, and social sanction (SanctionSociale) which captures truthfulness (Véracité) e.g liar and propriety (Bienseance) e.g corrupt. Another sub-concept of *Action* is *Insult*, a concept regrouping insults which do not involve a specific evaluation of the insulted object but rather derogatory terms for groups of individuals, as opposed to appraisal-oriented insults, e.g jackass (imbécile), carrying specific judgment.

Target (Cible) classifies the targets of hateful statements by taking into account the following characteristics: Sexual Orientation, Ethnical Origin, National Origin, Sex, Religion, and Immigration Status.

Context is a larger umbrella for several concepts extracted from the modalities ontology. It describes: the aspectual context (ContexteValidationAspectuel) of Action given by the author and which can be for example a process, an event or a state ; the enunciating context (ContexteValidationEnonciatif) which can be Collectif (collective) in cases of factual statements and (so called) global or wide spread opinions or Individuel (individual) when the author or another

Fig. 4. Main concepts of the ontology

source cited by the author is implied in a hateful action; the temporal context (ContexteValidationTemporel), capturing the absolute or relative temporal of statements which can be in the past, the present or the future of the enuniciator; the modal context (ContexteValidationModal), which carries the author's engagements towards its statements.

Morever, the statement can describe several aspects: the desirability of the statement (Appreciatif), the will (Boulique), to what extent the statement is considered moral (Déontique), or the degree of uncertainty assigned to the statement or its negation (Epistémique).

Orientation captures the hatred-related orientation of speech, with three categories: hateful and not-hateful, and undetermined for ambiguous cases. In addition to those concepts, three relations enrich the structure of the ontology of hate: *hasTarget, hasOrientation* and *hasContext* capture the target, the orientation and the context of a statement, respectively. The ontology of hate is represented in OWL [15] and is composed of 61 concepts structured on a 8-levels hierarchy; the model also has 4 ObjectProperties.

5. Hate detection and limitations of the model

The ontology highlights the main concepts for hate detection and fig. 5 shows two examples where hatred message are detected thanks to the identification of hate-specific targets and actions, respectively.

The main purpose of this work is to develop a formal model for online hate analysis and detection. Practical applications for hate detection will require annotation procedures, through which concepts of the ontology will be related to textual segments. However, practical cases using the ontology of hate should take into account its limitations and the fact that hate speech is contextually embedded, so that comments that in one community are perceived and hateful or offensive are not so in another community [21].

From a practical perspective, detection of specific targets, identification of implicit targets, analysis of irony and sarcasm are the main technical challenges for effective hate speech detection. The following examples, collected from Twitter from February to May 2020, illustrate those challenges.

Political opinions are not part of hate speech characteristics and the example, despite its inherent violence, should thus not be considered as a hateful speech: *Oui! MLP.Philipe Vardon et GÉNÉRATION IDENTITAIRE. Contre la pourriture Estrosi UMPS - LR.*.

The presence of hate target but without violent action should be considered, again, as non hateful. *Quand il s'agit d'entrenir les étrangers parasites on est toujours sûr de trouver la gauche*.

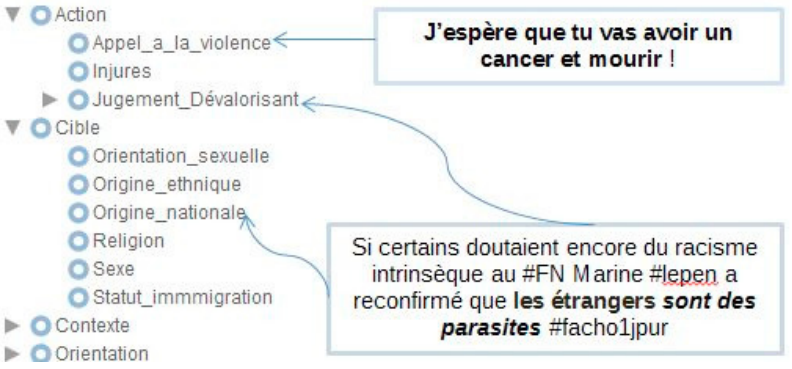

Fig. 5. Examples of hate detection

The difficulty of detecting implicit or tacit targets is illustrated by the next example: *Attendrons-nous qu'il soit trop tard ? Dans ma profession de foi écrite pour la même campagne législative de 1975 : Les Français ne sont ni xénophobes ni racistes. [...] Mais ils ne supportent plus que la France soit colonisée, exploitée, terrorisée.*

Moreover, sarcasm and irony potentially flip the hateful/no hateful orientation of sentences, as shown fereatfer: *Que se passe-t-il ? Allah est grand ou il n'est pas grand ? Faudrait savoir tout de même ! Il a quand même créé la terre plate !*

Examples above underline the importance of context and the difficulty of lexical analysis to detect between hateful and non hateful contents.

6. Conclusion and future work

This paper presents ongoing efforts intended to build a formal model for hate detection in French messages. It discusses the methodology and resources used to build the ontology and a corpora-driven mechanism for empirical validation of the model. The ontology models the main concepts needed to analyse hate speech. The model is intended to support the development of hate detection approaches able to overcome the limitations of existing approaches, which tend to rely heavily on the use of dictionaries of hateful lexical units or coarse categorizations of hate-related concepts. The next step is the evaluation of this ontology in the context of realistic use cases and to design procedures for the practical exploitation of the model, such as annotation of corpora and training of classifiers.

Acknowledgements

This document has been produced in the context of the FLYER project funded by the ASTRID research program.

References

[1] Abdaoui, A., Azé, J., Bringay, S., Poncelet, P., 2017. Feel: a french expanded emotion lexicon. Language Resources and Evaluation 51, 833–855.

[2] Badjatiya, P., Gupta, S., Gupta, M., Varma, V., 2017. Deep learning for hate speech detection in tweets, in: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 759–760.

[3] Battistelli, D., Amardeilh, F., 2009. Knowledge claims in scientific literature, uncertainty and semantic annotation: A case study in the biological domain, in: Proceedings of the Semantic Authoring, Annotation and Knowledge Markup Workshop (SAAKM 2009), Los Angeles, United States.

[4] Berlet, C., 2001. When hate went online, in: Northeast Sociological Association Spring Conference in April, Citeseer. pp. 1–20.

[5] Binns, R., Veale, M., Van Kleek, M., Shadbolt, N., 2017. Like trainer, like bot? inheritance of bias in algorithmic content moderation, in: International Conference on Social Informatics, Springer. pp. 405–415.

[6] Damiani, M., Battistelli, D., 2013. Enunciative and modal variations in newswire texts in french: From guideline to automatic annotation, in: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pp. 223–227.

[7] Dragos, V., Battistelli, D., Kelodjoue, E., 2018. Beyond sentiments and opinions: exploring social media with appraisal categories, in: 2018 21st International Conference on Information Fusion (FUSION), IEEE. pp. 1851–1858.
[8] El-Beze, M., Jackiewicz, A., Hunston, S., 2010. Opinions, sentiments et jugements d'évaluation. Traitement Automatique des Langues 51, 7–17.
[9] ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W.Y., Belding, E., 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media, in: Twelfth International AAAI Conference on Web and Social Media.
[10] Fortuna, P., Nunes, S., 2018. A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR) 51, 1–30.
[11] Founta, A.M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., Leontiadis, I., 2019. A unified deep learning architecture for abuse detection, in: Proceedings of the 10th ACM Conference on Web Science, pp. 105–114.
[12] de Gibert, O., Perez, N., García-Pablos, A., Cuadros, M., 2018. Hate speech dataset from a white supremacy forum. arXiv preprint arXiv:1809.04444 .
[13] Hughey, M.W., Daniels, J., 2013. Racist comments at online news sites: a methodological dilemma for discourse analysis. Media, Culture & Society 35, 332–347.
[14] Kumar, R., Ojha, A.K., Malmasi, S., Zampieri, M., 2018. Benchmarking aggression identification in social media, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pp. 1–11.
[15] McGuinness, D.L., Van Harmelen, F., et al., 2004. Owl web ontology language overview. W3C recommendation 10, 2004.
[16] Mondal, M., Silva, L.A., Benevenuto, F., 2017. A measurement study of hate speech in social media, in: Proceedings of the 28th ACM Conference on Hypertext and Social Media, pp. 85–94.
[17] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y., 2016. Abusive language detection in online user content, in: Proceedings of the 25th international conference on world wide web, pp. 145–153.
[18] Nockleby, J.T., 2000. Hate speech. Encyclopedia of the American constitution 3, 1277–1279.
[19] Pereira-Kohatsu, J.C., Quijano-Sánchez, L., Liberatore, F., Camacho-Collados, M., 2019. Detecting and monitoring hate speech in twitter. Sensors 19, 4654.
[20] Piolat, A., Bannour, R., 2009. Emotaix: un scénario de tropes pour l'identification automatisée du lexique émotionnel et affectif. LAnnee psychologique 109, 655–698.
[21] Saleem, H.M., Dillon, K.P., Benesch, S., Ruths, D., 2017. A web of hate: Tackling hateful speech in online social spaces. arXiv preprint arXiv:1709.10159 .
[22] Salminen, J., Almerekhi, H., Milenković, M., Jung, S.g., An, J., Kwak, H., Jansen, B.J., 2018. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media, in: Twelfth International AAAI Conference on Web and Social Media.
[23] Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., Stranisci, M., 2018. An italian twitter corpus of hate speech against immigrants, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).
[24] Schmidt, A., Wiegand, M., 2017. A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pp. 1–10.
[25] Serra, J., Leontiadis, I., Spathis, D., Stringhini, G., Blackburn, J., Vakali, A., 2017. Class-based prediction errors to detect hate speech with out-of-vocabulary words, in: Proceedings of the First Workshop on Abusive Language Online, pp. 36–40.
[26] Silva, L., Mondal, M., Correa, D., Benevenuto, F., Weber, I., 2016. Analyzing the targets of hate in online social media, in: Tenth International AAAI Conference on Web and Social Media.
[27] Vernier, M., Monceaux, L., Daille, B., Dubreil, E., 2009. Catégorisation des évaluations dans un corpus de blogs multi-domaine. Revue des Nouvelles Technologies de l'Information , 45–70.
[28] White, P.R., 2015. Appraisal theory. The international encyclopedia of language and social interaction , 1–7.
[29] Wright, L., Ruths, D., Dillon, K.P., Saleem, H.M., Benesch, S., 2017. Vectors for counterspeech on twitter, in: Proceedings of the first workshop on abusive language online, pp. 57–62.