Multilingual and Multitarget Hate Speech Detection in Tweets

Patricia Chiril, Farah Benamara, Véronique Moriceau, Marlène Coulomb-Gully, Abhishek Kumar

Multilingual and Multitarget Hate Speech Detection in Tweets

Patricia Chiril[1] Farah Benamara[1] Véronique Moriceau[1, 2]
Marlène Coulomb-Gully[3] Abhishek Kumar[4]
(1) IRIT, Université de Toulouse, France
(2) LIMSI, Univ. Paris Sud, Université Paris Saclay, France
(3) LERASS, Université de Toulouse, France
(4) Indian Institute of Science, India
patricia.chiril@irit.fr, benamara@irit.fr, moriceau@limsi.fr,
marlene.coulomb@univ-tlse2.fr, abhishekkumar12@iisc.ac.in

RÉSUMÉ

Les réseaux sociaux sont un espace où les utilisateurs sont libres d'exprimer leurs opinions ce qui donne lieu à la diffusion de messages haineux ou insultants qui doivent être modérés. Nous proposons dans cet article une approche supervisée pour la détection automatique de message haineux dans une perspective multilingue. Nous nous intéressons en particulier à la haine exprimée à l'encontre de deux types de cibles (des immigrants et des femmes) dans des tweets en anglais, ainsi qu'aux messages sexistes dans des tweets en anglais et en français. Divers modèles d'apprentissage automatique ont été développés, allant de modèles à base de traits, à des approches neuronales. Nos expérimentations montrent des résultats encourageants pour les deux langues.

ABSTRACT

Social media networks have become a space where users are free to relate their opinions and sentiments which may lead to a large spreading of hatred or abusive messages which have to be moderated. This paper proposes a supervised approach to hate speech detection from a multilingual perspective. We focus in particular on hateful messages towards two different targets (immigrants and women) in English tweets, as well as sexist messages in both English and French. Several models have been developed ranging from feature-engineering approaches to neural ones. Our experiments show very encouraging results on both languages.

MOTS-CLÉS : Réseaux sociaux, Détection de message haineux, Sexism, apprentissage supervisée.

KEYWORDS: Social media, Hate speech detection, Sexism, supervised learning.

1 Motivation

Social media networks such as Facebook, Twitter, blogs and forums, have become a space where users are free to relate events, personal experiences, but also opinions and sentiments about products, events or other people. This may lead to a large spreading of hatred or abusive messages which have to be moderated. In particular, theses messages may express threats, harassment, intimidation or "disparage a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic" (Nockleby, 2000). Although some countries, such as the United States, where hate speech is protected under the First Amendment as freedom of expression (Massaro, 1990), many other countries, such as France, have laws prohibiting it, laws that extend to the internet and social media. For instance, since the French law of 27 January 2017 related

to equality and citizenship, penalties due to discrimination are doubled (sexism is now considered as an aggravating factor). Gender equality has also been declared "major national cause" for the five-year period mandate of French president Emmanuel Macron [1]. In this context, it is important to automatically detect hateful messages on social platforms and possibly to prevent the widespreading of gender/racial stereotypes, especially towards young people.

From a computational point of view, hate speech detection is casted as a binary classification task : given a message, classify it as conveying a hateful content or not. Most studies focus on offensive contents in general while others on specific type of hate (like racism or hate speech against the LGBT community) relying on feature-based engineering or neural approaches (see (Schmidt & Wiegand, 2017)) for a comprehensive survey). Data are mainly tweets written in English, although some recent studies attempt to detect hate speech in Spanish (Anzovino *et al.*, 2018; Basile *et al.*, 2019), German (Ross *et al.*, 2017), Italian (Corazza *et al.*, 2018), Slovene (Fišer *et al.*, 2017) and Dutch (Jha & Mamidi, 2017), the latter focusing on benevolent sexist messages containing expressions such as *for a girl* or *like a man*. Other studies propose to tackle hate speech from a multilingual perspective (e.g., English and Spanish at IberEval2018 (Anzovino *et al.*, 2018)), but do not consider any cross-language experiments, as participants' models are trained and tested on each language separately. Finally, concerning French, hate speech detection only focuses on racist (Valette, 2004) or abusive messages (Papegnies *et al.*, 2017).

In this paper, we focus on (1) automatic hate speech detection towards *two different targets* – immigrants and women – and (2) automatic sexism detection *from a multilingual perspective*, namely in English and French tweets. For English, the data consists of tweets annotated as conveying hate speech against both immigrants and women, as part of HateEval@SemEval2019 (Basile *et al.*, 2019) (henceforth HS). For French, the data consists also of tweets, but annotated only for sexism (henceforth SEXISM). The main contributions of this paper are the following :

1. A new French dataset annotated for sexism detection.
2. A multitarget hate speech detection system. We propose both features-based models (relying on both language-dependent and language independent features) and a neural model to measure to what extent hate speech detection is target-dependent. When using the same model, our results show that HS achieve better results than SEXISM.
3. A multilingual hate speech detection. We also experiment with multilingual embeddings by training on one language and testing on the other in order to measure how the proposed models are language dependent. Our results are encouraging and open the door to hate speech detection in languages that lack annotated data for hate speech.

The paper is organized as follows. Section 2 presents the current state of the art, Section 3 describes our data, Section 4 the models and the experiments we carried out on multitarget detection while Section 5 on the multilingual experiments. We conclude providing some perspectives for future work.

2 Related work

Both sexism and racism can be expressed at different linguistic granularity levels going from lexical to discursive (Cameron, 1992) : e.g, women are often designated through their relationship with men or motherhood or by physical characteristics. Sexism can also be hostile or benevolent where messages are subjectively positive expressed in the form of a compliment (Glick & Fiske, 1996). Basically, sexism may be expressed explicitly or implicitly (see the following tweets from our French data) using different pragmatic devices, including :

1. http://www.egalite-femmes-hommes.gouv.fr/marlene-schiappa-presente-ses-priorites-

— Negative opinion, abusive message : *Meuf tu connais rien au foot. Tais toi. Contente de fan girler sur les joueurs et de mouiller sur MBappé ....*
— Stereotype : *C'est bon t'es une femme forte, te manque que la cuisine pour atteindre la perfection*
— Humor, irony : *Le fait maison c'est toujours mieux. La preuve, on préfère toujours sa femme à sa prostituée. #humour.*
— Benevolent sexism : *Elle court vite pour une femme*.

Same devices can also be employed towards immigrants, like the following tweet taken from the English data that illustrates a stereotype : *Illegals are dumping their kids heres o they can get welfare, aid and U.S School Ripping off U.S Taxpayers #SendThemBack! Stop Alowing illegals to Abuse the Taxpayer #Immigration*.

Most of the classifiers employed in hate speech detection still rely on supervised learning, and when creating a new classifier, one may manually design and encode different types of features from the data instances which will then be directly fed to the classical algorithms (Naive Bayes, Logistic Regression, Random Forest, SVM) or use deep learning methods that will automatically learn abstract features from data instances. Within the Automatic Misogyny Identification shared task at IberEval 2018, the best results were obtained with Support Vector Machine models with different feature configurations. There are also a few notable neural networks techniques deployed in order to detect hate speech in tweets that outperform the existing models : in (Badjatiya *et al.*, 2017) the authors used three methods (Convolutional Neural Network (CNN), Long short-term memory and FastText) combined with either random or GloVe word embeddings. In (Zhang & Luo, 2018) the authors implemented two deep neural network models (CNN + Gated Recurrent Unit layer and CNN + modified CNN layers for feature extraction) in order to classify social media text as racist, sexist, or non-hateful.

For most of the harassment and hate speech classification tasks, the most used information is depicted by the surface-level features (e.g. Bag of Words), the majority of authors choosing to include n-grams in the feature sets due to their high prediction rate. Due to the noise present in the data (especially on social media), many authors choose to combine the n-grams with a large section of additional features : linguistic features that take into consideration the POS information, dependency relations (long-distance relationship in between words), or word embeddings, which have the advantage of having similar vector representations for different, but semantically similar words. Since the task of hate speech detection and sentiment analysis are closely related, several approaches incorporate the latter as a supplementary classification step, assuming that generally negative sentiment relates to a hateful message (Dinakar *et al.*, 2012; Sood *et al.*, 2012).

Hate speech detection is a particularly difficult task mostly because in different contexts, the meaning of a message might change as it can be highly dependent on knowledge about the world. Because of this, in (Dinakar *et al.*, 2012), the authors present an approach in which they use automatic reasoning over aspects of the world. As it might be difficult to obtain knowledge about the world, the information about an utterance (meta information) may be used in order to refine unsatisfactory classification. For example, in (Waseem & Hovy, 2016), by using the users gender information the results were significantly improved, as the authors found that it is more likely for men to post hate speech messages [2]. This idea was further developed in (Hasanuzzaman *et al.*, 2017) where the authors introduced demographic aware information (age, gender and location) in order to tackle racism and confirm an important increase in performance. Another important feature is based on the

2. In spite of these findings and due to the difficulty of accurately identifying the gender of the user, we do not find this method favorable from an ethical perspective as we can encourage a gender bias in the system.

assumption that a user known for posting hateful messages is more likely to do so again in the future, thus by using the number of profane words in the users previous messages the detection performance improves (Dadvar *et al.*, 2013).

As far as we know, no work have addressed neither sexism detection in French, nor multitarget hate speech detection.

3 Data

Our data come from two corpora. The first one, HS-IW, is an already existing corpus containing English tweets annotated for hate speech against immigrants and women, as part of the HatEval task at SemEval2019. The second corpus, SEXISM, is new and contains French tweets collected between October 2017 and May 2018 with specific keywords such as *#balancetonporc, #sexisme*, names of politician women and men, insults, etc. The tweets have been labelled as sexist or non sexist by 3 annotators (2 female and 1 male annotators [3]). 329 tweets have been labelled by all annotators and the inter-annotator agreement is 0.89 (Cohen's Kappa). For these tweets, the final labels have been assigned according to a majority vote. Table 1 shows the distribution of the tweets for both tasks (hate speech and sexism detection).

Task	#hate	#nonHate	Total
HS-IW (English)	5,512	7,559	13,071
SEXISM (French)	659	2,426	3,085

TABLE 1 – Tweet distribution in both French and English datasets

4 Multitarget hate speech detection

Automatically labelling tweets as hateful/not hateful or sexist/not sexist is a challenging task because the language of tweets is full of grammatically and/or syntactic errors, it lacks conversational context, might consist of only one or a few words and because they can be indirectly hateful (through the use of sarcasm or irony) it makes the task of text-based feature extraction difficult. For both corpora, several models have been built, all tested using 10-cross-validation to better compare our results in cross-lingual experiments. In the next sections, we detail our models and then give our results.

4.1 Models

To measure to what extent hate speech detection is target-dependent, we propose several models raging from standard bag of words (our baseline), features-based models to neural model. For all the models, due to the noise in the data, we performed standard text pre-processing : removing user mentions, URLs, RT, stop words, degraded stop words and the words containing less than 3 characters were filtered out. For HS-IW, all the remaining words were stemmed using the Snowball Stemmer [4], while for SEXISM, tweets have been lemmatized using the French MSTParser [5]. We also experimented without stems and lemmas, but the results were not conclusive.

Baseline. In all experiments, we used as our baseline unigrams, bigrams and trigrams Tf/IDF (we ignored the terms that appear in less than 4 tweets, as well as the terms that appear in more than 80% of the tweets).

Feature-based models. We relied on state of the art features that have shown to be useful in hate speech detection. Our features include the following :

3. They are master degree's students in Communication and Gender.
4. http://snowballstem.org
5. http://alpage.inria.fr/statgram/frdep/fr_stat_dep_mst.html

— *Surface features :* such as the tweet length in words, the presence or absence of punctuation marks (sequence of question/exclamation marks), the presence of URLs and @user mentions.
— *Sentiment features* : The idea is to test whether identifying user's opinion can better classify his attitude as hateful or non-hateful. We took into consideration several existing lexicons : AFINN (Nielsen, 2011), SentiWordNet (Esuli & Sebastiani, 2006), Liu and Hu opinion lexicon [6], HurtLex (a multilingual hate word lexicon divided in 17 categories) (Bassignana *et al.*, 2018) and a lexicon containing 1 818 profanity English words created by combining a manually built offensive words list, the noswearing dictionary [7] and an offensive word list [8]. In the final models we chose to include only HurtLex and the lexicon we built, as none of the other models outperformed our baseline model. For the French corpus, we chose to use HurtLex, as it already contains hate words translated into French.
— *Emojis features* : We relied on a manually built emojis lexicon that contains 1 644 emojis along with their polarity among positive, negative and neutral.

We experiment with several combinations of the features above, and we finally keep the most relevant ones by applying the Chi2 feature selection algorithm. The best performing features have been used to train four classifiers (C_1, C_2 for the task of hate speech detection and C_3, C_4 for the task of sexism detection). For each classifier, we tried several machine learning algorithms (Naive Bayes, Logistic Regression, Support Vector Machine, Decision Tree and Random Forest) in order to evaluate and select the best performing one. Hereby, the hate speech baseline is a Random Forest (the number of trees in the forest = 360 with a maximum depth of the tree = 600) and the sexism baseline is a Support Vector Machine (linear kernel, C = 0.1). For C_2, best results have been obtained when using Random Forest only for intermediate classification, whose output were then combined and passed onto a final Extreme Gradient Booster classifier. The four classifiers are as follows :

— C_1 : combines the length of the tweet with the number of words in the profanity lexicon with a baseline architecture as described above
— C_2 : on top of C_1 features we also used the number of positive and negative emojis and emoticons and we perform linear dimensionality reduction by means of truncated Singular Value Decomposition (latent semantic analysis on TF/IDF matrices).
— C_3 : combines the length of the tweet with the number of words in the HurtLex lexicon on top of a baseline architecture
— C_4 : the same features as C_3 but with a C_2 system architecture

Neural model. The last model used a Bidirectional LSTM with an attention mechanism that attends over all hidden states and generates attention coefficients [9]. The hidden states were then averaged using the attention coefficients in order to generate the final state which was then fed to a one-layer feed-forward network for obtaining the final label prediction. For the task of hate speech detection, we used pre-trained on tweets [10] Glove embeddings with an embedding dimension of 200 (Pennington *et al.*, 2014), while for the task of sexism detection we used pre-trained on Wikipedia and Common Crawl FastText French word vectors with an embedding dimension of 300 (Grave *et al.*, 2018)). We experimented with different hidden state vector sizes, dropout values and attention vector sizes. The results reported in this paper were obtained by using 300 hidden units, an 150 attention vector, a dropout of 50% and the Adam optimizer with a learning rate of 10^{-3}. For the BiLSTM we used a Relu activation function and we run all the experiments for maximum 100 epochs, with a patience of

6. https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon
7. https://www.noswearing.com/dictionary
8. http://www.cs.cmu.edu/~biglou/resources/bad-words.txt
9. We also experimented with other neural architectures, like CNN, but the results were lower.
10. We also experimented with pre-trained on Wikipedia word vectors, however the accuracy decreased by 3%

10 and batch size of 64 [11].

4.2 Results

Since the number of sexist instances in the French corpus is relatively small, the results presented in this paper were obtained by using 10-cross validation. Table 2 shows how the experiments were set up and presents the results in terms of accuracy (A), macro-averaged F-score (F), precision (P) and recall (R). The best results in terms of macro-averaged F-score (the evaluation metric used for ranking at SemEval) are presented in bold, while the columns left empty were intentionally left so, as we employed same system architectures with different features for the two tasks. Overall, our results show that when using the same model, the results achieved for the task of hate speech detection are better than the results for sexism detection.

	Hate speech detection				**Sexism detection**			
	A	F	P	R	A	F	P	R
Baseline	0.772	0.762	0.764	0.669	0.827	0.676	0.734	0.335
C_1	**0.788**	**0.780**	**0.785**	**0.684**	–	–	–	–
C_2	0.781	0.778	0.754	0.723	–	–	–	–
C_3	–	–	–	–	0.830	0.441	0.751	0.306
C_4	–	–	–	–	**0.822**	**0.688**	**0.665**	**0.386**
BiLSTM + attention	0.736	0.727	0.709	0.646	0.77	0.497	0.416	0.522

TABLE 2 – Hate speech detection and sexism detection results in both HS and SEXISM corpora

Among the systems, C_1 represents our best performing one for the task of hate speech detection, while C_4 performed best for the task of sexism detection.

Error analysis : A manual error analysis of the instances for which our best performing model and manual annotation differ shows that in the misclassification of hateful instances intervene several factors : the presence of off-topic tweets, the lack of context (as some words that trigger hate in certain contexts may have different connotations in others) and implicit hate speech that employs stereotypes or metaphors in order to convey hatred. We also identified tweets for which we question the original label when taking into account the class definition. Below, we have provided some examples.

Example 1 (HS-IW) : Although in the first tweet (annotated as not hateful) the user talks about Donald Trump, which doesn't fit in the targeted categories (immigrants or women), the annotation raises problems when trying to classify tweets such as the second one (annotated as hateful).

— I love my religious brothers and sisters, but @realDonaldTrump, FUCK YOU, YOU'RE NOT EVEN A REAL THEOCRAT YOU FAT USLESS BITCH.
— @menzemerized_ Worse i have proof. A picture i took of you and one you took of me on the same night. Useless ungreatful kunt !

Example 2 (SEXISM) : Both of the following tweets were misclassified due to the lack of context and knowledge about the world. In the first tweet, as we don't have enough information about the "liberté d'importuner" movement, we aren't able to properly classify the disagreement of the user with Catherine Deneuve's statements. The same problem arises in the second tweet, as the speech employs irony.

— Ce que je pense de la "liberté d'importuner". #Sexisme #CatherineDeneuve #Tribune C'est pas parce que vous aimez la soumission qu'on doit toutes apprécier. L'avis des vieilles bourgeoises qui ne prennent plus le métro sur les frotteurs, on s'en passe.

11. The hyperparameters were tuned on the validation set (20% of the training dataset), such that the best validation error was produced.

— Merkel en Allemagne. Thatcher et maintenant #TheresaMay au Royaume-Uni. En France une femme présidente ? Folie ! Décadence !

5 Multilingual hate speech detection

We also experimented with multilingual embeddings : Glove bilingual word embeddings [12] obtained as described in (Ferreira *et al.*, 2016) as well as French and English FastText word vectors mapped into the same embedding space following the alignment approach presented in (Smith *et al.*, 2017). For the experiments we used the same BiLSTM model described in Section 4.1, firstly by using the HS-IW English corpus for training and the SEXISM French corpus for testing, and secondly by using jointly the two corpora (HS-IW and 30% of the original SEXISM corpus) for training and testing on the remaining SEXISM corpus. Table 3 shows how the experiments were set up and presents the results in terms of accuracy (A) and macro-averaged F-score (F), the best result in terms of accuracy being presented in bold.

Corpus		**FastText**		**Glove**	
Train	Test	A	F	A	F
English	French	0.783	0.445	0.732	0.485
English + French	French	**0.790**	**0.461**	0.766	0.479

TABLE 3 – Multilingual hate speech detection results

The multilingual experiments results are somewhat comparable to the results obtained when training and testing on the French data (cf. Table 2). This is very encouraging as one can rely on external annotated data for sexism in other languages to learn a model on a different language. Of course, these results have to be confirmed as for the moment we do not have the actual distribution of the tweets in the SemEval corpus (the number of tweets that convey hate towards immigrants and the number of tweets that convey hate towards women).

Error analysis : The error analysis shows that in the absence of context and knowledge about the world (the #balancetonporc movement, as well as the persons to which the author of the tweet is referring to) and without employing irony detection systems, we misclassify (as non-sexist) tweets such as the following one :

— Donc on va avoir une conférence à Sciences Po avec Raphaël Enthoven, Aurore Bergé, Elisabeth Lévy et Pierre-Oliver Sur pour se demander comment #balancetonporc favorise la délation et la "mise au pilori" des accusés wow so much progressisme et ouverture.

6 Conclusion

This paper proposed several models that can be used in order to identify messages that convey hate and proved the portability of these systems for the task of detecting sexist messages in French. As far as we know, this is the first work on sexism detection in French on Twitter data, this study serving as a first step towards improving the task. In our future work we plan on studying ways to retrieve contextual information, and as the results seemed promising, we also plan on experimenting more in a multilingual embedding space.

Acknowledgments

This work has been funded by Maison des Sciences de l'Homme et de la Société de Toulouse under the project AMeSexTo.

12. http://www.cs.cmu.edu/~afm/projects/multilingual_embeddings.html

Références

ANZOVINO M., FERSINI E. & ROSSO P. (2018). Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, p. 57–64 : Springer.

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*, p. 759–760 : International World Wide Web Conferences Steering Committee.

BASILE V., BOSCO C., FERSINI E., NOZZA D., PATTI V., RANGEL F., ROSSO P. & SANGUINETTI M. (2019). Semeval-2019 task 5 : Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)* : Association for Computational Linguistics", location = "Minneapolis, Minnesota.

BASSIGNANA E., BASILE V. & PATTI V. (2018). Hurtlex : A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, p. 1–6 : CEUR-WS.

CAMERON D. (1992). *Feminism and Linguistic Theory.* Palgrave Macmillan.

CORAZZA M., MENINI S., ARSLAN P., SPRUGNOLI R., CABRIO E., TONELLI S. & VILLATA S. (2018). Comparing Different Supervised Approaches to Hate Speech Detection. In *EVALITA 2018*, Turin, Italy.

DADVAR M., TRIESCHNIGG D., ORDELMAN R. & DE JONG F. (2013). Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, p. 693–696 : Springer.

DINAKAR K., JONES B., HAVASI C., LIEBERMAN H. & PICARD R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, **2**(3), 18.

ESULI A. & SEBASTIANI F. (2006). Sentiwordnet : A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06*, p. 417–422.

FERREIRA D. C., MARTINS A. F. & ALMEIDA M. S. (2016). Jointly learning to embed and predict with multiple languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, volume 1, p. 2019–2028.

FIŠER D., ERJAVEC T. & LJUBEŠIĆ N. (2017). Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in slovene. In *Proceedings of the First Workshop on Abusive Language Online*, p. 46–51.

GLICK P. & FISKE S. T. (1996). The ambivalent sexism inventory : Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, **70**(3), 491–512.

GRAVE E., BOJANOWSKI P., GUPTA P., JOULIN A. & MIKOLOV T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

HASANUZZAMAN M., DIAS G. & WAY A. (2017). Demographic word embeddings for racism detection on twitter. In *IJCNLP*.

JHA A. & MAMIDI R. (2017). When does a compliment become sexist ? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, p. 7–16.

MASSARO T. M. (1990). Equality and freedom of expression : The hate speech dilemma. *Wm. & Mary L. Rev.*, **32**, 211.

NIELSEN F. Å. (2011). A new ANEW : evaluation of a word list for sentiment analysis in microblogs. In M. ROWE, M. STANKOVIC, A.-S. DADZIE & M. HARDEY, Eds., *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts' : Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, p. 93–98.

NOCKLEBY J. T. (2000). Hate speech. In *Encyclopedia of the American Constitution (2nd ed., edited by Leonard W. Levy, Kenneth L. Karst et al.*, p. 1277–1279.

PAPEGNIES E., LABATUT V., DUFOUR R. & LINARÈS G. (2017). Detection of abusive messages in an on-line community. In *14ème Conférence en Recherche d'Information et Applications (CORIA)*, p. 153–168.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543.

ROSS B., RIST M., CARBONELL G., CABRERA B., KUROWSKY N. & WOJATZKI M. (2017). Measuring the reliability of hate speech annotations : The case of the european refugee crisis. In *Proceedings of NLP4CMC III : 3rd Workshop on Natural Language Processing for Computer-Mediated Communication (Bochum)*, p. 6–9.

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*, p. 1–10.

SMITH S. L., TURBAN D. H. P., HAMBLIN S. & HAMMERLA N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR*, **abs/1702.03859**.

SOOD S. O., CHURCHILL E. F. & ANTIN J. (2012). Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, **63**(2), 270–285.

VALETTE M. (2004). Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur internet. In *Colloque International sur le Document Electronique*, p. 215–230 : Centre de recherche en Ingénierie Multilingue, INaLCO.

WASEEM Z. & HOVY D. (2016). Hateful symbols or hateful people ? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, p. 88–93.

ZHANG Z. & LUO L. (2018). Hate speech detection : A solved problem ? the challenging case of long tail on twitter. *arXiv preprint arXiv :1803.03662*.