Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pages 69–75
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

The Role of Computational Stylometry in Identifying (Misogynistic) Aggression in English Social Media Texts

Antonio Pascucci[1], **Raffaele Manna**[1], **Vincenzo Masucci**[2], **Johanna Monti**[1]
"L'Orientale" University of Naples - UNIOR NLP Research Group[1], Expert System Corp.[2]
Via Duomo 219 Naples (Italy)[1], Via Nuova Poggioreale 60 Naples (Italy)[2]
{apascucci,rmanna,jmonti}@unior.it, vmasucci@expertsystem.com

Abstract

In this paper, we describe UniOr_ExpSys team participation in TRAC-2 (Trolling, Aggression and Cyberbullying) shared task, a workshop organized as part of LREC 2020. TRAC-2 shared task is organized in two sub-tasks: *Aggression Identification* (a 3-way classification between "Overtly Aggressive", "Covertly Aggressive" and "Non-aggressive" text data) and *Misogynistic Aggression Identification* (a binary classifier for classifying the texts as "gendered" or "non-gendered"). Our approach is based on linguistic rules, stylistic features extraction through stylometric analysis and Sequential Minimal Optimization algorithm in building the two classifiers.

Keywords: Computational Stylometry, Aggression Identification, Social Media

1. Introduction

The spread of offensive and hate speech on social media is one of the issues that mostly concerns the scientific community. The number of hate and offensive posts and comments on social media is growing day by day and the measures adopted by social media managers are often not enough. Most of the time, haters'accounts are simply temporarily blocked, and no other effective measures to combat the phenomenon are taken. In this paper, we describe our participation in TRAC-2 (Ritesh Kumar and Zampieri, 2020) workshop shared task and the results we achieved. TRAC-2 workshop shared task (now in its second edition), focuses on trolling, aggression and cyberbullying detection in a given corpus built ad hoc by the task organizers and is organized in two sub-tasks: *Aggression Identification* task and *Misogynistic Aggression Identification* task. TRAC-2 workshop shared task includes texts in three different languages: Bangla, Hindi and English for both sub-tasks. The participants are allowed to compete for the tasks and the languages they prefer. Considering the importance of linguistic knowledge in our approach, we decided to participate only in the two English sub-tasks (since we don't have linguistic knowledge in Bangla and Hindi). The method we use for text data classification, indeed, is based on a hybrid approach of Computational Stylometry, Machine Learning and Linguistic Rules. This research has been carried out in the context of two innovative industrial PhD projects in cooperation between the "L'Orientale" University of Naples and Expert System Corp. (a semantic intelligence company that creates artificial intelligence, cognitive computing and semantic technology software). That's the reason why we chose the name "UniOr_ExpSys" for our team. The paper is organized as follows: in Section 2 we show Related work in Hate and Offensive speech detection. Section 3 focuses on methodology and data. Results are in Section 4 and Conclusions are in Section 5.

2. Related work

Over the last few years, hate speech (HS) and offensive speech (OS) detection, has generated interest in scholars (for a survey, see (Schmidt and Wiegand, 2017) and (Fortuna and Nunes, 2018)). The advent of social media represents the main cause of the HS and OS spread. Social networks are an extremely efficient means of communication, but, unfortunately, not everyone makes proper use of them. Increasing vulgarity in online conversations has emerged as a relevant issue in society as well as in science (Ramakrishnan et al., 2019). The difference between HS and OS is subtle but significant and can be summarized as: HS is deemed to be harmful on the basis of defined *protected attributes* such as race, disability, sexuality and so on. In other words, HS is the intention to denigrate "a person or persons on the basis of (alleged) membership in a social group identified by attributes such as race, ethnicity, gender, sexual orientation, religion, age, physical or mental disability, and others" (Britannica, 2015); instead, OS can be described as a speech that "Causes someone to feel hurt, angry, or upset : rude or insulting"[1].

Research on detecting HS presence in social media has been carried out by (Malmasi and Zampieri, 2017). The scholars investigated the dataset built by (Davidson et al., 2017), composed of 14,509 English tweets annotated by three annotators into one of the following three classes: HATE (tweets containing HS), OFFENSIVE (tweets containing OS) and OK (non-offensive tweets). (Malmasi and Zampieri, 2017) used a linear Support Vector Machine to perform multi-class classification and achieved the best performance of 0.78 of text correctly classified with character 4-grams feature. A very ambitious project is *Contro l'odio* (literally *Against hate*), a web platform for monitoring and contrasting discrimination and HS against immigrants in Italy (Capozzi et al., 2019). The classifier they built is trained with the *Italian Hate Speech Corpus* (IHSC) (Sanguinetti et al., 2018), a collection of about 6,000 HS tweets. *Contro l'odio* project extends the research outcomes that emerged from the *Italian Hate Map* project (Musto et al., 2016), combining computational linguistics methods that

[1] `https://www.merriam-webster.com/dictionary/offend`

allow users to access a huge amount of information through interactive maps. (De Smedt et al., 2018) proposed a report on multilingual cross-domain (Extremism, Jidahism, Sexism and Racism) perspectives on online HS detection to identify common features of HS across domains. The scholars exploited different techniques (sentiment analysis, text classification, keyword extraction, and collocation extraction) and argued that it is hard to come up with a linguistic definition of HS, because there is no standardized "list of bad words", and if there is, then perpetrators are very creative in coining new offensive terminology.

Cyberbullying is also part of HS and OS, especially if we consider that social media represent real breeding grounds in which new and increasingly sophisticated forms of cyberbullying are being developed. The detection and classification of textual cyberbullying on social media has been well investigated in (Dinakar et al., 2011), (Xu et al., 2012), (Dadvar et al., 2013), and (Burnap and Williams, 2015). With the aim of monitoring the presence of cyberbullying in online texts, CREEP's project (Menini et al., 2019) main goal is to support supervising persons (e.g., educators) at identifying potential cases of cyberbullying. Stylistic features extraction in cyberbullying texts has been also investigated in (Pascucci et al., 2019) with a focus on features that belong to ten different cyberbullying categories characterized by text. Interesting research has been carried out by (Sprugnoli et al., 2018), who built a corpus of WhatsApp chats through a role-play by three classes of students aged 12 and 13 made of 14,600 tokens. In their corpus, the scholars distinguish four cyberbullying roles (Harasser, Victim, Bystander-defender, Bystander-assistant) and different classes of insults or discrimination, such as Body Shame, Sexism, Racism and Sexual Harassment. Their data have been annotated by two annotators and 1,203 cyberbullying expressions have been identified, corresponding to almost 6,000 tokens (41.1% of the whole corpus).

Italian scientific community pays a great deal of attention to HS and OS detection shared task, and a few linguistic resources (Sanguinetti et al., 2018), (Poletto et al., 2017), and (Del Vigna et al., 2017) have been developed regarding HS Facebook and Twitter comments in Italian.

The following is a short and certainly not exhaustive list that includes HS and OS shared tasks organized in the last few years:

- *HaSpeeDe* (Bosco et al., 2018), a shared task on HS detection, based on two datasets from two different online social platforms differently featured from the linguistic and communicative point of view. The shared task has been organized in the context of EVALITA 2018 (a periodic evaluation campaign of natural language processing and speech tools for the Italian language);

- *Germeval* (Wiegand et al., 2018), classification of German tweets from Twitter. It included a coarse-grained binary classification task and a fine-grained multiclass classification task;

- *AMI* (Fersini et al., 2018), a shared task on automatic misogyny identification divided in two subtasks: Subtask A on misogyny identification and Subtask B about misogynistic behaviour categorization and target classification. *AMI* shared task has been organized in the context of EVALITA 2018;

- *Hateval* (Basile et al., 2019), a shared task on multilingual detection of HS against immigrants and women in twitter organized as part of SemEval 2019. The shared task involved a total of 74 participants to detect HS in the dataset and to distinguish if the incitement was against an individual rather than a group;

- *Offenseval* (Zampieri et al., 2019b), also organized in the context of SemEval 2019, focuses on identifying and categorizing OS in social media. The task was based on a dataset (OLID - Offensive Language Identification Dataset) (Zampieri et al., 2019a) built ad hoc for this occasion. *Offenseval* was organized in three sub-tasks: in sub-task A, the goal was to discriminate between offensive and non-offensive posts. In subtask B, the focus was on the type of offensive content in the post, and in sub-task C, systems had to detect the target of the offensive posts. The 2020 *Offenseval* edition will be held as part of COLING 2020.

- *TRAC-1* (Kumar et al., 2018a), the first workshop on trolling, aggression and cyberbullying. TRAC-1 shared task (Kumar et al., 2018b) has been organized as part of COLING 2018 conference. TRAC-1 included a shared task on Aggression Identification (Kumar et al., 2018a). The task was to develop a classifier that could make a 3-way classification between Overtly Aggressive (OAG), Covertly Aggressive (CAG), or Non-Aggressive (NAG) text data in Hindi and English. It involved 130 teams, but only 30 of these submitted their systems. Besides, only 20 teams decided to submit their system description paper. TRAC-1 shared task organizers provided two test sets for Hindi and English: the first one was composed of 916 English Facebook comments and 970 Hindi Facebook comments. Additionally, 1,257 English tweets and 1,194 Hindi tweets have been provided as the surprise test set. The three best performing teams in English language in TRAC-1 shared task are: *vista.ue* (Raiyani et al., 2018), *Julian* (Risch and Krestel, 2018), and *saroyehun* (Aroyehun and Gelbukh, 2018). In Table 1 the three systems performances are reported in terms of F1-weighted.

	saroyehun	**Julian**	**vista.ue**
Facebook Test set	0.642	0.601	0.581
Surprise Test set	0.592	0.599	0.600

Table 1: Performances achieved by the three TRAC-1 best teams on the TRAC-1 Facebook test set and the Surprise test set for English language

TRAC-2 takes its cue from TRAC-1 workshop.

3. Methodology and Data

In this section, we describe our approach to text classification and TRAC-2 shared task data.

3.1. Methodology

Our approach to text analysis and features extraction is a hybrid approach of Computational Stylometry (CS), Machine Learning (ML) and Linguistic Rules (LR).

CS can be described as a set of techniques that allow scholars to find out information about the authors of texts through an automatic linguistic analysis of texts. One of the main assumptions in CS is that each author operates choices which are influenced by sociological (age, gender and education level) and psychological (personality, mental health and being a native speaker or not) factors (Daelemans, 2013) which determine a unique writing style. With this in mind, it is natural that stylistic features play a fundamental role in detecting author's traits. Considering that stylistic features detected over the years by the scholars are at least one hundred, we summarize in a short list some main stylistic features studied in literature: sentence length (Argamon et al., 2003), vocabulary richness (De Vel et al., 2001), word length distributions (Zheng et al., 2006), punctuation (Baayen et al., 1996), use of a specific class of verbs or adjectives, use of first/third person, n-grams, readability index (Lucisano and Piemontese, 1988), use of metaphors.

Concerning ML, it is known that there are so many definitions, but the most exhaustive and concise is: ML is the computer ability to learn from data and consists in making predictions on unknown data on the basis of parameters identified during the training process.

Lastly, the LR writing process is carried out thanks to COGITO©, Expert System's semantic intelligence software, by which it is possible to write rules to process the texts and extract all the characteristics. An important aspect of the software is that it allows to perform word-sense disambiguation, that is crucial in text analysis, exploiting the power of its semantic network. Our standard approach to text analysis consists of the following steps:

- *Linguistic Definition of Stylometric Features*: since each author operates grammatical choices when writing a text, we organize all the grammatical characteristics of the texts under study in a taxonomy to detect the authorial fingerprint based on the grammatical choices done. This first step is carried out thanks to COGITO©, that allows us to write LR;
- *Semantic Engine Development*: we train the semantic engine to extract the features from the analyzed texts. The semantic engine is implemented thanks to COGITO©'s semantic network (*Sensigrafo*) - that can operate word-sense disambiguation - with the addition of the rules we built;
- *Training Set Analysis*: the training set is analysed and all features (based on the grammatical choices done by the writer) are extracted;
- *ML*: In the last step, we exploit the features extracted to train the model to detect these features in the dataset. ML process is carried out exploiting WEKA platform (Hall et al., 2009) (a software with machine learning tools and algorithms for data analysis) thanks to which it is possible to build a classifier with the support of one of the algorithms available.

3.2. Task description and Data

TRAC-2 workshop shared task (now in its second edition), focuses on trolling, aggression and cyberbullying detection in a given corpus build ad hoc by the task organizers and is organized in two sub-tasks:

- Sub-task-A: *Aggression Identification* task, for which participant have to build a 3-way classifier to detect if the texts are (OAG), (CAG), or (NAG);
- Sub-task-B: *Misogynistic Aggression Identification* task, for which participants have to build a binary classifier for classifying texts as Gendered (GEN) or Non-Gendered (NGEN).

As we reported, TRAC-2 shared task included also a second SubTask (*Misogynistic Aggression Identification*), as opposed to TRAC-1, which included only the *Aggression Identification* SubTask. TRAC-2 shared task includes texts in three different languages: Bangla, Hindi and English (as opposed to TRAC-1, which didn't include Bangla) for both sub-tasks (Bhattacharya et al., 2020). The participants are allowed to compete for the tasks and the languages they prefer. As we mentioned in Section 3.1, building ad hoc LR and exploiting our semantic network plays a crucial role in our approach, so considering that we have no linguistic knowledge in Bangla and Hindi, we decided to take part only in the two English sub-tasks.

3.2.1. Evaluation Metric

The systems submitted to TRAC-2 shared task have been evaluated on the basis of weighted macro-averaged F-scores. It means that the individual F-score of each class has been weighted by the proportion of the concerned class in the test set. The final F-score represents the average of these individual F-scores of each class.

3.2.2. Preprocessing

As usual in social media text data analysing, we cleaned the texts before analysying them. We removed @ symbol (it means that we also removed all mentions), we also removed hashtags (#), URLs, and emojis.

3.2.3. Training set and Dev set analysis

TRAC-2 English shared task training set is composed of 4,217 text data labelled both for SubTask A and for SubTask B. Besides this, a Dev set composed of 1,064 text data even those labelled for both SubTasks was also delivered. In order to detect the best performing algorithm between Random Forest (RF) (Liaw et al., 2002), Simple Logistic (SL) (Peng et al., 2002), and Sequential Minimal Optimization (SMO) (Platt, 1998), we built three different classifiers. Firstly, we train the three different model with the Training set for both SubTasks and we tested it on the Dev set. The results are shown in Table 2 (SubTask A) and Table 3 (SubTask B).

3.2.4. Cross-validation

Cross-validation is a method used to test the performance of a model. The 10-folds cross-validation phase also confirmed that SMO classifier performances were better than those of the classifiers trained with the other two algorithms

Classifier	Precision	Recall	F-Measure
RF	0.537	0.495	0.498
SL	0.472	0.449	0.454
SMO	0.546	0.528	**0.530**

Table 2: Evaluation on SubTask A Dev set using SubTask A Training set as training, where all performances reported should be read as weighted

Classifier	Precision	Recall	F-Measure
RF	0.659	0.618	0.616
SL	0.630	0.595	0.594
SMO	0.663	0.630	**0.630**

Table 3: Evaluation on SubTask B Dev set using SubTask A Training set as training, where all performances reported should be read as weighted

(RF and SL). The results of the 10-folds cross-validation test on both SubTasks Training sets are shown in Table 4 (SubTask A) and Table 5 (SubTask B).

Classifier	Precision	Recall	F-Measure
RF	0.510	0.508	0.501
SL	0.503	0.505	0.496
SMO	0.569	0.523	**0.527**

Table 4: 10-folds Cross-validation on SubTask A Training set, where all performances reported should be read as weighted

Classifier	Precision	Recall	F-Measure
RF	0.595	0.592	0.589
SL	0.645	0.644	**0.642**
SMO	0.642	0.642	**0.642**

Table 5: 10-folds Cross-validation on SubTask B Training set, where all performances reported should be read as weighted

Considering the performances achieved in both Dev set evaluation tests and in the two 10-folds cross-validation tests, we decided to analyze the Test set with the classifier we built with the support of the SMO algorithm.

3.2.5. TRAC-2 Test set

The Test set developed by Trac-2 shared task organizers is composed of 1,200 text data to be labelled in both SubTasks. As we mentioned above, in SubTask A it is possible to label text data as: OAG, CAG, or NAG. In SubTask B texts can be labelled as GEN or NGEN. Despite each team was allowed to submit up to three systems for evaluation, we decided to submit just one for both SubTasks. The decision originated from the fact that the SMO algorithm was the best performing algorithm since the analysis TRAC-2 training and dev set. As shown above, other classifiers trained with other algorithms achieved worse performances.

4. Results

In this section, we show the results achieved by UniOr_ExpSys in both SubTasks. In the following few lines, we describe our hybrid approach of CS, ML and LR. Thanks to COGITO©we are able to build ad hoc linguistic rules to recognize stylistic features in texts. After this process, we train a semantic engine to extract the aforementioned features. The semantic engine is implemented thanks to the semantic network with the addition of the rules we built. Then, the training set is analysed and all features are extracted. In the last step, we exploit the features extracted to train the model to detect these features in the dataset. For the ML process, we exploit the WEKA platform and we built a classifier with the support of the SMO Algorithm. Please note that our system is trained with TRAC- 2 training set and TRAC - 1 dataset with regard to SubTask A and only with TRAC-2 training set with regard to SubTask B. The results achieved in TRAC-2 SubTask A (*Aggression Identification task*) and TRAC-2 SubTask B (*Misogynistic Aggression Identification task*) are shown in Table 6 and Table 7 respectively.

System	F1 (weighted)	Accuracy
CS-LR-SMO	**0.6291**	0.62

Table 6: Results for Sub-task EN-A.

System	F1 (weighted)	Accuracy
CS-LR-SMO	**0.6733**	0.6183

Table 7: Results for Sub-task EN-B.

4.1. Error analysis

It is important to highlight that our approach pays close attention to linguistic and stylistic aspects. Each feature is extracted thanks to the linguistic analysis of texts. In several instances, it has not been possible to extract stylistic features characterizing that specific category of texts (especially because texts were too short). Another fundamental aspect required by our approach is represented by balanced data, both in the training set and in the test set. Balanced data would have allowed a better training phase, with positive effects also on the classifier performances. Nevertheless, we are happy about the results we achieved in TRAC-2 participation and we thank the task organizers for the exciting competition in which we participated. In the future, exploring deep learning techniques for classifying these kinds of text data is certainly necessary.

Figure 1 and Figure 2 show the confusion matrices of both SubTasks classifiers.

As we can see in the SubTask A confusion matrix (Figure 1), CAG class text data are well classified, with the only exception of 15 instances incorrectly classified. The class that achieved the worst performance is NAG, which includes Non-Aggressive texts, but 156 have been classified as CAG and even 74 as OAG. With regard to SubTask B confusion matrix (Figure 2), GEN text data are quite well classified, while there is a big issue with NGEN: slightly more than

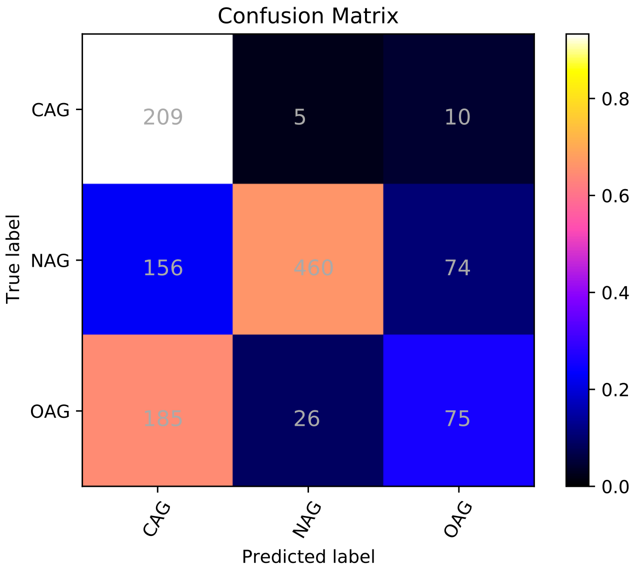

Figure 1: Sub-task EN-A, confusion matrix of the CS-LR-SMO model

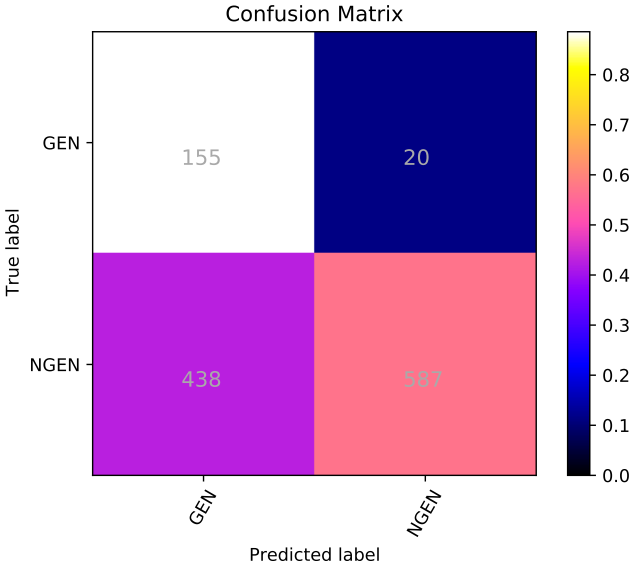

Figure 2: Sub-task EN-B, confusion matrix of the CS-LR-SMO model

half text data have been correctly classified, and this has undermined the performance of our binary classifier.

5. Conclusions

In this paper, we have shown the results achieved during the participation at TRAC-2 shared task workshop, organized as part of LREC 2020. The shared task is organized in two SubTasks: *Aggression Identification task*, for which participant have to build a 3-way classifier to detect if the texts are i) Overtly Aggressive (OAG), ii) Covertly Aggressive (CAG), or iii) Non-Aggressive (NAG) and Sub-task-B: *Misogynistic Aggression Identification task*, for which participants have to build a binary classifier for classifying texts as i) Gendered (GEN) or ii) Non-Gendered (NGEN). We use a hybrid approach based on CS, ML and LR, which focuses on stylistic features extraction to identify the features that characterize texts belonging to the different categories. With regard to *Aggression Identification task* we achieved 0.629072 of F1-weighted, and with regard to *Misogynistic Aggression Identification task* we achieved 0.673321.

6. Acknowledgements

This research has been partly supported by the PON Ricerca e Innovazione 2014-20 and the POR Campania FSE 2014-2020 funds. Authorship contribution is as follows: Antonio Pascucci is author of Sections 1, 2, and 3. Sections 4 and 5 are in common between Antonio Pascucci and Raffaele Manna. This research has been developed in the framework of two Innovative Industrial PhD projects in Computational Stylometry (CS) by "L'Orientale" University of Naples in cooperation with Expert System Corp. We are grateful to Vincenzo Masucci and Expert System Corp. for providing COGITO©for research and to Prof. Johanna Monti for supervising the research.

7. Bibliographical References

Argamon, S., Šarić, M., and Stein, S. S. (2003). Style mining of electronic messages for multiple authorship discrimination: first results. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 475–480. ACM.

Aroyehun, S. T. and Gelbukh, A. (2018). Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 90–97.

Baayen, H., Van Halteren, H., and Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–132.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., Rosso, P., and 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*, pages 54–63.

Bhattacharya, S., Singh, S., Kumar, R., Bansal, A., Bhagat, A., Dawer, Y., Lahiri, B., and Ojha, A. K. (2020). Developing a multilingual annotated corpus of misogyny and aggression.

Bosco, C., Dell'Orletta, F., Poletto, F., Sanguinetti, M., and Tesconi, M. (2018). Overview of the evalita 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9. CEUR.

Britannica, E. (2015). Britannica academic. *Encyclopædia Britannica Inc.*

Burnap, P. and Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

Capozzi, A. T., Lai, M., Basile, V., Poletto, F., Sanguinetti, M., Bosco, C., Patti, V., Ruffo, G., Musto, C., Polignano, M., et al. (2019). Computational linguistics against hate: Hate speech detection and visualization on social media in the" contro l'odio" project. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR-WS.

Dadvar, M., Trieschnigg, D., Ordelman, R., and de Jong, F. (2013). Improving cyberbullying detection with user context. In *Advances in Information Retrieval*, pages 693–696. Springer.

Daelemans, W. (2013). Explanation in computational stylometry. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 451–462. Springer.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

De Smedt, T., Jaki, S., Kotzé, E., Saoud, L., Gwóźdź, M., De Pauw, G., and Daelemans, W. (2018). Multilingual cross-domain perspectives on online hate speech. *arXiv preprint arXiv:1809.03944*.

De Vel, O., Anderson, A., Corney, M., and Mohay, G. (2001). Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4):55–64.

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., and Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

Dinakar, K., Reichart, R., and Lieberman, H. (2011). Modeling the detection of textual cyberbullying. In *The Social Mobile Web*, pages 11–17.

Fersini, E., Nozza, D., and Rosso, P. (2018). Overview of the evalita 2018 task on automatic misogyny identification (ami). *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:59.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018a). Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018b). Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling (TRAC)*, Santa Fe, USA.

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

Lucisano, P. and Piemontese, M. E. (1988). Gulpease: una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e città*, 3(31):110–124.

Malmasi, S. and Zampieri, M. (2017). Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427*.

Menini, S., Moretti, G., Corazza, M., Cabrio, E., Tonelli, S., and Villata, S. (2019). A system to monitor cyberbullying based on message classification and social network analysis. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 105–110.

Musto, C., Semeraro, G., de Gemmis, M., and Lops, P. (2016). Modeling community behavior through semantic analysis of social data: The italian hate map experience. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 307–308.

Pascucci, A., Masucci, V., and Monti, J. (2019). Computational stylometry and machine learning for gender and age detection in cyberbullying texts. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–6. IEEE.

Peng, C.-Y. J., Lee, K. L., and Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.

Poletto, F., Stranisci, M., Sanguinetti, M., Patti, V., and Bosco, C. (2017). Hate speech annotation: Analysis of an italian twitter corpus. In *4th Italian Conference on Computational Linguistics, CLiC-it 2017*, volume 2006, pages 1–6. CEUR-WS.

Raiyani, K., Gonçalves, T., Quaresma, P., and Nogueira, V. B. (2018). Fully connected neural network with advance preprocessor to identify aggression over facebook and twitter. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 28–41.

Ramakrishnan, M., Zadrozny, W., and Tabari, N. (2019). Uva wahoos at semeval-2019 task 6: Hate speech identification using ensemble machine learning. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 806–811.

Risch, J. and Krestel, R. (2018). Aggression identification using deep learning and data augmentation. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 150–158.

Ritesh Kumar, Atul Kr. Ojha, S. M. and Zampieri, M. (2020). Evaluating aggression identification in social media. In Ritesh Kumar, et al., editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*, Paris, France, may. European Language Resources Association (ELRA).

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., and 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)*.

Schmidt, A. and 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. Association for Computational Linguistics*, pages 1–10, Valencia, Spain.

Sprugnoli, R., Menini, S., Tonelli, S., Oncini, F., and Piras, E. (2018). Creating a whatsapp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59.

Wiegand, M., Siegel, M., and Ruppenhofer, J. (2018). Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval*.

Xu, J.-M., Jun, K.-S., Zhu, X., and Bellmore, A. (2012). Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666. Association for Computational Linguistics.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019a). Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019b). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

Zheng, R., Li, J., Chen, H., and Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American society for information science and technology*, 57(3):378–393.