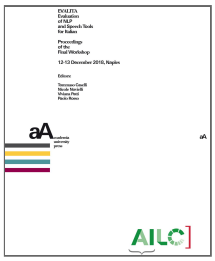

Tommaso Caselli, Nicole Novielli, Viviana Patti and Paolo Rosso (dir.)

EVALITA Evaluation of NLP and Speech Tools for Italian
Proceedings of the Final Workshop 12-13 December 2018, Naples

Accademia University Press

Automatic Identification of Misogyny in English and Italian Tweets at EVALITA 2018 with a Multilingual Hate Lexicon

Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile and Viviana Patti

DOI: 10.4000/books.aaccademia.4724
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2018
Published on OpenEdition Books: June 5, 2019
Series: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic EAN: 9788831978699

http://books.openedition.org

Printed version
Date of publication: December 1, 2018

Electronic reference
PAMUNGKAS, Endang Wahyu ; et al. *Automatic Identification of Misogyny in English and Italian Tweets at EVALITA 2018 with a Multilingual Hate Lexicon* In: *EVALITA Evaluation of NLP and Speech Tools for Italian: Proceedings of the Final Workshop 12-13 December 2018, Naples* [online]. Torino: Accademia University Press, 2018 (generated 05 octobre 2023). Available on the Internet: <http://books.openedition.org/aaccademia/4724>. ISBN: 9788831978699. DOI: https://doi.org/10.4000/books.aaccademia.4724.

Automatic Identification of Misogyny in English and Italian Tweets at EVALITA 2018 with a Multilingual Hate Lexicon

Endang Wahyu Pamungkas[1], **Alessandra Teresa Cignarella**[1,2], **Valerio Basile**[1] **and Viviana Patti**[1]
[1]Dipartimento di Informatica, Università degli Studi di Torino
[2]PRHLT Research Center, Universitat Politècnica de València
{pamungka | cigna | basile | patti}@di.unito.it

Abstract

English. In this paper we describe our submission to the shared task of Automatic Misogyny Identification in English and Italian Tweets (AMI) organized at EVALITA 2018. Our approach is based on SVM classifiers and enhanced by stylistic and lexical features. Additionally, we analyze the use of the novel *HurtLex* multilingual linguistic resource, developed by enriching in a computational and multilingual perspective of the hate words Italian lexicon by the linguist Tullio De Mauro, in order to investigate its impact in this task.

Italiano. *Nel presente lavoro descriviamo il sistema inviato allo shared task di Automatic Misogyny Identification (AMI) ad EVALITA 2018. Il nostro approccio si basa su classificatori SVM, ottimizzati da feature stilistiche e lessicali. Inoltre, analizziamo il ruolo della nuova risorsa linguistica* HurtLex*, un'estensione in prospettiva computazionale e multilingue del lessico di* parole per ferire *in italiano proposto dal linguista Tullio De Mauro, per meglio comprendere il suo impatto in questo tipo di task.*

1 Introduction

Hate Speech (HS) can be based on race, skin color, ethnicity, gender, sexual orientation, nationality, or religion, it incites to violence and discrimination, abusive, insulting, intimidating, and harassing. Hateful language is becoming a huge problem in social media platforms such as Twitter and Facebook (Poland, 2016). In particular, a type of cyberhate that is increasingly worrying nowadays is the use of hateful language that specifically targets women, which is normally referred to as: MISOGYNY (Bartlett et al., 2014).

Misogyny can be linguistically manifested in numerous ways, including social exclusion, discrimination, hostility, threats of violence and sexual objectification (Anzovino et al., 2018). Many Internet companies and micro-blogs already tried to tackle the problem of blocking this kind of online contents, but, unfortunately, the issue is far from being solved because of the complexity of the natural language[1] (Schmidt and Wiegand, 2017). For the above-mentioned reasons, it has become necessary to implement targeted NLP techniques that can be automated to treat hate speech online and misogyny.

The first shared task specifically aimed at Automatic Misogyny Identification (AMI) took place at IberEval 2018[2] within SEPLN 2018 considering English and Spanish tweets (Fersini et al., 2018a). Hence, the aim of the proposed shared task is to encourage participating teams in proposing the best automatic system firstly to distinguish misogynous and non-misogynous tweets, and secondly to classify the type of misogynistic behaviour and judge whether the target of the misogynistic behaviour is a specific woman or a group of women. In this paper, we describe our submission to the 2nd shared task of Automatic Misogyny Identification (AMI)[3] organized at EVALITA 2018, organized in the same manner but focusing on Italian tweets, rather than Spanish and English as in the IberEval task.

2 Task Description

The aim of the AMI task is to detect misogynous tweets written in English and Italian (Task A) (Fersini et al., 2018b). Furthermore, in Task

[1]https://www.nytimes.com/2013/05/29/business/media/facebook-says-it-failed-to-stop-misogynous-pages.html
[2]https://sites.google.com/view/ibereval-2018
[3]https://amievalita2018.wordpress.com/

B, each system should also classify each misogynous tweet into one of five different misogyny behaviors (STEREOTYPE, DOMINANCE, DERAILING, SEXUAL HARASSMENT, AND DISCREDIT) and two targets of misogyny classes (active and passive). Participants are allowed to submit up to three runs for each language. Table 1 shows the dataset label distribution for each class. Accuracy will be used as an evaluation metric for Task A, while macro F-score is used for Task B.

The organizers provided the same amount of data for both languages: 4,000 tweets in the training set and 1,000 in the test set. The label distribution for Task A is balanced, while in Task B the distribution is highly unbalanced for both misogyny behaviors and targets.

3 Description of the System

We used two Support Vector Machine (SVM) classifiers which exploit different kernels: linear and radial basis function (RBF) kernels.

SVM with Linear Kernel. Linear kernel was used to find the optimal hyperplane when SVM was firstly introduced in 1963 by Vapnik et al., long before Cortes and Vapnik (1995) proposed to use the kernel trick. Joachims (1998) recommends to use linear kernel for text classification, based on the observation that text representation features are frequently linearly separable.

SVM with RBF Kernel. Choosing the kernel is usually a challenging task, because its performance will be dataset dependent. Therefore, we also experimenteed with a Radial Basis Function (RBF) kernel, which has been already proven as an effective classifier in text classification problems. The drawback of RBF kernels is that they are computationally expensive and obtain a worse performance in big and sparse feature matrices.

3.1 Features

We employed several lexical features, performing a simple preprocessing step including tokenization and stemming, using the NLTK (Natural Language Toolkit) library[4]. A detailed description of the features employed by our model follows.

Bag of Words (BoW). We used bags of words in order to build the tweets representation. Before producing the word vector, we changed all the characters from upper to lower case. Our vector space consists of the count of unigrams and bigrams as a representation of the tweet. In addition, we also employed **Bag of Hashtags (BoH)** and **Bag of Emojis (BoE)** features, which are built by using the same technique as BoW, focusing on the presence of hashtags and emojis.

Swear Words. This feature takes into account the presence of a swear word and the number of its occurrences in the tweet. For English, we took a list of swear words from `www.noswearing.com`, while for Italian we gathered the swear word list from several sources[5] including a translated version of `www.noswearing.com`'s list and a list of swear words from Capuano (2007).

Sexist Slurs. Beside swear words, we also considered sexist words, that are specifically targeting women. We used a small set of sexist slurs from previous work by Fasoli et al. (2015). We translated and expanded that list manually for our Italian systems. This feature has a binary value, 1 when at least one sexist slur presence on tweet and 0 when there is no sexist slur on tweet.

Women Words. We manually built a small set of words containing synonyms and several words related to word “woman" in English and “donna" in Italian. Based on our previous work (Pamungkas et al., 2018), these words were effective to detect the target of misogyny on the tweet. Similar to sexist slur feature, this feature also has binary value show the presence of women words on tweet.

Surface Features. We also considered several surface level features including: **upper case** character count, number of **hashtags**, number of **URLs**, and the **length** of the tweet counting the characters.

Hate Words Lexicon. HurtLex (Bassignana et al., 2018) is a multilingual lexicon of hate words, built starting from a list of words compiled manually (De Mauro, 2016). The lexicon is semi-automatically translated into 53 languages, and the lexical items are divided into 17 categories (see Table 2). For our system configuration, we exploited the presence of the words in each category as a single feature, thus obtaining 17 single features, one for each HurtLex category.

[4] `https://www.nltk.org/`

[5] `https://www.parolacce.org/2016/12/20/dati-frequenza-turpiloquio/` and `https://it.wikipedia.org/wiki/Turpiloquio_nella_lingua_italiana`

Task A			Task B		
	English	**Italian**		**English**	**Italian**
Misogynistic	1,785/460	1,828/512	Stereotype	179/140	668/175
			Dominance	148/124	71/61
			Derailing	92/11	24/2
			Sexual Harassment	352/44	431/170
			Discredit	1,014/141	634/104
			Active	1,058/401	1,721/446
			Passive	727/59	96/66
Not misogynistic	2,215/540	2,172/488	No class	2,215/540	2,172/488
Total				4,000/1,000	4,000/1,000

Table 1: Dataset label distribution (training/test).

Category	Description
PS	Ethnic Slurs
RCI	Location and Demonyms
PA	Profession and Occupation
DDP	Physical Disabilities and Diversity
DDF	Cognitive Disabilities and Diversity
DMC	Moral Behavior and Defect
IS	Words Related to Social and Economic antage
OR	Words Related to Plants
AN	Words Related to Animals
ASM	Words Related to Male Genitalia
ASF	Words Related to Female Genitalia
PR	Words Related Prostitution
OM	Words Related Homosexuality
QAS	Descriptive Words with Potential Negative Connotations
CDS	Derogatory Words
RE	Felonies and Words Related to Crime and Immoral Behavior
SVP	Words Related to the Seven Deadly Sins of the Christian Tradition

Table 2: HurtLex Categories.

4 Experimental Setup

We experimented with different sets of features and kernels to find the best configuration of the two SVM classifiers (one for each language of the task). A 10-fold cross validation was carried out to tune our systems based on accuracy. Our submitted systems configuration can be seen in Table 3.

Run #3 for both languages uses the same configuration of our best system at the IberEval task. (Fersini et al., 2018a).

The best result on the English training set has been obtained by run #1, where we used the RBF kernel (0.765 accuracy), while for Italian the best result has been obtained by runs #2 and #3 with the Linear kernel (0.893 accuracy). Different sets of categories from *HurtLex* were able to improve the classifier performance, depending on the language.

In order to classify the category and target of misogyny (Task B), we adopted the same set of features as Task A. Therefore, we did not build new systems specifically for Task B.

We experimented with different selections of categories from the HurtLex lexicon, and identified the most useful for the purpose of misogyny identification. As it can be seen in Table 3, the main categories are: physical disabilities and diversity (DDP), words related to prostitution (PR), words referring to male genitalia (ASM) and female genitalia (ASF). But also: derogatory words (CDS), words related to felonies and crime, and also immoral behavior (RE).

Language	**English**			**Italian**		
Systems	run1	run2	**run3**	run1	run2	**run3**
Accuracy	0.765	0.72	0.744	0.786	0.893	0.893
Bag of Word	-	✓	-	-	✓	✓
Bag of Hashtags	-	-	-	-	-	✓
Bag of Emojis	-	-	-	-	-	✓
S.W. Count	✓	-	✓	✓	-	-
S.W. Presence	✓	-	✓	✓	-	-
Sexist Slurs	✓	✓	✓	✓	✓	-
Woman Word	✓	✓	✓	✓	✓	-
Hashtag	-	-	✓	-	✓	-
Link Presence	✓	✓	✓	-	-	-
Upper Case Count	✓	-	-	✓	✓	-
Text Length	-	✓	-	✓	-	-
ASF Count	✓	✓	-	✓	✓	✓
PR Count	-	-	-	✓	✓	✓
OM Count	✓	✓	-	-	-	-
DDF Count	-	-	-	-	-	-
CDS Count	✓	✓	-	✓	✓	-
DDP Count	✓	✓	-	-	-	✓
AN Count	✓	✓	-	-	-	-
ASM Count	-	-	-	✓	✓	-
DMC Count	-	-	-	-	-	-
IS Count	✓	✓	-	-	-	-
OR Count	-	-	-	-	-	-
PA Count	✓	✓	-	-	-	-
PS Count	-	-	-	-	-	-
QAS Count	-	-	-	-	-	-
RCI Count	-	-	-	-	-	-
RE Count	-	-	-	✓	✓	-
SVP Count	-	-	-	-	-	-
Kernel	RBF	Linear	RBF	RBF	Linear	Linear

Table 3: Feature Selection for all the submitted systems.

5 Results

Table 4 shows our system performance based on the test sets. Our best system in Task A ranked 3^{rd} in Italian (0.839 in accuracy for run3) and 13^{th} in English (0.621 in accuracy for run3). Interestingly, our best result on both languages were obtained by the best configuration submitted at the IberEval campaign. However, our English system performance was way worse compared to the result of IberEval (accuracy = 0.814). We will try to analyze this problem in the Section 6.

ITALIAN

Rank	Team	Accuracy
1	bakarov.c.run2	0.844
2	bakarov.c.run1	0.842
3	14-exlab.c.run3	0.839
4	bakarov.c.run3	0.836
5	14-exlab.c.run2	0.835
6	StopPropagHate.c.run1	0.835
7	*AMI-BASELINE*	0.830
8	StopPropagHate.u.run2	0.829
9	SB.c.run1	0.824
10	RCLN.c.run1	0.824
11	SB.c.run3	0.823
12	SB.c.run	0.822

ENGLISH

Rank	Team	Accuracy
1	hateminers.c.run1	0.704
2	hateminers.c.run3	0.681
3	hateminers.c.run2	0.673
4	resham.c.run3	0.651
5	bakarov.c.run3	0.649
6	resham.c.run1	0.648
7	resham.c.run2	0.647
8	ITT.c.run2.tsv	0.638
9	ITT.c.run1.tsv	0.636
10	ITT.c.run3.tsv	0.636
11	himani.c.run2.tsv	0.628
12	bakarov.c.run2	0.628
13	14-exlab.c.run3	0.621
14	himani.c.run1.tsv	0.619
15	himani.c.run3.tsv	0.614
16	14-exlab.c.run1	0.614
17	SB.c.run2.tsv	0.613
18	bakarov.c.run1	0.605
19	*AMI-BASELINE*	0.605
20	StopPropagHate.c.run1.tsv	0.593
21	SB.c.run1.tsv	0.592
22	StopPropagHate.u.run3.tsv	0.591
23	StopPropagHate.u.run2.tsv	0.590
24	RCLN.c.run1	0.586
25	SB.c.run3.tsv	0.584
26	14-exlab.c.run2	0.500

Table 4: Official Results for Subtask A.

In Task B, most of the submitted systems struggled to classify the misogynous tweets into the five categories and discriminate whether the target is active or passive. Both subtasks for both languages have very low baselines (below 0.4 for English and around 0.5 for Italian). Several under-represented classes such as DERAILING and DOMINANCE are very difficult to be detected in category classification (See Table 1 for details). Similarly, the label distribution was very unbalanced for target classification, where most of the misogynous tweets are attacking a specific target (ACTIVE).

ITALIAN

Rank	Team	Avg.	Cat.	Targ.
1	bakarov.c.run1	0.493	0.555	0.432
2	*AMI-BASELINE*	0.487	0.534	0.440
3	14-exlab.c.run3	0.485	0.552	0.418
4	14-exlab.c.run2	0.482	0.550	0.415
5	bakarov.c.run3	0.478	0.536	0.421
6	bakarov.c.run2	0.463	0.499	0.426
7	SB.c.run.tsv	0.449	0.485	0.414
8	SB.c.run1.tsv	0.448	0.483	0.414
9	RCLN.c.run1	0.448	0.473	0.422
10	SB.c.run2.tsv	0.446	0.480	0.411
11	14-exlab.c.run1	0.292	0.164	0.420

ENGLISH

Rank	Team	Avg.	Cat.	Targ.
1	himani.c.run3.tsv	0.406	0.361	0.451
2	himani.c.run2.tsv	0.377	0.323	0.431
3	*AMI-BASELINE*	0.370	0.342	0.399
4	hateminers.c.run3	0.369	0.302	0.435
5	hateminers.c.run1	0.348	0.264	0.431
6	SB.c.run2.tsv	0.344	0.282	0.407
7	himani.c.run1.tsv	0.342	0.280	0.403
8	SB.c.run1.tsv	0.335	0.282	0.389
9	hateminers.c.run2	0.329	0.229	0.430
10	SB.c.run3.tsv	0.328	0.269	0.387
11	resham.c.run2	0.322	0.246	0.399
12	resham.c.run1	0.316	0.235	0.397
13	bakarov.c.run1	0.309	0.260	0.357
14	resham.c.run3	0.283	0.214	0.353
15	RCLN.c.run1	0.280	0.165	0.395
16	ITT.c.run2.tsv	0.276	0.173	0.379
17	bakarov.c.run2	0.275	0.176	0.374
18	14-exlab.c.run1	0.260	0.124	0.395
19	bakarov.c.run3	0.254	0.151	0.356
20	14-exlab.c.run3	0.239	0.107	0.371
21	ITT.c.run1.tsv	0.238	0.140	0.335
22	ITT.c.run3.tsv	0.237	0.138	0.335
23	14-exlab.c.run2	0.232	0.205	0.258

Table 5: Official Results for Subtask B.

Several features which focus on the use of offensive words were proven to be useful in English. For Italian, a simple tweet representation which involves Bag of Words, Bag of Hashtags, and Bag of Emojis already produced a better result than the baseline. Some of the HurtLex categories that were improving the system's performance during training did not help the prediction on the test set (ASF, OM, CDS, DDP, AN, IS, PA for English and CDS, ASM for Italian). However, similarly to the Spanish case, the system configuration which utilized ASF, PR, and DDP obtained the best result in Italian.

6 Discussion

We performed an error analysis on the gold standard test set, and analyzed 160 Italian tweets that our best system configuration mislabelled. The label "misogynistic" was wrongly assigned to 147 instances (false positives, 91.9% of the errors), while the contrary happened only 13 times (false negatives, 8.1% of the errors). The same situation happened in the English dataset, but with a less striking impact, with 228 false positives (60.2% of the errors), 151 false negatives (39.8% of the errors). In this section we conduct a qualitative error analysis, identifying and discussing several factors that contribute to the misclassification.

Presence of swear words. We encountered a lot of "bad words" in the dataset of this shared task for both English and Italian. In case of abusive context, the presence of swear words can help to spot abusive content such as misogyny. However, they could also lead to false positives when the swear word is used in a casual, not offensive context (Malmasi and Zampieri, 2018; Van Hee et al., 2018; Nobata et al., 2016). Consider the following two examples containing the swear word "bitch" in different contexts:

> 1. Im such a fucking cunt bitch and i dont even mean to be goddammit

> 2. Bitch you aint the only one who hate me, join the club, stand in the corner, and stfu.

In Example 1, the swear word "bitch" is used just to arouse interest/show off, thus not directly insulting the other person. This is a case of *idiomatic swearing* (Pinker, 2007). In Example 2, the swear word "bitch" is used to insult a specific target in an abusive context, an instance of *abusive swearing* (Pinker, 2007). Resolving swearing context is still a challenging task for automatic system which contributing to the difficulties of this task.

Reported speech. Tweets may contain misogynistic content as an indirect quote of someone else's words, such as in the following example:

> 3. Quella volta che mia madre mi ha detto quella cosa le ho risposto "Mannaggia! Non sarò mai una brava donna schiava zitta e lava! E adesso?!" Potrei morire per il dispiacere.
> → *That time when my mom told me that thing and I answered "Holy s**t! I will never be a good slave who shuts up and cleans! What now?"*

According to task guidelines this should not be labeled as a misogynistic tweet, because it is not the user himself who is misogynistic. Therefore, instances of this type tend to confuse a classifier based on lexical features.

Irony and world knowledge. In Example 3, the sentence "Potrei morire per il dispiacere."[6] is ironic. Humor is very hard to model for automatic systems — sometimes, the presence of figurative language even baffles human annotators. Moreover, external world knowledge is often required in order to infer whether an utterance is ironic (Wallace et al., 2014).

Preprocessing and tokenization. In computer-mediated communication, and specifically on Twitter, users often resort to a language type that is closer to speech, rather than written language. This is reflected in less-than-clean orthography, with forms and expressions that imitate the verbal face-to-face conversation.

> 4. @ @ @ @ x me glob prox2aa colpiran tutti incluso nemicinterno.. esterno colpopiúduro saràculograndе che bevetropvodka e inoltre x questiondisoldi progetta farmezzofallirsudfinitestampe: ciò nnvàben xrchèindebolis
> → *4 me glob next2aa will hit everyone included internalenemy.. external harderhit willbebigass who drinkstoomuchvodka and also 4 mattersofmoney isplanning tomakethesouthfailwithprintings: dis notgood causeweaken*

In Example 4, preprocessing steps like tokenization and stemming are particularly hard to perform, because of the lack of spaces between one word and the other and the confused orthography. Consequently all the classification pipeline is compromised and error-prone.

Gender of the target. As defined in the Introduction, we know that misogyny is a specific type of hateful language, targeting women. However, detecting the gender of the target is a challenging task in itself, especially in Twitter datasets.

> 5. @realDonaldTrump shut the FUCK up you infected pussy fungus.

> 6. @TomiLahren You're a fucking skank!

Both examples use bad words to abuse their targets. However, the first example is labeled as not misogyny since the target is Donald Trump (man), while the second example is labeled as misogyny with the target Tomi Lahren (woman).

[6] Translation: I could die for heartbreak.

7 Conclusions

Here we draw some considerations based on the results of our participation to the EVALITA 2018 AMI shared task. In order to test the multilingual potential of our model, one of the systems we submitted for Italian at EVALITA (run #3) was based on our best model for Spanish at IberEval. Based on the official results, this system performed well for Italian, consisting of features such as: BoW, BoE, BoH and several HurtLex categories specifically related to the hate against women. Concerning English, we obtained lower results in EVALITA in comparison to IberEval with the same system configuration. It is worth mentioning that even if the training set for the AMI EVALITA task was substantially bigger, in absolute terms all the AMI's participants at EVALITA obtained worse scores than the ones obtained by the IberEval's teams.

Acknowledgments

Valerio Basile and Viviana Patti were partially supported by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media-IhatePrejudice*, S1618_L2_BOSC_01).

References

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In *Proc. of the 23rd Int. Conf. on Applications of Natural Language & Information Systems*, pages 57–64. Springer.

Jamie Bartlett, Richard Norrie, Sofia Patel, Rebekka Rumpel, and Simon Wibberley. 2014. Misogyny on twitter. *Demos*.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A Multilingual Lexicon of Words to Hurt. In *Proc. of the 5th Italian Conference on Computational Linguistics (CLiC-it 2018)*, Turin, Italy. CEUR.org.

Romolo Giovanni Capuano. 2007. *Turpia: sociologia del turpiloquio e della bestemmia*. Riscontri (Milano, Italia). Costa & Nolan.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Tullio De Mauro. 2016. Le parole per ferire. *Internazionale*. 27 settembre 2016.

Fabio Fasoli, Andrea Carnaghi, and Maria Paola Paladino. 2015. Social acceptability of sexist derogatory and sexist objectifying slurs across contexts. *Language Sciences*, 52:98–107.

Elisabetta Fersini, Maria Anzovino, and Paolo Rosso. 2018a. Overview of the Task on Automatic Misogyny Identification at IberEval. In *Proceedings of 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, pages 57–64. CEUR-WS.org, September.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018b. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, and Viviana Patti. 2018. 14-ExLab@ UniTo for AMI at IberEval2018: Exploiting Lexical Knowledge for Detecting Misogyny in English and Spanish Tweets. In *Proc. of 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*.

Steven Pinker. 2007. *The stuff of thought: Language as a window into human nature*. Penguin.

Bailey Poland. 2016. *Haters: Harassment, Abuse, and Violence Online*. Potomac Press.

Anna Schmidt and Michael Wiegand. 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*, pages 1–10.

Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2018. Automatic detection of cyberbullying in social media text. *arXiv preprint arXiv:1801.05617*.

Byron C. Wallace, Laura Kertz, Eugene Charniak, et al. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 512–516.