Cross-domain and Cross-lingual Abusive Language Detection: a Hybrid Approach with Deep Learning and a Multilingual Lexicon

Endang Wahyu Pamungkas, Viviana Patti
Dipartimento di Informatica
University of Turin, Italy
{pamungka,patti}@di.unito.it

Abstract

The development of computational methods to detect abusive language in social media within variable and multilingual contexts has recently gained significant traction. The growing interest is confirmed by the large number of benchmark corpora for different languages developed in the latest years. However, abusive language behaviour is multifaceted and available datasets are featured by different topical focuses. This makes abusive language detection a domain-dependent task, and building a robust system to detect general abusive content a first challenge. Moreover, most resources are available for English, which makes detecting abusive language in low-resource languages a further challenge. We address both challenges by considering ten publicly available datasets across different domains and languages. A hybrid approach with deep learning and a multilingual lexicon to cross-domain and cross-lingual detection of abusive content is proposed and compared with other simpler models. We show that training a system on general abusive language datasets will produce a cross-domain robust system, which can be used to detect other more specific types of abusive content. We also found that using the domain-independent lexicon HurtLex is useful to transfer knowledge between domains and languages. In the cross-lingual experiment, we demonstrate the effectiveness of our joint-learning model also in out-domain scenarios.

1 Introduction

Detecting online abusive language in social media messages is gaining increasing attention from scholars and stakeholders, such as governments, social media platforms and citizens. The spread of online abusive content negatively affects the targeted victims, has a chilling effect on the democratic discourse on social networking platforms and negatively impacts those who speak for freedom and non-discrimination. *Abusive language* is usually used as an umbrella term (Waseem et al., 2017), covering several sub-categories, such as cyberbullying (Van Hee et al., 2015; Sprugnoli et al., 2018), hate speech (Waseem and Hovy, 2016; Davidson et al., 2017), toxic comments (Wulczyn et al., 2017), offensive language (Zampieri et al., 2019a) and online aggression (Kumar et al., 2018). Several datasets have been proposed having different topical focuses and specific targets, e.g., misogyny or racism. This diversity makes the task to detect general abusive language difficult. Some studies attempted to bridge some of these subtasks by proposing cross-domain classification of abusive content (Wiegand et al., 2018a; Karan and Šnajder, 2018; Waseem et al., 2018).

Another prominent challenge in abusive language detection is the multilinguality issue. Even if in the last year abusive language datasets were developed for other languages, including Italian (Bosco et al., 2018; Fersini et al., 2018b), Spanish (Fersini et al., 2018b), and German (Wiegand et al., 2018b), most studies so far focused on English. Since most popular social media such as Twitter and Facebook goes multilingual, fostering their users to interact in their primary language, there is a considerable urgency to develop a robust approach for abusive language detection in a multilingual environment, also for guaranteeing a better compliance to governments demands for counteracting the phenomenon (see, e.g., the recently issued EU commission *Code of Conduct on countering illegal hate speech online* (EU Commission, 2016). Cross-lingual classification is an approach to transfer knowledge from resource-rich languages to resource-poor ones. It has been applied to sentiment analysis (Zhou et al., 2016), a related task to abusive language detection. However, there is still not much work focused on cross-

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 363–370
Florence, Italy, July 28 - August 2, 2019.

Dataset	Label	Language	Topical Focus	Train	Test	PIR
Harassment (Golbeck et al., 2017)	**H - harassing**, N - non-harassing	EN	Harassing content, including racist and misogynistic contents, offensive profanities and threats	14,252	6,108	0.26
Waseem (Waseem and Hovy, 2016)	**racism, sexism**, none	EN	Racism and Sexism	11,542	4,947	0.31
OffensEval (Zampieri et al., 2019b)	**OFF - offensive**, NOT - not offensive	EN	Offensive content, including insults, threats, and posts containing profane language or swear words	13,240	860	0.33
HatEval (Basile et al., 2019)	**1 - hateful**, 0 - not hateful	EN, ES	Hate speech against women and immigrants	9,000 (EN) 4,500 (ES)	2,971 (EN) 1,600 (ES)	0.42 0.41
AMI Evalita (Fersini et al., 2018a)	**1 - misogynous**, 0 - not misogynous	EN, IT	Misogynous content	4,000 (EN) 4,000 (IT)	1,000 (EN) 1,000 (IT)	0.45 0.47
AMI IberEval (Fersini et al., 2018b)	**1 - misogynous**, 0 - not misogynous	EN, ES	Misogynous content	3,251 (EN) 3,307 (ES)	726 (EN) 831 (ES)	0.47 0.50
GermEval (Wiegand et al., 2018b)	**offensive**, other	DE	Offensive content, including insults, abuse, and profanity	5,009	3,532	0.34

Table 1: Twitter abusive language datasets in four languages: original labels, language(s) featured, topical focus, distribution of train and test set and positive instance rate (PIR).

lingual abusive language classification.

In this study, we conduct an extensive experiment to explore cross-domain and cross-lingual abusive language classification in social media data, by proposing a hybrid approach with deep learning and a multilingual lexicon. We exploit several available Twitter datasets in different domains and languages. We present three main contributions in this work. First, we characterize the available datasets as capturing various phenomena related to abusive language, and investigate this characterization in cross-domain classification. Second, we explored the use of a domain-independent, multilingual lexicon of abusive words called *HurtLex* (Bassignana et al., 2018) in both cross-domain and cross-lingual settings. Last, we take advantage of the availability of multilingual word embeddings to build a joint-learning approach in the cross-lingual setting. All code and resources are available at https://github.com/dadangewp/ACL19-SRW.

2 Related Work

Some work has been done in the cross-domain classification of abusive language. Wiegand et al. (2018a) proposed to use high-level features by combining several linguistic features and lexicons of abusive words in the cross-domain classification of abusive microposts from different sources. Waseem et al. (2018) use multi-task learning for domain transfer in a cross-domain hate speech detection task. Recently, Karan and Šnajder (2018) also addressed cross-domain classification in several abusive language datasets, testing the framework of Frustratingly Simple Domain Adaptation (FEDA) (Daume III, 2007) to transfer knowledge between domains.

Meanwhile, cross-lingual abusive language detection has not been explored yet by NLP scholars. We only found a few works describing participating systems developed for recent shared tasks on the identification of misogynous (Basile and Rubagotti, 2018) and offensive language (van der Goot et al., 2018), where some experiment in a cross-lingual setting is proposed. Basile and Rubagotti (2018) used the *bleaching* approach (van der Goot et al., 2018) to conduct cross-lingual experiments between Italian and English when participating to the automatic misogyny identification task at EVALITA 2018 (Fersini et al., 2018a). Schneider et al. (2018) used multilingual embeddings in a cross-lingual experiment related to GermEval 2018 (Wiegand et al., 2018b).

3 Data

We consider ten different publicly abusive language datasets and benchmark corpora from shared tasks. Some shared tasks (HatEval, AMI Evalita and AMI IberEval) provided data in two languages. Table 1 summarizes the datasets' characteristics. We binarize the label of these datasets into abusive (bold) and not-abusive. For the cross-lingual experiments, we include datasets from four languages: English, Italian, Spanish, and German. We split all datasets into training and testing by keeping the original split when provided, and splitting the distribution randomly (70% for training and 30% for testing) otherwise.

Dataset		LSVC + BoW				LSVC + BoW + HL				LSTM + WE				LSTM + WE + HL			
Test	Train	P	R	F_1	Δ	P	R	F_1	Δ	P	R	F_1	Δ	P	R	F_1	Δ
Harassment	Waseem	.325	.233	.271	.103	.337	.264	.296	.079	.291	.467	.359	.033	.290	.524	**.373**	.045
	HatEval	.389	.119	.183	.191	.374	.116	.177	.198	.341	.308	.324	.068	.332	.379	**.354**	.064
	OffensEval	.320	.508	.393	-.019	.322	.516	.396	-.021	.333	.443	.380	.012	.314	.567	**.404**	.014
	Harassment	.547	.284	.374		.540	.288	.375		.510	.319	.392		.464	.380	.418	
Waseem	Harassment	.729	.022	.043	.688	.720	.034	.065	.669	.464	.111	.179	.587	.491	.149	**.229**	.520
	HatEval	.620	.109	.186	.545	.672	.113	.194	.540	.496	.213	.299	.467	.453	.318	**.374**	.375
	OffensEval	.461	.390	**.422**	.309	.453	.391	.420	.314	.444	.282	.345	.421	.419	.411	.415	.334
	Waseem	.817	.662	.731		.819	.665	.734		.760	.771	.766		.711	.790	.749	
HatEval	Harassment	.485	.181	.264	.339	.513	.229	.317	.290	.523	.308	.387	.216	.514	.394	**.446**	.158
	Waseem	.505	.490	.497	.106	.477	.558	.514	.093	.481	.636	**.548**	.055	.494	.609	.546	.058
	OffensEval	.450	.646	.531	.072	.451	.656	.534	.073	.452	.603	.516	.087	.457	.704	**.554**	.050
	HatEval	.449	.919	.603		.453	.919	.607		.444	.939	.603		.441	.955	.604	
OffensEval	Harassment	.301	.104	.155	.422	.321	.113	.167	.406	.525	.133	.213	.395	.406	.179	**.249**	.349
	Waseem	.440	.246	.316	.261	.462	.254	**.328**	.245	.403	.225	.289	.319	.400	.175	.244	.354
	HatEval	.372	.225	.281	.296	.381	.233	.289	.284	.392	.371	.381	.227	.371	.529	**.436**	.162
	OffensEval	.616	.542	.577		.626	.529	.573		.667	.558	.608		.551	.654	.598	

Table 2: Results on cross-domain abusive language identification (only in English).

We also provide further information about the captured phenomena of every dataset. Based on this information, we can compare the nature and topical focus of the dataset, which potentially affect the cross-domain experimental results. Some datasets have a broader coverage than the others, focussing on more general phenomena, such as OffensEval (Zampieri et al., 2019b), and GermEval (Wiegand et al., 2018b). However, there are also some shared phenomena between datasets, such as racism and sexism in Waseem (Waseem and Hovy, 2016) and HatEval (Basile et al., 2019). AMI datasets contain the most specific phenomenon, only focusing on misogyny. The positive instance rate (PIR) denotes the ratio of abusive instances to all instances of the dataset.

4 Cross-domain Classification

In this experiment, we investigate the performance of machine learning classifiers which are trained on a particular dataset and tested on different datasets ones. We focus on investigating the influence of captured phenomena coverage between datasets. We hypothesize that a classifier which is trained on a broader coverage dataset and tested on narrower coverage dataset will give better performance than the opposite. Furthermore, we analyse the impact of using the HurtLex lexicon (Bassignana et al., 2018) to transfer knowledge between domains. HurtLex is a multilingual lexicon of hate words, originally built from 1,082 Italian hate words compiled in a manual fashion by the linguist Tullio De Mauro (De Mauro, 2016). This lexicon is semi-automatically extended and translated into 53 languages by using BabelNet (Navigli and Ponzetto, 2012), and the lexical items are divided into 17 categories such as homophobic slurs, ethnic slurs, genitalia, cognitive and physical disabilities, animals and more[1].

Model. In this experiment, we employ two models. First, we exploit a simple traditional machine learning approach by using linear support vector classifier (LSVC) with unigram representation as a feature. Second, we utilize a long short-term memory (LSTM) neural model consisting of several layers, starting with a word embedding layer (32-dimensions) without any pre-trained model initialization[2]. This embedding layer is followed by LSTM networks (16-units), whose output is passed to a dense layer with ReLU activation function and dropout (0.4). The last section is a dense layer with sigmoid activation to produce the final prediction. We experiment with HurtLex by concatenating its 17 categories as one hot encoding representation to both LSVC-based and LSTM-based systems.

Data and Evaluation We use four English datasets, namely Harassment, Waseem, HatEval, and OffensEval [3]. We evaluate the system performance based on precision, recall, and F-score on the positive class (abusive class).

Results. Table 2 shows the results of the cross-domain experiment. We test every dataset with three systems which are trained on three other datasets. We also run in-domain scenario to compare the delta between in-domain and out-domain performance and measure the drop in per-

[1] `http://hatespeech.di.unito.it/resources.html`

[2] We experimented the use of pre-trained models (i.e. GloVe, word2vec, and FastText), but the result is lower compared to a self-trained model based on training set.

[3] AMI datasets are excluded due to the low number of instances.

formance. Not surprisingly, the performance on out-domain datasets is always lower (except in two cases when the Harassment dataset is used as test set). Overall, LSTM-based systems performed better than LSVC-based systems. The use of HurtLex also succeeded in improving the performance on both LSVC-based and LSTM-based systems. We can see that HurtLex is able to improve the recall in most of the cases. Our further investigation shows that systems with HurtLex are able to detect more abusive contents, noted by the increases of true positives. The OffensEval training set always achieves the best performance when tested on three other datasets. On the other hand, the Harassment dataset always presents the larger drop in performance when used as training data. Training the models on the Harassment dataset lead to a very low result even in the in-domain setting. The highest result on the Harassment dataset is only .418 F-score, achieved by LSTM with HurtLex [4], while when trained on the other datasets our models are able to reach above .600 F-score. Upon further investigation, we found, that Golbeck et al. (2017) only used a limited set of keywords, which contributes to limit their dataset coverage. Overall, we argue that there are good arguments in favor of our hypothesis that a system trained on datasets with a broader coverage of phenomena will be more robust to detect other kinds of abusive language (see the OffensEval results).

5 Cross-lingual Classification

We aim to experiment with cross-lingual abusive language classification. As far as our knowledge goes, there is still no work which focuses on investigating the feasibility of cross-lingual classification in the abusive language area. We will explore two scenarios, in-domain and out-domain classification, in four different languages, namely English, Spanish, Italian, and German. Again, we will test HurtLex in this experiment.

Model. We build four systems for each in-domain and out-domain experiments. One system of each scenario is built based on LSVC with unigram features, while three other systems are built based on a LSTM architecture. Here we describe three systems which are based on LSTM:

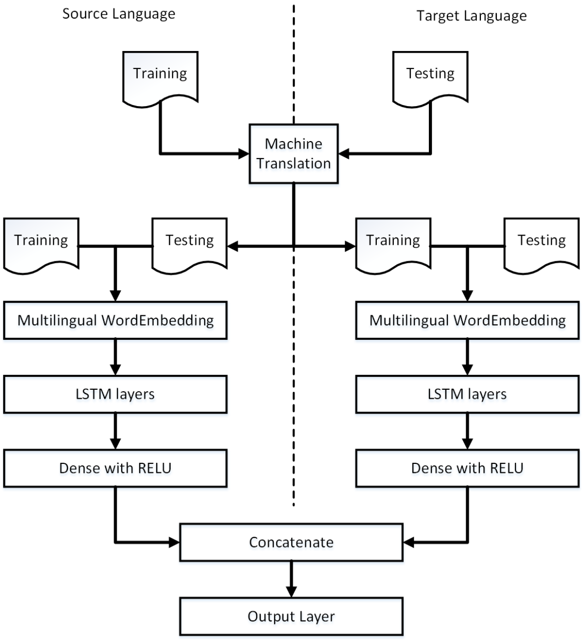

Figure 1: Joint-learning model architecture.

(a). **LSTM + WE.** First, we exploit LSTM with monolingual word embedding. We adopt a similar model as in cross-domain classification where we use machine translation (Google Translate[5]) to translate training data from source to target language. In this model, we use pre-trained word embedding from FastText[6].

(b). **JL + ME.** We also propose a joint-learning model with multilingual word embedding. We take advantage of the availability of multilingual word embeddings[7] to build a joint-learning model. Figure 1 summarize how the data is transformed and learned in this model. We create bilingual training data automatically by using Google Translate to translate the data in both directions (training from source to target language and testing from target to source language), then using it as training data for the two LSTM-based architectures (similar architecture of the model in cross-domain experiment). We concatenate these two architectures before the output layer, which produces the final prediction. In the, we expect to reduce some of the noise from the translation while keeping the original structure of the training set.

[4] Marwa et al. (2018) claimed to get a higher result, but that paper did not give a complete information about system configuration they used.

[5] http://translate.google.com/

[6] https://fasttext.cc/

[7] https://github.com/facebookresearch/MUSE

Dataset		LSVC + BoW				LSTM + WE				JL + ME			JL + ME + HL		
Test	Train	P	R	F_1	Δ	P	R	F_1	Δ	P	R	F_1	P	R	F_1
EN-Evalita	IT-Evalita	.491	.739	.590	.004	.479	.824	.605	.019	.480	.935	**.635**	.501	.761	.605
	ES-IberEval	.561	.704	.624	-.030	.551	.615	.581	.081	.550	.711	.620	.543	.763	**.635**
	EN-Evalita	.557	.637	.594		.518	.917	.662		-	-	-	-	-	-
IT-Evalita	EN-Evalita	.209	.125	.156	.698	.179	.129	.150	.682	.453	.520	.484	.491	.502	**.497**
	ES-IberEval	.611	.611	.611	.243	.583	.287	.385	.447	.698	.387	.506	.666	.654	**.660**
	IT-Evalita	.786	.934	.854		.714	.996	.832		-	-	-	-	-	-
ES-IberEval	EN-Evalita	.640	.545	.589	.151	.524	.829	.642	.118	.627	.721	.670	.604	.798	**.687**
	IT-Evalita	.575	.528	.550	.190	.474	.455	.464	.296	.587	.636	.610	.586	.696	**.637**
	ES-IberEval	.739	.742	.740		.761	.759	.760		-	-	-	-	-	-

Table 3: Results on in-domain (AMI) cross-lingual abusive language identification (EN, ES,IT).

Dataset		LSVC + BoW				LSTM + WE				JL + ME			JL + ME + HL		
Test	Train	P	R	F_1	Δ	P	R	F_1	Δ	P	R	F_1	P	R	F_1
EN-Waseem	ES-HatEval	.498	.353	.413	.318	.519	.524	.522	.244	.591	.414	.487	.532	.523	**.528**
	IT-Evalita	.470	.248	.325	.406	.481	.199	.282	.484	.497	.156	.238	.566	.311	**.401**
	DE-GermEval	.547	.323	.406	.325	.505	.388	**.439**	.327	.545	.182	.273	.350	.456	.396
	EN-Waseem	.817	.662	.731		.760	.771	.766		-	-	-	-	-	-
ES-HatEval	EN-Waseem	.464	.286	.354	.308	.489	.323	.389	.284	.332	.708	**.452**	.426	.351	.384
	IT-Evalita	.517	.234	.323	.239	.620	.443	.517	.156	.626	.506	.559	.602	.647	**.623**
	DE-GermEval	.495	.429	.459	.203	.450	.503	.475	.198	.510	.341	.409	.516	.446	**.478**
	ES-HatEval	.606	.730	.662		.615	.744	.673		-	-	-	-	-	-
IT-Evalita	EN-Waseem	.311	.700	.431	.423	.300	.709	.422	.410	.306	.836	**.448**	.301	.743	.428
	ES-HatEval	.502	.538	.519	.335	.424	.724	.534	.298	.439	.829	**.574**	.462	.724	.564
	DE-GermEval	.569	.268	.364	.490	.486	.377	.425	.407	.369	.730	.490	.593	.590	**.592**
	IT-Evalita	.786	.934	.854		.714	.996	.832		-	-	-	-	-	-
DE-GermEval	EN-Waseem	.442	.178	.254	.196	.421	.189	.261	.311	.436	.136	.208	.456	.188	**.266**
	ES-HatEval	.438	.254	.321	.129	.398	.607	.481	.091	.361	.726	**.482**	.395	.359	.377
	IT-Evalita	.371	.656	.474	-.024	.369	.730	.490	.082	.362	.862	**.510**	.354	.909	.509
	DE-GermEval	.578	.369	.450		.799	.446	.572		-	-	-	-	-	-

Table 4: Results on out-domain cross-lingual abusive language identification (EN, ES, IT, DE).

(c). **JL + ME + HL.** Finally, we also experiment the use of HurtLex in our joint-learning model, by simply concatenating its representation into both LSTM model in source and target language.

Dataset and Evaluation We use the AMI datasets (with topical focus on misogyny identification) for the in-domain experiment, in three languages, i.e. English (EN-Evalita), Spanish (ES-Ibereval), and Italian (IT-Evalita). For English, we decide to use the Evalita one due to its larger size. For the out-domain experiment, we use Waseem (EN), HatEval (ES), AMI-Evalita (IT-Evalita in the table, IT), and GermEval (DE). We use precision, recall, and F-score in positive class as evaluation metric.

Results. Table 3 shows the results of the in-domain experiments, while out-domain results can be seen in Table 4. For the in-domain experiment, our joint-learning based systems are able to outperform two other systems based on LSVC and LSTM with monolingual embeddings. Furthermore, HurtLex succeeded to improve the system performance, except when systems are tested on English datasets. LSCV models were outperformed by deep learning-based systems in the out-domain experiment. Our joint-learning based system always gives the best performance on all settings (except when trained on GermEval and tested on Waseem, where LSTM with monolingual embeddings performs better). HurtLex is only able to improve 7 out of 12 results based on F-score, where in most cases it succeeds to improve the recall. This result is consistent with in cross-domain experiments in Section 3. The out-domain results are generally lower than in-domain ones. A lot of variables could influence the difficulty of the out-domain scenario, which calls for deeper investigations. Some of them are discussed in Section 6.

6 Discussion

We discuss some of the challenges which contribute to make the cross-domain and cross-lingual abusive language detection tasks difficult. In particular we will focus on some issues related to the presence of swear words in these kinds of texts.

The different uses of swear words. As described in Section 3, the datasets we considered have different focuses w.r.t. the abusive phenomena captured, and this impacts on the lexical distribution in each dataset. Based on a further analysis we observed that in datasets with a general topical focus such as OffensEval, the abusive tweets are marked by some common swear words such as “fuck”, “shit”, and “ass”. While in datasets featured by a specific hate target, such as the AMI dataset (misogyny), the lexical keywords in abusive tweets are dominated by specific sexist slurs such as “bitch”, “cunt”, and “whore”. This finding is consistent with the study of (ElSherief et al., 2018), which conducted an analysis on hate speech in social media based on its target. Furthermore, the pragmatics of swearing could also change from one dataset to another, depending on the topical features.

Translation issues. As we expected, the use of automatic machine translation (via Google Translate) in our pipeline can give rise to errors in the cross-lingual setting. In particular, we observed errors in translations from English to other languages (Italian and Spanish) on some swear words, which are usually important clues to detect abusive content. See for instance the following cases from the EN-AMI Evalita dataset:

> *Original tweet (EN):*
> Punch that girl right in the **skank**
> *Translated tweet (IT):*
> Pugno quella ragazza proprio nella **Skank**
>
> *Original tweet (EN):*
> Apparently, you can turn a **hoe** into a housewife
> *Translated tweet (ES):*
> Aparentemente, puedes convertir una **azada** en una ama de casa.

Translating swearing is indeed challenging. In the first example, Google Translate is unable to provide an Italian translation for the English word “skank” (a proper translation could be “sciacquetta” or “sciattona”, which means “slut”). We found 134 occurrences of the word “skank” in EN-AMI Evalita and 185 in the EN-HatEval dataset. The second example shows, instead, a problem related to context and disambiguation issues. Indeed, the word “hoe” here is used informally in its derogatory sense, meaning *“A woman who engages in sexual intercourse for money”* (synonyms: slut, whore, floozy)[8]. But, disregarding the context, it is translated in Spanish by relying on a different conventional meaning (hoe as *agricultural and horticultural hand tool*). The term “hoe” is also very frequent in the EN-AMI Evalita (292 occurrences) and EN-HatEval dataset (348 occurrences).

7 Conclusion and Future Work

In this study, we conduct an exploratory experiment on abusive language detection in cross-domain and cross-lingual classification scenarios. We focus on social media data, exploiting several datasets across different domains and languages. Based on the cross-domain experiments, we found that training a system on datasets featured by more general abusive phenomena will produce a more robust system to detect other more specific kinds of abusive languages. We also observed that HurtLex is able to transfer knowledge between domains by improving the number of true positives. In the cross-lingual experiment, our joint-learning systems outperformed the other systems in most cases also in the out-domain setting. The results presented here succeed to shed some light regarding the issues and difficulties of this research direction. As future work, we aim at exploring more deeply the issue related to different coverage, topical focuses and abusive phenomena characterizing the datasets in this field, taking a semantic ontology-based approach to clearly represent the relations between concepts and linguistic phenomena involved. This will allow us to further explore and refine the idea that combining some datasets can produce a more robust system to detect abusive language across different domains. We also found that detecting out-domain abusive content cross-lingual is really challenging, and the use of domain-independent resources to transfer knowledge between domains and languages an interesting issue to be further explored. Finally, we will further investigate the different uses and contexts of swearing, which seems to play a key role in the abusive language detection task (Holgate et al., 2018), with impact also on experiments in cross-domain and cross-lingual settings.

Acknowledgements

We would like to deeply thank Valerio Basile for his valuable and useful feedback on earlier versions of this paper. The work of Endang W. Pamungkas and Viviana Patti was partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01).

[8] https://www.urbandictionary.com/define.php?term=Hoe

References

Angelo Basile and Chiara Rubagotti. 2018. Crotonemilano for AMI at evalita2018. A performant, cross-lingual misogyny detection system. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 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, Minneapolis, Minnesota, USA. ACL.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*.

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the EVALITA 2018 hate speech detection task. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018.*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017.*, pages 512–515. AAAI Press.

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

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth M. Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018.*, pages 42–51. AAAI Press.

EU Commission. 2016. Code of conduct on countering illegal hate speech online.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the evalita 2018 task on automatic misogyny identification (AMI). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018.*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the task on automatic misogyny identification at ibereval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018.*, volume 2150 of *CEUR Workshop Proceedings*, pages 214–228. CEUR-WS.org.

Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference, WebSci 2017, Troy, NY, USA, June 25 - 28, 2017*, pages 229–233. ACM.

Rob van der Goot, Nikola Ljubesic, Ian Matroos, Malvina Nissim, and Barbara Plank. 2018. Bleaching text: Abstract features for cross-lingual gender prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 383–389. Association for Computational Linguistics.

Eric Holgate, Isabel Cachola, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Why swear? analyzing and inferring the intentions of vulgar expressions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4405–4414, Brussels, Belgium. Association for Computational Linguistics.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. 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. Association for Computational Linguistics.

Tolba Marwa, Ouadfel Salima, and Meshoul Souham. 2018. Deep learning for online harassment detection in tweets. In *2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pages 1–5. IEEE.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Julian Moreno Schneider, Roland Roller, Peter Bourgonje, Stefanie Hegele, and Georg Rehm. 2018. Towards the automatic classification of offensive language and related phenomena in german tweets. In *14th Conference on Natural Language Processing KONVENS 2018*, page 95.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a WhatsApp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59, Brussels, Belgium. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Zeerak Waseem, James Thorne, and Joachim Bingel. 2018. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. *Online Harassment*, page 29.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018a. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018b. Overview of the germeval 2018 shared task on the identification of offensive language. In *14th Conference on Natural Language Processing KONVENS 2018*, page 1.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399. International World Wide Web Conferences Steering Committee.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1412, Berlin, Germany. Association for Computational Linguistics.