Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pages 106–112
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

FlorUniTo@TRAC-2: Retrofitting Word Embeddings on an Abusive Lexicon for Aggressive Language Detection

Anna Koufakou, Valerio Basile, Viviana Patti
Florida Gulf Coast University, University of Turin, University of Turin
akoufakou@fgcu.edu, valerio.basile@unito.it, viviana.patti@unito.it

Abstract

This paper describes our participation to the TRAC-2 Shared Tasks on Aggression Identification. Our team, FlorUniTo, investigated the applicability of using an abusive lexicon to enhance word embeddings towards improving detection of aggressive language. The embeddings used in our paper are word-aligned pre-trained vectors for English, Hindi, and Bengali, to reflect the languages represented in the shared task datasets. The embeddings are retrofitted to a multilingual abusive lexicon, HurtLex. We experimented with an LSTM model using the original as well as the transformed embeddings and different language and setting variations. Overall, our systems placed toward the middle of the official rankings based on weighted F1 score. Furthermore, the results on the development and test sets show promise for this novel avenue of research.

Keywords: embeddings, retrofitting, abusive lexicon

1. Introduction

Abusive language is a broad term encompassing several linguistic patterns linked to phenomena such as offensive language, aggressive language or hate speech. Abusive language is a strong signal to detect problematic use of languages, e.g., in cases of cyberbullying, misogyny, racism, or trolling. *Aggressive language* is any form of natural language written or spoken with the intention of hurt. It is typically offensive, although a growing number of studies are recently modeling covert, or implicit, abuse (Caselli et al., 2020).

In the Natural Language Processing (NLP) field, the automatic detection of abusive language, and related phenomena such as aggressiveness and offensiveness, is traditionally approached in a supervised fashion, with or without the support of language resources such as lexicons and dictionaries. A large part of recent research in many NLP tasks has employed deep learning based on word or character embeddings, for example, see Gambäck and Sikdar (2017), Pavlopoulos et al. (2017), Mishra et al. (2018), and Zhang et al. (2018), among others. Abusive language detection is no exception, as highlighted by the characteristics of the participating systems to the most recent and popular evaluation campaigns for offensive language (Zampieri et al., 2019) and hate speech detection (Basile et al., 2019).

In this paper, we describe the systems we submitted for detecting aggression in the context of the TRAC-2 shared task (Kumar et al., 2020) This task was designed as a two-fold open challenge on detecting aggression in English, Hindi, and Bengali social media posts and then detecting misogynistic aggression in the same posts. Our systems utilize word embeddings that are retrofitted to an abusive language lexicon (Bassignana et al., 2018) and then used by an Long Short-Term memory (LSTM) network model to predict labels. We retrofitted the word embeddings so that words that are found in the same categories in the lexicon end up closer together in the vector space. The retrofitting technique has been applied before for semantic lexicons, however it has never been applied to hate or abusive lexicons or similar.

2. Related Work

Recent related work in this field has focused on various tasks in different languages, for example: abuse (Waseem et al., 2017), gender- or ethnic-based hate speech (Basile et al., 2019), misogyny (Fersini et al., 2018) and aggression (Kumar et al., 2018), among others. The methods in recent literature utilize deep learning based on a plethora of models (e.g. RNNs or CNNs or BERT, see for instance Mishra and Mishra (2019) on Hate Speech Identification) and have used character or word embeddings, subword units, etc. A recent survey is given by Mishra et al. (2019).

Language resources also provide substantial support to tasks like abusive language detection (Wiegand et al., 2018), misogyny identification (Pamungkas et al., 2018a; Pamungkas et al., 2018b) or hate speech detection (Davidson et al., 2017). **HurtLex**[1] (Bassignana et al., 2018) is a multilingual lexicon of offensive words, created by semi-automatically translating a handcrafted resource in Italian by linguist Tullio De Mauro (called *Parole per Ferire*, "words to hurt" (De Mauro, 2016)) into 53 languages. Lemmas in HurtLex are associated to 17 non-mutually exclusive categories, plus a binary macro-category indicating whether the lemma reflects a stereotype. The number of lemmas in any language of HurtLex is in the order of thousands, depending on the language, and they are divided into the four principal parts of speech: noun, adjective, verb, and adverb. The lexicon includes *conservative* entries with a higher level of confidence of being offensive, and *inclusive* entries. In this work, we employ HurtLex version 1.2, comprising 8,228 entries for English, 2,209 for Hindi, and 994 for Bengali. It should be noted that code-switching is a phenomenon in HurtLex, with many English lemmas present in the Hindi and Bengali lexicons.

Retrofitting. Although embeddings have been shown to be successful in the wider field of NLP and specifically in abusive language detection, they do not take into account semantic relationships among words, such as synonyms or antonyms. One well-cited technique that addresses this issue is the retrofitting technique proposed by Faruqui et al.

[1] https://github.com/valeriobasile/hurtlex

(2015). This technique uses belief propagation to transform the original embeddings based on relationships it finds in a lexicon so that words that are related end up closer together in the vector space. The original paper used semantic lexicons such as the Paraphrase database (Ganitkevitch et al., 2013, PPDB) to extract synonym relationships, while in this paper we utilize an abusive lexicon and leverage its categorization of the words. Earlier work of an author of this paper examined the use of retrofitting in the context of abusive language detection (Koufakou and Scott, 2019; Koufakou and Scott, 2020), but that work used semantic lexicons, rather than an abusive lexicon. Outside of this field, retrofitting has been successfully applied in other applications, for example the classification of cancer pathology reports by Alawad et al. (2018), utilizing medical resources for a lexicon.

More recently, other methods have been presented that are related to retrofitting: for example, Mrkšić et al. (2017) proposed ATTRACT-REPEL, which utilizes the semantic lexicon to use antonym in addition to synonym relationships. Such methods, however, are based on opposition relations, which are unfit to be adapted to a resource like a hate lexicon. Therefore, we found that the retrofitting method is the most efficient and easy to implement for our purpose, being applicable to a hate lexicon with slight modifications.

3. Methodology and Data

The multilingual annotated data provided by the TRAC-2 workshop organizers are described in Bhattacharya et al. (2020). They included data in three different languages: English, Hindi, and Bengali. The shared task comprises two sub-tasks: sub-task A was Aggression Identification and sub-task B was Misogynistic Aggression Identification. For sub-task A, the data provided were labeled as "Overtly Aggressive" (OAG), "Covertly Aggressive" (CAG) and "Non-aggressive" (NAG). The data came as 5,000 annotated records from social media each in Bangla (in both Roman and Bangla script), Hindi (in both Roman and Devanagari script) and English for training and validation (development set). The data for sub-task B were the same records as for sub-task A with annotations for "Gendered" (GEN) or "Non-Gendered" (NGEN).

We used the TRAC-2 data for all experiments, and also augmented the train set with training data from TRAC-1 (Kumar et al., 2018), the first edition of the shared task, when applicable. For example, we used the English train data from TRAC-1 and English train data from TRAC-2 for sub-task A English. In contrast, we did not use any additional train data for the Bengali tasks (A or B), as it was not available in TRAC-1. Table 1 shows the distribution of the three labels for each of the sets we used for training (train) and validation (dev) related to sub-task A. As shown in the table, the 'Non Aggressive' (NAG) label is the vast majority for all sets, except in the case of the augmented Hindi train dataset (denoted in the table as HIN train++).

Table 2 shows the distribution of GEN versus NGEN labels for the data and sub-task B. As we see in Table 2, the data are very imbalanced with the 'Non Gendered' (NGEN) class as the vast majority label. For this sub-task, we only used TRAC-2 data, as TRAC-1 did not have specific gender labels for their data.

Set	OAG	CAG	NAG	Total
EN train	435	453	3,375	4,263
EN train^{++}	3,143	4,693	8,426	16,262
EN dev	113	117	836	1,066
HIN train	910	829	2,245	3,984
HIN train^{++}	5,766	5,698	4,520	15,984
HIN dev	208	211	578	997
BEN train	850	898	2,078	3,826
BEN dev	217	218	522	957

Table 1: Label distribution for datasets used in Sub-task A: Aggression Identification. $^{++}$ denotes that the train set from TRAC-2 was augmented with the equivalent TRAC-1 train set.

Set	GEN	NGEN	Total
EN train	309	3,954	4,263
EN dev	73	993	1,066
HIN train	661	3,323	3,984
HIN dev	152	845	997
BEN train	712	3,114	3,826
BEN dev	191	766	957

Table 2: Label distribution for datasets used in Sub-task B: Misogyny Identification. Only TRAC-2 data was used.

For our experiments, we started with pre-processing and tokenizing the text. For pre-processing the text, we used the Ekphrasis tool (Baziotis et al., 2017) and regular expressions adapted from Raiyani et al. (2018), a paper from TRAC-1. We also normalized emojis[2] and applied basic tokenization. We applied pre-trained embedddings to the resulting vocabulary. The embeddings that worked the best for the data according to our experimentation were the models provided by FastText[3]. We specifically used the 25-dimensional[4] aligned version of the word embeddings in English, Hindi, and Bengali, in order to encode code-switched messages.

A few examples of the data provided by the shared task that show the code-switching and different labels are included below:

> *jitne wrong kah rhe hn wo sare bi sexual hn.because its prove that all homophofic are always homosexual* (from the English train set, NAG, and GEN)

> *Bhai I just hope that jab aapki beti college ke pehle din entrance exam me rank laake admission le tab koi chutiya bewdaa class me jaake ye na bole ki wo meri property hai,sab durr rehna usse.* (from the Hindi train set, OAG and NGEN)

[2] https://pypi.org/project/emoji

[3] https://fasttext.cc

[4] We started experimenting with larger embeddings, but, due to time constraints, we participated to the shared task with the 25-dimensional setting. We are currently carrying out further experiments with larger models.

Best review. Khup negatives reviews milale Kabir singh la.. Filmi corporation chya Suchitra tyagi tar Vish okalay tichya review madhe.. Go to masses man.. (from the English train set, truncated, NAG and NGEN)

We experimented with retrofitting these embeddings to the HurtLex lexicon (Bassignana et al., 2018). In particular, we considered the relationship between words that belong to the same category in HurtLex, and applied retrofitting based on such symmetric relation. We only considered "conservative" entries in HurtLex, which are supposed to contain less ambiguous terms and therefore less noise, although inducing a smaller coverage.

For each word in our vocabulary, we looked up the relative lemma in HurtLex and found the unique categories the word belongs to. We then created a set of words, which is the union of all words in these categories. This set of words becomes the lexicon for retrofitting. Finally, for all the vectors corresponding to these words, we applied a retrofitting process using code similar to the code found online provided by the original paper[5]. We kept all constants and steps in the method the same as in the original code.

As the data came in three different languages, we experimented with different combinations of languages for the embeddings: for example, we applied English-only pre-trained embeddings to English data, as well as English and Hindi pre-trained embeddings to English data. Also, as Hurtlex contains lexicons for different languages, we were able to experiment with English, Hindi, and Bengali combinations for the retrofitting as well. For example, we used English and Hindi word aligned vectors, for all terms that have a match in our vocabulary. We then retrofitted these vectors using first Hurtlex English and then Hurtlex Hindi as the lexicon.

We implemented our models with the Keras library for Python. First we used an Embedding Layer with Trainable set to True, which fed into an LSTM with 8 nodes. This was followed by a dropout of 0.5, and finally a dense layer with softmax or sigmoid activation, corresponding to the sub-task. As loss functions, we used categorical cross entropy or binary entropy according to the sub-task, Adam optimizer, 10 epochs, and batch size of 64 (if we used only the TRAC-2 train set) or 256 (if we augmented the train set with the equivalent TRAC-1 train set).

Among the submissions for Sub-task A in Bengali, we also introduced a baseline system based on a Support Vector Machine trained on unigrams and TF-IDF, for comparison, since three runs were allowed for submission.

4. Results

We participated to both sub-tasks in all three languages provided by the shared task. In all settings, our systems ranked toward the middle of the official rankings based on weighted F1 score.

4.1. Sub-task A: Aggression Identification

In sub-task A, our systems ranked 9th out of 16 in English, and 5th out of 10 in Hindi and Bengali. Table 3 shows the results of all the runs we submitted. We included accuracy as well as macro-averaged F1 score in order to get a clearer picture of the experimental results. We also provided the weighted F1 for the best system as it was provided by the organizers of the task (weighted F1 was the only metric provided for other systems).

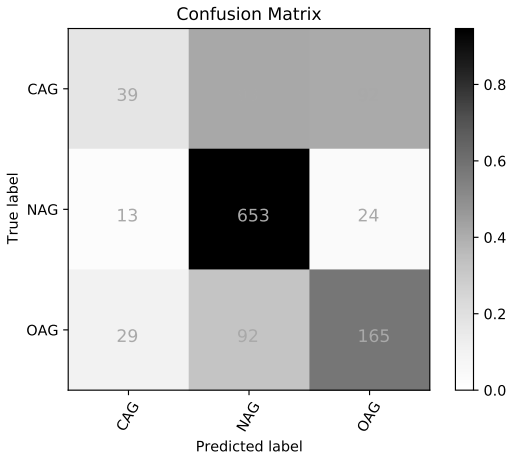

Figure 1: Confusion matrix for our best run in Sub-task EN-A (EN+HIN embeddings, no retrofitting).

In this task, we used several combinations of pre-trained word embeddings from FastText, and retrofitted them on HurtLex extracting the categories of words in one or more language at a time. The combinations that worked better during development, therefore subject to the final submission to the shared task, all involved English, even on the Hindi and Bengali data, probably because the beneficial effect of the larger coverage of resources in such language. For example, using only Bengali FastText word-aligned vectors covered only 25 percent of the vocabulary in the BEN-A sub-task data, but after adding the English FastText word-aligned vectors the coverage of the vocabulary was 60 percent.

Moreover, for some runs, we concatenated the training set of TRAC-1 to the training data, which proved beneficial in the Hindi case according to weighted F1 and to English according to macro F1. We observed that the performance on the development set was always improved by augmenting the train set with the TRAC-1 data. Additionally, the effect of retrofitting in this task is mixed, sometimes helping the performance, while other times lowering it.

Another interesting observation comes from the confusion matrices. For English sub-task A (see Figure 1), the matrix shows NAG ('Non Aggressive') as the large majority, which is what we observed in the TRAC-2 train and development sets (see Table 1). In contrast, for Hindi sub-task A, the confusion matrix (see Figure 2) shows that OAG ('Overtly Aggressive') is the larger class. Moreover, our best models are slightly biased towards overt aggression in English and Bengali (Figures 1 and 3), but biased towards covert aggression in Hindi (Figure 2).

[5] https://github.com/mfaruqui

Language	System Embeddings	HurtLex	Augmented with	F1 (weighted)	F1 (macro)	Accuracy
English	EN+HIN			**.677**	.564	**.714**
	EN+HIN	EN+HIN		.622	.512	.670
	EN+HIN	EN+HIN	TRAC-1	.676	**.585**	.676
	Best TRAC-2 system			.802	-	-
Hindi	EN+HIN			.725	**.650**	.714
	EN+HIN	EN+HIN		.705	.629	.695
	EN+HIN	EN+HIN	TRAC-1	**.726**	.649	**.720**
	Best TRAC-2 system			.813	-	-
Bengali	none (SVM)			.742	.671	.758
	EN+BEN			**.746**	**.672**	**.763**
	EN+BEN	EN+BEN		.730	.644	.750
	Best TRAC-2 system			.821	-	-

Table 3: Results for Sub-task A: Aggression Identification.

Language	System Embeddings	HurtLex	F1 (weighted)	F1 (macro)	Accuracy
English	EN	EN	.830	.628	.847
	EN	EN (5 cat.)	.829	.620	.848
	EN+HIN	EN+HIN (5 cat.)	**.838**	**.649**	**.852**
	Best TRAC-2 system		.871		
Hindi	EN+HIN		.770	.768	**.774**
	EN+HIN	EN+HIN (5 cat.)	**.771**	**.769**	**.774**
	HIN	HIN (5 cat.)	.762	.760	.765
	Best TRAC-2 system		.878		
Bengali	EN+BEN		**.869**	**.761**	.872
	EN+BEN	EN+BEN (5 cat.)	.867	.748	**.877**
	BEN	BEN (5 cat.)	.860	.736	.870
	Best TRAC-2 system		.939		

Table 4: Results for Sub-task B: Misogynistic Aggression Identification.

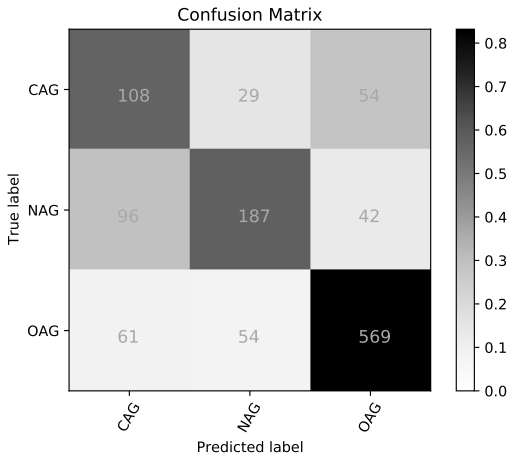

Figure 2: Confusion matrix for our best run in Sub-task HIN-A (EN+HIN embeddings, retrofitted on EN+HIN Hurtlex, additional data from TRAC-1).

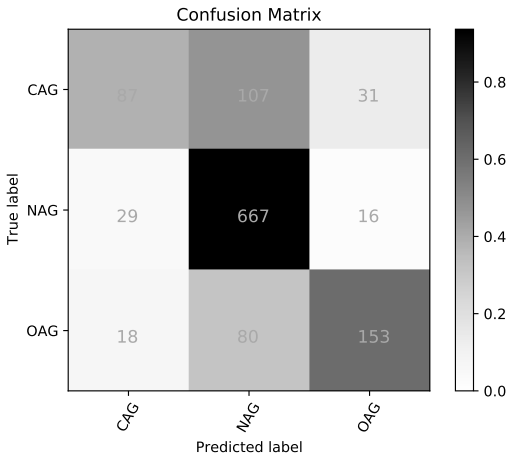

Figure 3: Confusion matrix for our best run in Sub-task BEN-A (EN+BEN embeddings, no retrofitting).

4.2. Sub-task B: Misogynistic Aggression Identification

Similarly to the previous section, Table 4 shows the results of all the systems we submitted for sub-task B. In sub-task B, our systems ranked 9th out of 15 for English, 6th out of 10 for Hindi, and 7th out of 8 for Bengali (all the rankings were based on weighted F1 score).

For this sub-task, as the focus is on gender, we explored using only categories in the HurtLex lexicon that relate to gender and misogyny. This is denoted in the Tables as "5 cat." which stands for "5 categories". This approach was

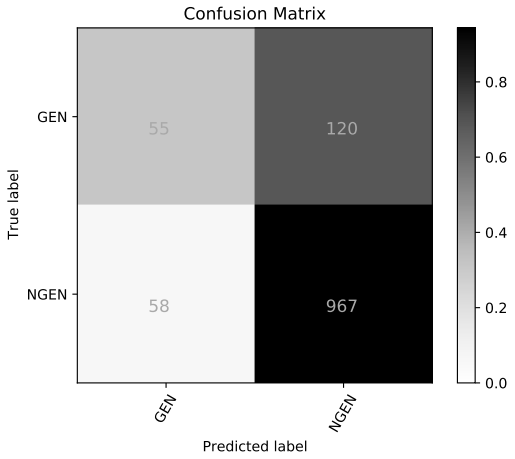

Figure 4: Confusion matrix for our best run in Sub-task EN-B (EN+HIN embeddings, retrofitted on 5 categories of EN+HIN HurtLex).

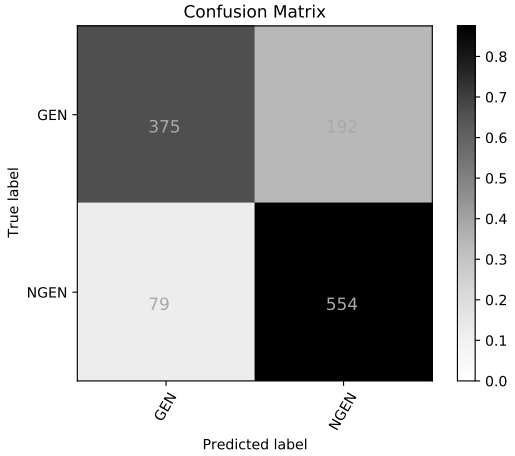

Figure 5: Confusion matrix for our best run in Sub-task HIN-B (EN+HIN embeddings, retrofitted on 5 categories of EN+HIN HurtLex).

inspired by Pamungkas et al. (2018b), who also explored the application of a selection of HurtLex categories to automatic misogyny identification.
Specifically, the categories we selected from HurtLex are the following, as in the aforementioned studies (Pamungkas et al., 2018a; Pamungkas et al., 2018b):

- ASF: female genitalia
- ASM: male genitalia
- DDF: physical disabilities and diversity
- DDP: cognitive disabilities and diversity
- PR: words related to prostitution

From the results, we observe that retrofitting using these categories in HurtLex leads to the best performance for English and Hindi when both English and Hindi parts of HurtLex are used as the lexicons for retrofitting, but not for Bengali (possibly due to the smaller coverage of the resource for this language).

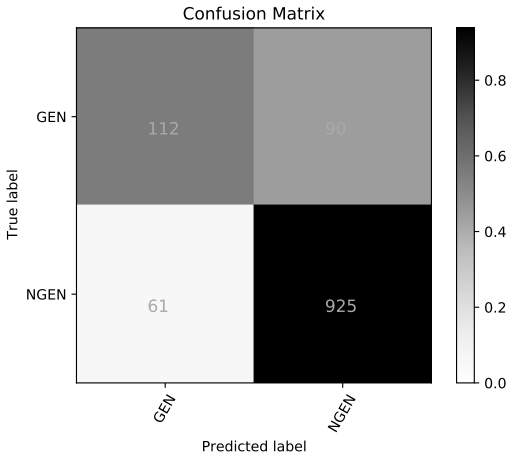

Figure 6: Confusion matrix for our best run in Sub-task BEN-B (EN+BEN embeddings, no retrofitting).

Looking more closely into the predictions of our best runs for this sub-task, the confusion matrices depicted in Figures 4–6 show a similar situation. In all three languages, our classifiers are quite conservative with respect to the *gendered* class, with roughly twice as many (depending on the language) GEN→NGEN misclassifications than NGEN→GEN.

5. Conclusion

In this report, we presented our systems submitted to the TRAC-2 shared task on aggression identification. We participated to both sub-tasks (aggression and gendered aggression) in the three languages proposed by the organizers.
The main novelty of our proposed approach is the use of a multilingual abusive lexicon, and the implementation of a retrofitting technique on pre-trained embeddings based on such lexicon. Although our methods yielded mixed results in the general aggression identification task (Sub-task A) compared to the method without retrofitting, we show that our approach is indeed beneficial when focused on a more narrow scope, namely misogynistic aggression identification.
Despite differences in coverage, the resources used by our models are available for all the languages proposed in this shared task, as well as many more languages. We even found that the different languages actually inform each other, especially in presence of code-switched data.
Future work includes exploring the effect of altering the retrofitting method and its parameters for its application to abusive lexicons as well as experimenting with different data and models. Given the success of using the categorized lexicon HurtLex for some of the subtasks, we also plan to explore the direct coding of lexical-level features based on the lexicon, in a complementary approach to retrofitting.

Acknowledgements

The work of V. Basile and V. Patti is partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01).

6. Bibliographical References

Alawad, M., Hasan, S. S., Christian, J. B., and Tourassi, G. (2018). Retrofitting word embeddings with the umls metathesaurus for clinical information extraction. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2838–2846. IEEE.

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.

Bassignana, E., Basile, V., and Patti, V. (2018). Hurtlex: A multilingual lexicon of words to hurt. In Elena Cabrio, et al., editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*.

Baziotis, C., Pelekis, N., and Doulkeridis, C. (2017). Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August. Association for Computational Linguistics.

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.

Caselli, T., Basile, V., Mitrović, J., Kartoziya, I., and Granitzer, M. (2020). I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language. In *Proceedings of LREC*.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of Eleventh International AAAI conference on Web and Social Media*.

De Mauro, T. (2016). Le parole per ferire. *Internazionale*. 27 settembre 2016.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., and Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado, May–June. Association for Computational Linguistics.

Fersini, E., Nozza, D., and Rosso, P. (2018). Overview of the EVALITA 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR-WS.org.

Gambäck, B. and Sikdar, U. K. (2017). Using Convolutional Neural Networks to Classify Hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Koufakou, A. and Scott, J. (2019). Exploring the use of lexicons to aid deep learning towards the detection of abusive language. In *Proceedings of the 2019 Workshop on Widening NLP, collocated with ACL*, pages 129–131.

Koufakou, A. and Scott, J. (2020). Lexicon-enhancement of embedding-based approaches towards the detection of abusive language. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*.

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

Kumar, R., Ojha, A. K., Malmasi, S., 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)*, France, May. European Language Resources Association (ELRA).

Mishra, S. and Mishra, S. (2019). 3idiots at hasoc 2019: Fine-tuning transformer neural networks for hate speech identification in indo-european languages. In *FIRE*.

Mishra, P., Yannakoudakis, H., and Shutova, E. (2018). Neural character-based composition models for abuse detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 1–10.

Mishra, P., Yannakoudakis, H., and Shutova, E. (2019). Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024*.

Mrkšić, N., Vulić, I., Séaghdha, D. Ó., Leviant, I., Reichart, R., Gašić, M., Korhonen, A., and Young, S. (2017). Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the association for Computational Linguistics*, 5:309–324.

Pamungkas, E. W., Cignarella, A. T., Basile, V., and Patti, V. (2018a). Automatic identification of misogyny in english and italian tweets at evalita 2018 with a multilingual hate lexicon. In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*. CEUR-WS.org.

Pamungkas, E. W., Cignarella, A. T., Basile, V., and Patti, V. (2018b). *14-ExLab@UniTo* for AMI at ibereval2018: Exploiting lexical knowledge for detecting misogyny in english and spanish tweets. 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 234–241. CEUR-WS.org.

Pavlopoulos, J., Malakasiotis, P., and Androutsopoulos, I. (2017). Deep learning for user comment moderation. In *Proceedings of the First Workshop on Abusive Language (ALW1)*, pages 25–35, Vancouver, BC, Canada, August. Association for Computational Linguistics.

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.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Langauge Online (ALW1)*.

Wiegand, M., Ruppenhofer, J., Schmidt, A., and Greenberg, C. (2018). 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*, June.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval)*.

Zhang, Z., Robinson, D., and Tepper, J. (2018). Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *Lecture Notes in Computer Science*. Springer Verlag.