Aggression Identification in Social Media: a Transfer Learning Based Approach

Faneva Ramiandrisoa, Josiane Mothe

Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pages 26–31
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Aggression Identification in Social Media: a Transfer Learning Based Approach

Faneva Ramiandrisoa[1,2], **Josiane Mothe**[1]
[1]IRIT, Université de Toulouse, France
[2] Université d'Antananarivo
{faneva.ramiandrisoa, josiane.mothe}@irit.fr

Abstract

The way people communicate have changed in many ways with the outbreak of social media. One of the aspects of social media is the ability for their information producers to hide, fully or partially, their identity during a discussion; leading to cyber-aggression and interpersonal aggression. Automatically monitoring user-generated content in order to help moderating it is thus a very hot topic. In this paper, we propose to use the transformer based language model BERT (*Bidirectional Encoder Representation from Transformer*) (Devlin et al., 2019) to identify aggressive content. Our model is also used to predict the level of aggressiveness. The evaluation part of this paper is based on the dataset provided by the TRAC shared task (Kumar et al., 2018a). When compared to the other participants of this shared task, our model achieved the third best performance according to the weighted F1 measure on both Facebook and Twitter collections.

Keywords: Information systems, Information retrieval, Social media, Cyber-agression, TRAC Trolling, Aggression and Cyberbulling

1. Introduction

Over the years, social media has become one of the key ways people communicate and share opinions (Pelicon et al., 2019). These platforms such as Twitter or WhatsApp, have changed the way people communicate (Décieux et al., 2019). Indeed, the ability to fully or partially hide their identity leads people to publish things that they probably would never say to someone face to face (Pelicon et al., 2019). Several studies have observed the proliferation of abusive language and increase of aggressive and potentially harmful contents on social media (Zhu et al., 2019). Although most of the forms of abusive language are not criminal, they can lead to a deterioration of public discourse and opinions, which can in turn generate a more radicalized society (Pelicon et al., 2019).

Some studies focus on the automatic detection of abusive language as a first step. Different types of abusive content detection have been defined and studied such as hate speech (Warner and Hirschberg, 2012), cyberbulling (Dadvar et al., 2013), aggression (Kumar et al., 2018a).

In parallel, different evaluation forums propose shared tasks to foster the development of systems to help abusive language detection. Among them, we can cite: TRAC (Kumar et al., 2018a), GermEval (Struß et al., 2019) and SemEval-2019 Task 6 (Zampieri et al., 2019).

The objective of SemEval-2019 Task 6 and GermEval is to detect offensive language in tweets, respectively in English and German. To solve these shared tasks, participants heavily rely on deep learning approaches as well as transfer learning using the transformer based language model BERT (Devlin et al., 2019); with good success (Struß et al., 2019; Zampieri et al., 2019).

As for the TRAC shared task, the objective is to detect aggression in Facebook and Twitter posts and comments. Deep learning approaches are also widely used in this shared task and achieved the best performance (Kumar et al., 2018a). However, no participant used transfer learning based on BERT model while this model achieved good performance on offensive language detection and on a wide range of Natural Language Processing (NLP) tasks. Indeed, BERT model broke several records for how well models can handle language-based tasks. Moreover, to the best of our knowledge, the BERT model has never been used on the TRAC dataset in the literature. This statement motivated us to conduce this work and evaluate a BERT model approach on the TRAC task.

In this paper, we proposed a model that uses transfer learning technique based on the on BERT model to address the problem of aggression identification on Facebook and Twitter content (more details in Section 3.). We evaluate the model on the dataset provided by the TRAC shared task. We also compare our model with the ones of the participants to the shared task. For this, we adopted the same rules as during the shared task (Kumar et al., 2018a).

The rest of this paper is organized as follows: Section 2. presents related work in the area of offensive detection and different existing shared tasks in this domain; Section 3. describes the methodology we propose for aggression detection; Section 4. describes in detail the TRAC dataset and evaluation measures we use for evaluation; Section 5. presents the results and discusses them; finally, Section 6. concludes this paper and presents some future work.

2. Related Work

Recent overviews of related work on the detection of abusive language are presented in (Schmidt and Wiegand, 2017) and (Mishra et al., 2019). (Schmidt and Wiegand, 2017) presents a survey on hate speech detection using Natural Language Processing (NLP). The authors report that supervised learning approaches are predominantly used for this later task. Support vector machines (SVM) and recurrent neural networks are the most widespread. The authors also report that features are widely used for hate speech detection, such as simple surface features (e.g. bag of words, n-grams, etc.), word generalization (e.g. word em-

bedding, etc.), knowledge-based features (e.g. ontology, etc.), ... (Mishra et al., 2019) report a survey of automated abuse detection methods as well as a detailed overview of datasets that are annotated for abuse. The authors notice that many researchers have exclusively relied on text based features for abuse detection while the recent state of the art approaches rely on word-level Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN).
Within shared tasks on abusive language detection, participants heavily use deep learning techniques that achieved good performances. This is the case for GermEval (Struß et al., 2019), SemEval-2019 Task 6 (Zampieri et al., 2019) and TRAC (Kumar et al., 2018a).
GermEval (Struß et al., 2019) is a shared task that focuses on the detection of offensive language on German tweets. During this shared task, the best performing system on the various sub-tasks of the challenge uses the transformer based language model BERT (Devlin et al., 2019), which convinced us to consider BERT in our work as well.
SemEval-2019 Task 6 (Zampieri et al., 2019) is a shared task that focused on identification and classification of offensive language in social media, more precisely on English tweets. During the SemEval-2019 Task 6, the transformer based language model BERT (Devlin et al., 2019) was also widely used and achieved top performances, and even in the case it did not achieve the best performance, overall it performed well.
Finally, TRAC (Kumar et al., 2018a) is a shared task that focuses on aggression identification considering both English and Hindi languages. The objective is to classify texts into three classes: **Non-Aggressive (NAG)**, **Covertly Aggressive (CAG)**, and **Overtly Aggressive (OAG)**. Facebook posts and comments are provided for training and validation, while, for testing, two different sets, one from Facebook and one from Twitter, were provided. The best performance during the shared task was achieved with deep learning approaches whether on Facebook test set or Twitter test set (Kumar et al., 2018a). During this shared task, apart from deep learning approaches, such as CNN + LSTM architecture (Ramiandrisoa, 2020), participants considered classical machine learning methods (e.g. Random Forests) based on features as in (Ramiandrisoa and Mothe, 2018; Arroyo-Fernández et al., 2018; Risch and Krestel, 2018). However, no team used BERT model for aggression detection and according to our knowledge, it was also never used on the TRAC dataset. In this paper, we propose to use this transformer based language model for aggression detection on TRAC dataset since it achieved good results on other shared tasks, specifically on abusive language detection and it has also advanced the state of the art for eleven Natural Language Processing (NLP) tasks (Devlin et al., 2019).
In the next Section, we describe the methodology we adopted as well as the TRAC dataset we used.

3. Methodology

According to related work where the transformer-based language model BERT (Devlin et al., 2019) achieves the top performance on offensive language and hate speech detection, we decided to adopt it for the aggression detection problem. For best understanding of our model, in this section, we provide first a short description of BERT model before describing our model.

3.1. BERT details

BERT or Bidirectional Encoder Representations from Transformers is a new method of pre-training language representations which obtains state-of-the-art results on a wide range of NLP tasks. Using BERT has two stages : pre-training and fine-tuning.
During pre-training, a deep bidirectional representation is trained on unlabeled data by jointly conditioning on both left and right context in all layers. Pre-training is fairly expensive but fortunately a number of pre-trained models were trained at Google on the same corpus data composed of BooksCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2,500M words). These pre-trained BERT models are publicly available on github[1], so most of NLP researchers do not need to pre-train their own model from scratch. Two model sizes of pre-trained BERT model are released which are BERT_{Base} and BERT_{Large}. The BERT_{Base} model contains 12 layers of size 768, 12 self-attention heads and 110M parameters, while the BERT_{Large} model contains 24 layers of size 1024, 16 self-attention heads and 340M parameters.
Compared to pre-training, fine-tuning is relatively inexpensive. Fine-tuning BERT model consists of consists of adding one additional output layer to the pre-trained model, then train it on labeled data from the downstream task to create a new model. With this method, there is no need of task-specific architecture modifications. In other words, the fine-tuning is a transfer learning of pre-trained BERT model. More details on BERT can be found in (Devlin et al., 2019).

3.2. Model details

In this work, we fine-tuned the BERT_{Large} model since it gives better performance than the BERT_{Base} model in a variety of tasks (Devlin et al., 2019).
As BERT is a pre-trained model, it requires a specific format for the input data. As input, it requires three sequences (of the same length): sequence of token IDs, sequence of mask IDs and sequence of segment IDs. In others words, we should convert all texts in our corpus into triplets of sequences.
In the following, we detail how to transform a given text into a triplet of sequences as illustrated in Figure 1:

1) Break text into sequence of tokens by using the BERT tokenizer. A maximum sequence length is fixed in order to have the same length for all sequences in the corpus. So longer sequences are truncated to the size of maximum sequence length minus two and shorter sequences are padded. In this paper, we set the maximum sequence length to 40 tokens because the maximum length of our preprocessed text is equal to 32 in the training set and 31 in the validation set. In other words, we do not cut any texts during training.

[1] `https://github.com/google-research/bert`, accessed on February, 04^{th} 2020

2) Add the token "[CLS]" at the beginning of the sequence of tokens and the token "[SEP]" at the end.

3) Convert each token in the sequence of tokens into ID by using also the BERT tokenizer. The result of the conversion is the sequence of token IDs.

4) Pad with 0 the sequence of token IDs with length less than the maximum sequence length fixed in step 1).

5) Build the sequence of mask IDs which is used to indicate which elements in the sequence of token IDs are real tokens and which are padding elements. The mask has 1 for real tokens and 0 for padding tokens. Figure 1 illustrates this process on an example.

6) Build the sequence of segment IDs which contains only 0 as elements because we classify a text. See Figure 1 for an illustrative example.

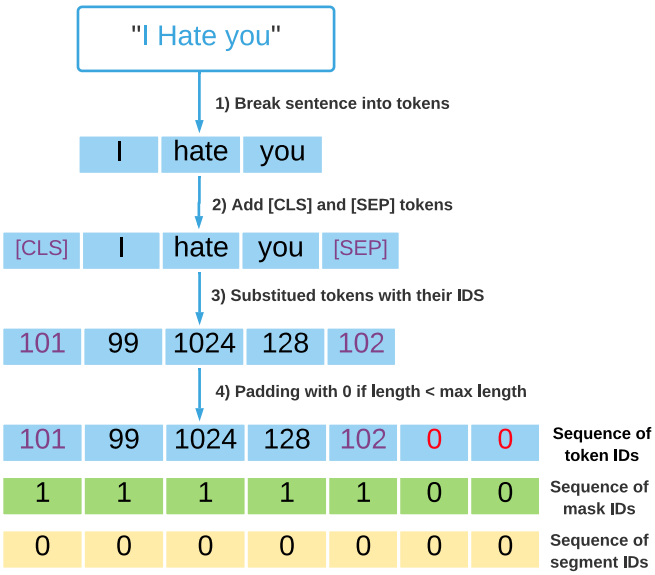

Figure 1: The sequence of token IDs, sequence of mask IDs and sequence of segment IDs from a text. In that illustrative example, the maximum sequence length is fixed to 7.

With regard to the output, a linear layer composed of three nodes is added. This is because there are three classes in the TRAC shared task dataset.

During training, more precisely fine-tuning, we used a batch size of 8, the Adam optimizer with a learning rate of 2e-5 and a number of epochs of 3 as parameters. For the implementation, we used the library pytorch-pretrained-bert[2]. Training was carried out on a Nvidia Geforce GTX 1080TI GPU and took about 39 minutes in total.

In the next sections, we report the evaluation framework and then the results of our fine-tuned BERT model.

[2] https://github.com/shehzaadzd/pytorch-pretrained-BERT, accessed on February, 04^{th} 2020

4. Evaluation framework

In this section, we detail the dataset we used in this paper to evaluate our model as well as how we preprocessed it for text cleaning; we also present the evaluation measure.

4.1. Data

4.1.1. Data Description

The dataset used in this work is the dataset provided for the TRAC shared task (Kumar et al., 2018a) which is a subset of dataset describes in (Kumar et al., 2018b). It consists in English and Hindi randomly sampled Facebook and Twitter comments. In this study, we focused on the English part only, which is detailed in Table 1.

In the dataset, comments are annotated with 3 levels of aggression:

- Non-Aggressive (NAG) : this label is used for data that is generally not intended to be aggressive and mostly used while wishing or supporting individuals or groups.
- Covertly Aggressive (CAG) : this label is used for data that contains hidden aggression and sarcastic negative emotions such as using metaphorical words to attack an individual or a group.
- Overtly Aggressive (OAG) : this label is used for data that contains open and direct aggression such as a direct verbal attack pointed towards any group or individual.

The dataset in the shared task was divided in three sets: training, validation and test. The training and validation sets are used to build models and are only composed of comments from Facebook. Considering English only, the training set is composed of 11,999 comments while the validation set is composed of 3,001 comments.

For the test set, two collections were given: the first is composed of 916 comments crawled from Facebook and the second is composed of 1,257 comments crawled from Twitter. The collection built from Twitter is what the organizers named the *surprise collection* and the idea behind this collection is to test the power of generalization of the developed model. Indeed, the model is trained on Facebook content but tested on both Facebook and Twitter contents.

Number of	Train	Validation	Test	
			Facebook	Twitter
texts	11,999	3,001	916	1,257
OAG	2,708	711	144	361
CAG	4,240	1,057	142	413
NAG	5,051	1,233	630	483

Table 1: Distribution of training, validation and testing data on English TRAC 2018 data collection.

4.1.2. Preprocessing

In this section, we describe the preprocessing steps we applied on Facebook and Twitter comments in order to clean

them before using it to learn the model when training and to evaluate it when testing.

Emoticon substitution : we used the online emoji project on github `https://github.com/carpedm20/emoji` [3] to map the emoticon unicode to substituted phrase. Then we treat the substituted phrase into regular English phrase.

HashTag segmentation : HashTags are commonly used in social media like Twitter, Instagram, Facebook,... In order to detect whether an HashTag contains abusive or offensive words, we used an open source word segmentation available on github `https://github.com/grantjenks/python-wordsegment` [4]. One example would be "#asshole" segmented as "asshole" which is offensive in this case.

Misc. : we converted all texts into lowercase. Also all "URL" is substituted by "http". And Finally, we removed all digit, punctuation, email and non UTF-8 word.

4.2. Evaluation measure

The evaluation metric used in this paper is the same measure as used in the TRAC shared task which is the weighted F1. The weighted F1 is equal to the average of the F1 (given by equation 1) of each class label; it is an weighted average, weighted by the number of instances for each class label.

$$F1 = 2\frac{R * P}{R + P} \quad (1)$$

where $P = \frac{TP}{TP+FP}$ is the precision, $R = \frac{TP}{TP+FN}$ is the recall, TP denotes the true positives, FP the false positives, and FN the false negatives.

5. Results

Table 2 (resp. Table 3) summarizes our results on Facebook (resp. on Twitter) test set. In each table, we can see the three best results from participants in the TRAC workshop and our model which is the fine-tuned of the large pre-trained BERT model.

On Facebook test set, the fine-tuned BERT model (our model) achieves a weighted F1 of 0.627, clearly exceeding the baseline and ranks our model 3rd when compared to the participants of the TRAC shared task.

Systems	Weighted F1
Saroyehun (Aroyehun and Gelbukh, 2018)	**0.642**
EBSI-LIA-UNAM (Arroyo-Fernández et al., 2018)	0.632
BERT-based model (ours)	0.627
DA-LD-Hildesheim (Modha et al., 2018)	0.618

Table 2: Results for the English task on Facebook test set. Bold value is the best performance.

On Twitter test set, the fine-tuned BERT model (our model) achieves a weighted F1 of 0.595, clearly exceeding the baseline and ranks also our model 3rd when compared to TRAC shared task participants.

[3] accessed on February, 04^{th} 2020

[4] accessed on February, 04^{th} 2020

Systems	Weighted F1
vista.ue (Raiyani et al., 2018)	**0.601**
Julian (Risch and Krestel, 2018)	0.599
BERT-based model (ours)	0.595
saroyehun (Aroyehun and Gelbukh, 2018)	0.592

Table 3: Results for the English task on Twitter test set. Bold value is the best performance.

In view of these results, our model can easily generalize from one social media platform to another one. Indeed, our model is trained on Facebook comments and achieved good performance, the same 3rd rank, when tested on both Facebook and Twitter comments. It is worth noticing that the systems that outperforms ours are not the same on the two collections, showing that there are less stable than ours. The next step is to test our model on other social media content.

5.1. Discussion

Figures 2 and 3 present the confusion matrices of our model on Facebook and Twitter test sets respectively. When analysing the results of our model according to weighted F1 on both test sets, we can see that our model mislabelled several NAG instances with CAG class. In general, our model shows better performance on classes with many training instances compared with classes with less training instances except with CAG class. Our model has some difficulty to identify the CAG class. Indeed, even though the OAG class has the smaller number of instances, the performance on the OAG class is better than on the CAG class which has more instances.

On the Facebook test set, CAG is the class where our model is less performing, with an F1 score of 0.36, followed by OAG class with an F1 score of 0.55 and NAG with 0.71. From the figure 2, we can see that it is hard for our model to distinguish CAG from NAG as it predicts 181 NAG instances as CAG. We can see this also holds between OAG and NAG where our model predicts 74 NAG instances as OAG. This second case may be due to the number of instances in the data set (used to train the model) because we have about 2 times more NAG cases than OAG cases.

On the Twitter test set, the most problematic class to identify was also CAG where our model got an F1 score of 0.38, followed by OAG with an F1 score of 0.66 and NAG with 0.73. Figure 3 shows that not only our model has some difficulty to distinguish CAG from NAG but also has some difficulty to distinguish CAG from OAG.

6. Conclusion and Future Work

This paper details the model we propose to solve aggression detection. It also reports the results we obtained on the TRAC English dataset (Facebook and Twitter based) (Kumar et al., 2018a). For this, we trained a neural network based classifier by fine-tuning the pre-trained BERT$Large$ model.

The evaluation shows that our model is able to detect aggression in social media content and achieves the 3rd best

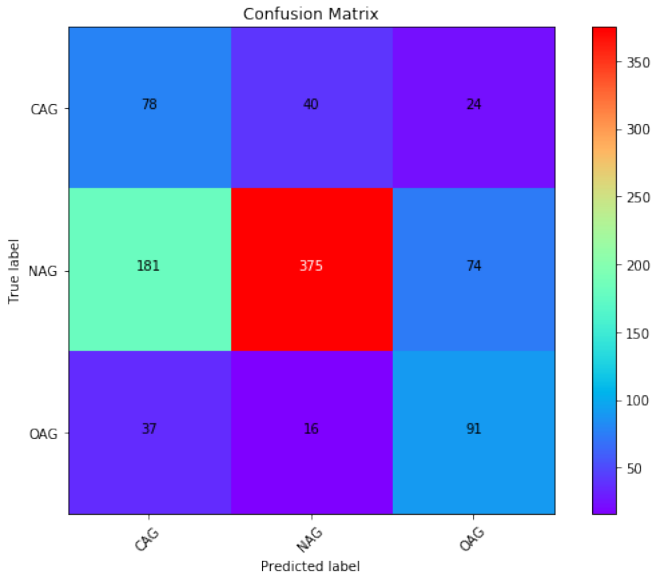

Figure 2: Heatmap of the confusion matrix of our model on Facebook test set.

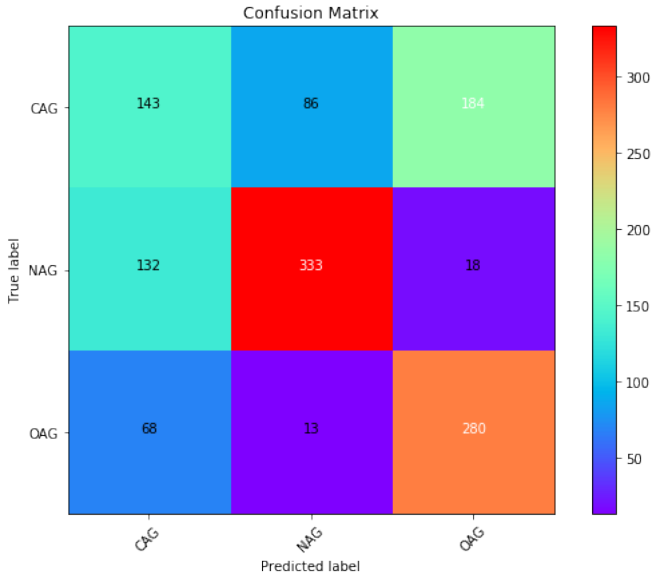

Figure 3: Heatmap of the confusion matrix of our model on Twitter test set.

result both on Facebook and Twitter test sets and this, even if the model is trained on Facebook comments only.

For Future work, we plan to apply our model to the second edition of the TRAC shared task[5]. Also we plan to improve our preporcessing step by enlarging the training set with data augmentation techniques or using external datasets because it has been shown to be effective in (Aroyehun and Gelbukh, 2018). As for information representation, the Information Nutritional Label could be worth investigating as well since it has been shown to be interesting to represent information for various IR tasks (Fuhr et al., 2018; Lespagnol et al., 2019), possibly combined with a key-phrase representation which is semantically richer than word representation (Mothe et al., 2018). We also plan to test our model on related collections, tasks, and sub-tasks in order to evaluate its robustness.

Ethical issue. While TRAC challenge has its proper ethical policies, detecting aggressive content from user's posts raises ethical issues that are beyond the scope of the paper.

[5] https://sites.google.com/view/trac2/home, accessed on February, 04^{th} 2020

7. Bibliographical References

Aroyehun, S. T. and Gelbukh, A. F. (2018). Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, TRAC@COLING 2018, Santa Fe, New Mexico, USA, August 25, 2018*, pages 90–97.

Arroyo-Fernández, I., Forest, D., Torres-Moreno, J., Carrasco-Ruiz, M., Legeleux, T., and Joannette, K. (2018). Cyberbullying detection task: the EBSI-LIA-UNAM system (ELU) at coling'18 TRAC-1. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, TRAC@COLING 2018, Santa Fe, New Mexico, USA, August 25, 2018*, pages 140–149.

Dadvar, M., Trieschnigg, D., Ordelman, R., and de Jong, F. (2013). Improving cyberbullying detection with user context. In *Advances in Information Retrieval - 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings*, pages 693–696.

Décieux, J. P., Heinen, A., and Willems, H. (2019). Social media and its role in friendship-driven interactions among young people: A mixed methods study. *YOUNG*, 27(1):18–31.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Fuhr, N., Giachanou, A., Grefenstette, G., Gurevych, I., Hanselowski, A., Jarvelin, K., Jones, R., Liu, Y., Mothe, J., Nejdl, W., et al. (2018). An information nutritional label for online documents. In *ACM SIGIR Forum*, volume 51, pages 46–66. ACM New York, NY, USA.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018a). Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, TRAC@COLING 2018, Santa Fe, New Mexico, USA, August 25, 2018*, pages 1–11.

Kumar, R., Reganti, A. N., Bhatia, A., and Maheshwari, T. (2018b). Aggression-annotated corpus of hindi-english code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*.

Lespagnol, C., Mothe, J., and Ullah, M. Z. (2019). Information nutritional label and word embedding to esti-

mate information check-worthiness. In *ACM SIGIR conference on research and development in information retrieval*, pages 941–944.

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

Modha, S., Majumder, P., and Mandl, T. (2018). Filtering aggression from the multilingual social media feed. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, TRAC@COLING 2018, Santa Fe, New Mexico, USA, August 25, 2018*, pages 199–207.

Mothe, J., Ramiandrisoa, F., and Rasolomanana, M. (2018). Automatic keyphrase extraction using graph-based methods. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pages 728–730.

Pelicon, A., Martinc, M., and Novak, P. K. (2019). Embeddia at semeval-2019 task 6: Detecting hate with neural network and transfer learning approaches. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 604–610.

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@COLING 2018, Santa Fe, New Mexico, USA, August 25, 2018*, pages 28–41.

Ramiandrisoa, F. and Mothe, J. (2018). IRIT at TRAC 2018. In *Workshop on Trolling, Aggression and Cyberbullying, in International Conference of Computational Linguistics (TRAC@COLING 2018), Santa Fe, New Mexico, USA, 25/08/2018*, pages 19–27, http://www.aclweb.org. Association for Computational Linguistics (ACL).

Ramiandrisoa, F. (2020). Aggression Identification in Posts - two machine learning approaches. In *Workshop on Machine Learning for Trend and Weak Signal Detection in Social Networks and Social Media*.

Risch, J. and Krestel, R. (2018). Aggression identification using deep learning and data augmentation. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, TRAC@COLING 2018, Santa Fe, New Mexico, USA, August 25, 2018*, pages 150–158.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, SocialNLP@EACL 2017, Valencia, Spain, April 3, 2017*, pages 1–10.

Struß, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M., and Klenner, M. (2019). Overview of germeval task 2, 2019 shared task on the identification of offensive language. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019*.

Warner, W. and Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada, June. Association for Computational Linguistics.

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@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 75–86.

Zhu, Y., Kiros, R., Zemel, R. S., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27.

Zhu, J., Tian, Z., and Kübler, S. (2019). UM-IU@LING at SemEval-2019 task 6: Identifying offensive tweets using BERT and SVMs. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 788–795, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.