Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pages 1–5
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Evaluating Aggression Identification in Social Media

Ritesh Kumar[1], Atul Kr. Ojha[2,3], Shervin Malmasi[4], Marcos Zampieri[5]
[1]Dr. Bhimrao Ambedkar University, Agra, [2]Charles University, Faculty of Mathematics and Physics Institute of Formal and Applied Linguistics, Prague & [3]Panlingua Language Processing LLP, New Delhi, [4]Amazon Inc., USA, [5]Rochester Institute of Technology, USA
ritesh78_llh@jnu.ac.in, shashwatup9k@gmail.com, shervin.malmasi@mq.edu.au, marcos.zampieri@rit.edu

Abstract

In this paper, we present the report and findings of the Shared Task on Aggression and Gendered Aggression Identification organised as part of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC - 2) at LREC 2020. The task consisted of two sub-tasks - aggression identification (sub-task A) and gendered aggression identification (sub-task B) - in three languages - Bengali, Hindi and English. For this task, the participants were provided with a dataset of approximately 5,000 instances from YouTube comments in each language. For testing, approximately 1,000 instances were provided in each language for each sub-task. A total of 70 teams registered to participate in the task and 19 teams submitted their test runs. The best system obtained a weighted F-score of approximately 0.80 in sub-task A for all the three languages. While approximately 0.87 in sub-task B for all the three languages.

Keywords: Aggression, Gendered Aggression, English, Hindi, Bengali, TRAC

1. Introduction

In recent years, there have been several studies exploring the computational modelling and automatic detection of abusive content in social media focusing on toxic comments[1], aggression (Kumar et al., 2018), cyberbullying (Xu et al., 2012; Dadvar et al., 2013), hate speech (Davidson et al., 2017), and offensive content (Zampieri et al., 2019a) to name a few. Prior studies have tackled abusive language identification in content from different platforms such as Twitter (Xu et al., 2012; Burnap and Williams, 2015; Davidson et al., 2017; Wiegand et al., 2018), Wikipedia comments[1], and Facebook (Kumar et al., 2018). A number of shared tasks been organized focusing on the automatic detection of offensive language (Struß et al., 2019; Zampieri et al., 2019b; Mandl et al., 2019), hate speech (Basile et al., 2019) and aggression (Kumar et al., 2018). These have motivated the creation of for various languages such as English, German, Hindi, Italian, Spanish, and others.

In this paper, we discuss the results of the second iteration of the TRAC shared task, organized as part of the Workshop on Trolling, Aggression and Cyberbullying at LREC 2020. The task consisted of two sub-tasks - aggression identification and gendered aggression identification on YouTube comments in three languages: Bengali, Hindi and English. To the best of our knowledge, TRAC-2 is the first shared task to include YouTube comments as training and testing data and the first shared task to include Bengali data. Both these novel aspects open new avenues for future research.

The remainder of this paper is organized as follows. Section 2. discusses related studies and shared tasks to TRAC-2. Section 3. presents the setup and schedule of TRAC-2 and Section 4. presents the dataset used in the competition. Section 5. presents the approaches used by participants of the competition and Section 6. presents and analyzes the results they obtained. Finally, 7. concludes this paper and presents avenues for future work.

[1]http://bit.ly/2FhLMVz

2. Related Work

Automatically identifying the various forms of abusive language online has been studied from different angles. Examples include trolling (Cambria et al., 2010; Kumar et al., 2014; Mojica, 2016; Mihaylov et al., 2015), flaming / insults (Sax, 2016; Nitin et al., 2012), radicalization (Agarwal and Sureka, 2015; Agarwal and Sureka, 2017), racism (Greevy and Smeaton, 2004; Greevy, 2004), misogyny ((Menczer et al., 2015; Frenda et al., 2019; Hewitt et al., 2016; Fersini et al., 2018; Anzovino et al., 2018; Sharifirad and Matwin, 2019)), online aggression (Kumar et al., 2018), cyberbullying (Xu et al., 2012; Dadvar et al., 2013), hate speech (Kwok and Wang, 2013; Djuric et al., 2015; Burnap and Williams, 2015; Davidson et al., 2017; Malmasi and Zampieri, 2017; Malmasi and Zampieri, 2018), and offensive language (Wiegand et al., 2018; Zampieri et al., 2019b). The terms used in the literature have overlapping properties as discussed in Waseem et al. (2017) and Zampieri et al. (2019a). The most important differences concern their target (e.g. hate speech is typically targeted at groups whereas cyberbulling targets individuals), which is represented in TRAC-2 Task B, and types (e.g. veiled or direct abuse), represented in TRAC-2 Task A.

Most related studies focus on English, but significant amount of work has been carried out for other languages too. This includes languages such as Arabic (Mubarak et al., 2020), German (Struß et al., 2019), Greek (Pitenis et al., 2020), Hindi (Mandl et al., 2019), and Spanish (Basile et al., 2019).

TRAC - 2 is the second iteration of the TRAC shared task on Aggression Identification (Kumar et al., 2018) hosted at the TRAC workshop at COLING 2018. The first edition of TRAC included English and Hindi data from Facebook and Twitter. It consisted of a three-way classification task with posts labelled as *overtly aggressive*, *covertly aggressive*, and *non-aggressive*. TRAC received 30 submissions and the results obtained by participants suggested that neural network-based systems and machine learning classifiers

Language	Train Sub-task A				Train Sub-task B			Test Set				
	TOTAL	NAG	CAG	OAG	TOTAL	NGEN	GEN	NAG	CAG	OAG	NGEN	GEN
Bengali	**4,783**	2,600	1,116	1,067	**4,783**	3,880	903	789	169	242	1005	195
English	**5,329**	4,211	570	548	**5,329**	4,947	382	690	224	286	1023	177
Hindi	**4,981**	2,823	1,040	1,118	**4,981**	4,168	813	316	215	669	700	500

Table 1: Number of instances in each class in the TRAC-2 datasets.

(e.g. SVMs) achieved comparable performance.
Shared tasks similar to TRAC have been organized in recent years. One such example is OffensEval (SemEval-2019 Task 6) (Zampieri et al., 2019b) which focused on offensive language identification. OffensEval featured three sub-tasks: offensive language identification, offensive type identification, and offense target identification building on the annotation model introduced in the OLID dataset (Zampieri et al., 2019a) for English. This multiple sub-task model has been adopted by other shared tasks such as GermEval for German (Struß et al., 2019), HASOC (Mandl et al., 2019) for English, German, and Hindi, and HatEval (Basile et al., 2019) for English and Spanish.

3. Task Setup and Schedule

Participants enrolled to participate in any combination of tracks and languages. The registered participants were sent the links to the annotated datasets along with a description of the format of the dataset. The participants were allowed to use additional data for training the system, with the condition that the additional dataset should be either publicly available or make available immediately after submission. Use of non-public additional data for training was not allowed. The participants were given around 6 weeks to experiment and develop the system. After the 6 weeks of release of train and development sets, the test set was released and the participants had 7 days to test and upload their system. The complete timeline of the shared task is given in Table 2.

Date	Event
December 30, 2019	Announcement and registration
January 25, 2020	Train and dev set release
March 5, 2020	Test set release
March 12, 2020	System submission
March 11, 2020	Declaration of results
March 31, 2020	System description paper

Table 2: TRAC-2 timeline.

We made use of CodaLab [2] for the evaluation. Each team was allowed to submit up to 3 system runs for evaluation and their best run was included in the final ranking presented in this report.

4. Dataset

The participants of the shared task were provided with a dataset of approximately 5,000 randomly sampled YouTube comments for training and approximately 1,000 comments for development in each of Bnagla, Hindi and English. For the sub-task on aggression identification, it annotated with 3 levels of aggression - Overtly Aggressive (OAG), Covertly Aggressive (CAG) and Non-Aggressive (NAG). For the second sub-task on gender identification, it was marked as gendered (GEN) or non-gendered (NGEN). For test, over 1,000 comments were provided[3]. The statistics of the complete dataset in each language is given in Table 1.

5. Participants and Approaches

A total of 70 participants registered for the shared task, with most of the teams registering to participate in both tracks and all the languages. Out of these, finally a total of 19 teams submitted their systems. All the teams who submitted their system were invited to submit the system description paper, describing the experiments conducted by them. Table 3, lists the participating teams and the language they took part in. Next we give a short description of the approach taken by each team for building their system. More details about the approaches could be found in the paper submitted by the respective teams.

- **abaruah** uses BERT, RoBERTa, DistilRoBERTa, and SVM-based classifiers for English. For Hindi and Bengali, multilingual BERT (M-BERT), XLM-RoBERTa and SVM classifiers were used.
- **AI_ML_NIT_Patna** uses Convolutional Neural Network and Long Short Term Memory with two different input text representations, FastText and One-hot embeddings. Their findings suggest that the LSTM model with FastText embedding performs better than other models for Hindi and Bengali datasets. On the other hand, the CNN model with FastText embedding gives better results for the English dataset.
- **FlorUniTo** uses word-embedding with an LSTM model.
- **Julian** uses multiple fine-tuned BERT models, based on bootstrap aggregating (bagging).
- **IRIT** uses the transformer-based language model BERT (Bidirectional Encoder Representation from Transformer) for two sub-tasks.
- **lastus** uses bidirectional Long Short Term Memory network (bi-LSTM) to build the purported model.
- **Ms8qQxMbnjJMgYcw** uses a single BERT-based system with two outputs for all tasks simultaneously.

[2] https://competitions.codalab.org/

[3] The complete dataset used for the shared task can be downloaded from the shared task website - https://sites.google.com/view/trac2/shared-task

Team	Bengali	English	Hindi	System Description Paper
Julian	✓	✓	✓	(Risch and Krestel, 2020)
abaruah	✓	✓	✓	(Baruah et al., 2020)
sdhanshu	✓	✓	✓	(Mishra et al., 2020)
Ms8qQxMbnjJMgYcw	✓	✓	✓	(Gordeev and Lykova, 2020)
FlorUniTo	✓	✓	✓	(Koufakou et al., 2020)
na14	✓	✓	✓	(Samghabadi et al., 2020)
AI_ML_NIT_Patna	✓	✓	✓	(Kumari and Singh, 2020)
asking28	✓	✓	✓	
Spyder	✓	✓	✓	(Datta et al., 2020)
zhixuan		✓		
lastus		✓		(Altın et al., 2020)
scmhl5		✓		(Liu et al., 2020)
IRIT		✓		(Ramiandrisoa and Mothe, 2020)
UniOr_ExpSys		✓		(Pascucci et al., 2020)
SAJA		✓		(Tawalbeh et al., 2020)
krishanthvs		✓		
bhanuprakash2708			✓	
saikesav564	✓			
debina			✓	
Total	**10**	**16**	**11**	**13**

Table 3: The teams that participated in the TRAC-2 shared task.

- **na14** uses an end-to-end neural model with attention on top of BERT that incorporates a multi-task learning paradigm addressing both sub-tasks simultaneously.
- **SAJA** uses transfer learning technique depending on universal sentence encoder (USE) embedding.
- **scmhl5** exploits the pre-trained Bert model to extract the text of each instance into a 768-dimensional vector of embeddings. Further it trains an ensemble of classifiers on the embedding features.
- **sdhansu** uses fine-tuning of various Transformer models on the different datasets. The utility of task label marginalization, joint label classification, and joint training on multilingual datasets as possible improvements to their models was also investigated. Their analysis suggests that the multilingual joint training approach is the best trade-off between computational efficiency and evaluation performance.
- **Spyder** uses three different models using Tf-Idf, sentiment polarity and machine learning-based classifiers.
- **UniOr_ExpSys** uses linguistic rules, stylistic features and a Sequential Minimal Optimization (SMO) algorithm in building their classifiers.

6. Results

In this section, we present the results of the experiments carried out by different teams during the shared task. In the task, the participants were allowed to use other datasets, in addition to the one provided by the organizers. However, because of the lack of similar alternative datasets, all the groups used only the dataset provided for the task. As we mentioned earlier, for for the final testing of the system, 1000 instances were given to participants in each language for each sub-task.

The teams' result on Bengali, English and Hindi dataset is demonstrated in Table 4. In sub-task A , the best system obtained a weighted F-score of approximately 0.82 for Bengali, 0.80 for English and 0.81 for Hindi. In other words, the best system obtained approximately 0.80 F-score for all the three languages. In sub-task B, the best system obtained a weighted F-score of approximately 0.93 for Bengali, 0.87 for English and and 0.87 for Hindi.

7. Conclusion

In this paper, we have presented the report of the Second Shared Task on Aggression Identification, organized with the TRAC-2 workshop at LREC-2020. The shared task feature two sub-tasks- aggression identification (sub-task A) in which systems were trained to discriminate between posts labeled as *overtly aggressive*, *covertly aggressive*, and *non-aggressive*, and gendered aggression identification (sub-task B) in which systems were trained to discriminate between *gendered* or *non-gendered* posts. Datasets in Bengali, Hindi and English were made available to participants. TRAC-2 received a very good response from the community which underlines the relevance of the task. More than 70 teams were registered and 19 teams submitted their systems. We found that most of the systems were developed using neural networks following the recent success of such approaches in recent related shared tasks (Zampieri et al., 2019b; Basile et al., 2019). The analysis of the performance of the best systems in the two sub-tasks shows that the three-way aggression identification task in sub-task A is still a challenging task for all languages in TRAC-2.

8. Acknowledgements

The dataset used in the shared task is being develop under a project title 'Communal and Misogynistic Aggression' (The ComMA Project), funded by an Unrestricted Research Gift by Facebook Research.

Team	Bengali		English		Hindi	
	Task A	Task B	Task A	Task B	Task A	Task B
Julian	0.821	0.938	0.802	0.851	0.812	0.878
abaruah	0.808	0.925	0.728	0.870	0.794	0.868
sdhanshu	0.780	0.927	0.759	0.857	0.779	0.849
Ms8qQxMbnjJMgYcw	0.771	0.929	0.756	0.871	0.776	0.838
FlorUniTo	0.745	0.868	0.677	0.837	0.726	0.770
na14	0.736	0.920	0.714	0.857	0.718	0.800
AI_ML_NIT_Patna	0.717	0.879	0.660	0.822	0.654	0.736
asking28	0.685	0.815	0.714	0.710	0.700	0.733
Spyder	0.448	-	0.430	-	0.594	-
zhixuan	-	-	0.739	0.856	-	-
lastus	-	-	0.724	0.819	-	-
scmhl5	-	-	0.663	0.851	-	-
IRIT	-	-	0.635	0.820	-	-
UniOr_ExpSys	-	-	0.629	0.673	-	-
SAJA	-	-	0.607	0.856	-	-
krishanthvs	-	-	0.441	0.737	-	-
bhanuprakash2708	-	-	-	-	0.140	0.413
saikesav564	0.468	-	-	-	-	-
debina	-	-	-	-	-	0.412

Table 4: Performance of teams on Bengali, English & Hindi Dataset

9. Bibliographical References

Agarwal, S. and Sureka, A. (2015). Using knn and svm based one-class classifier for detecting online radicalization on twitter. In *International Conference on Distributed Computing and Internet Technology*.

Agarwal, S. and Sureka, A. (2017). Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website.

Altın, L. S. M., Bravo, A., and Saggion, H. (2020). Lastus/taln at trac - 2020 trolling, aggression and cyberbullying. In Ritesh Kumar, et al., editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*.

Anzovino, M., Fersini, E., and Rosso, P. (2018). Automatic identification and classification of misogynistic language on twitter. In Max Silberztein, et al., editors, *Natural Language Processing and Information Systems*.

Baruah, A., Das, K., Barbhuiya, F., and Dey, K. (2020). Aggression identification in english, hindi and bangla text using bert, roberta and svm. In Ritesh Kumar, et al., editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of SemEval*.

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

Cambria, E., Chandra, P., Sharma, A., and Hussain, A. (2010). Do not feel the trolls. In *ISWC, Shanghai*.

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

Datta, A., Si, S., Chakraborty, U., and Naskar, S. K. (2020). Spyder: Aggression detection on multilingual tweets. In Ritesh Kumar, et al., editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of ICWSM*.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., and Bhamidipati, N. (2015). Hate Speech Detection with Comment Embeddings. In *Proceedings of WWW*.

Fersini, E., Nozza, D., and Rosso, P. (2018). Overview of the evalita 2018 task on automatic misogyny identification (AMI). In Tommaso Caselli, et al., editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. (EVALITA 2018)*.

Frenda, S., Ghanem, B., Montes-y Gómez, M., and Rosso, P. (2019). Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5).

Gordeev, D. and Lykova, O. (2020). Bert of all trades, master of some. In Ritesh Kumar, et al., editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*.

Greevy, E. and Smeaton, A. F. (2004). Classifying racist texts using a support vector machine. In *Proceedings of the ACM SIGIR*.

Greevy, E. (2004). *Automatic text categorisation of racist webpages*. Ph.D. thesis, Dublin City University.

Hewitt, S., Tiropanis, T., and Bokhove, C. (2016). The problem of identifying misogynist language on twitter (and other online social spaces). In *Proceedings of the*

8th ACM Conference on Web Science, WebSci '16.

Koufakou, A., Basile, V., and Patti, V. (2020). Florunito@trac-2: Retrofitting word embeddings on an abusive lexicon for aggressive language detection. In Ritesh Kumar, et al., editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*.

Kumar, S., Spezzano, F., and Subrahmanian, V. (2014). Accurately detecting trolls in slashdot zoo via decluttering. In *Proceedings of ASONAM*.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018). Benchmarking Aggression Identification in Social Media. In *Proceedings of TRAC*.

Kumari, K. and Singh, J. P. (2020). Ai_ml_nit_patna @ trac - 2: Deep learning approach for multi-lingual aggression identification. In Ritesh Kumar, et al., editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*.

Kwok, I. and Wang, Y. (2013). Locate the Hate: Detecting Tweets Against Blacks. In *Proceedings of AAAI*.

Liu, H., Burnap, P., Alorainy, W., and Williams, M. (2020). Scmhl5 at trac-2 shared task on aggression identification: Bert based ensemble learning approach. In Ritesh Kumar, et al., editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*.

Malmasi, S. and Zampieri, M. (2017). Detecting Hate Speech in Social Media. In *Proceedings of RANLP*.

Malmasi, S. and Zampieri, M. (2018). Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation (FIRE)*.

Menczer, F., Fulper, R., Ciampaglia, G. L., Ferrara, E., Ahn, Y., Flammini, A., Lewis, B., and Rowe, K. (2015). Misogynistic Language on Twitter and Sexual Violence. In *Proceedings of the ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM)*.

Mihaylov, T., Georgiev, G. D., Ontotext, A., and Nakov, P. (2015). Finding opinion manipulation trolls in news community forums. In *Proceedings of CoNLL*.

Mishra, S., Prasad, S., and Mishra, S. (2020). Multilingual joint fine-tuning of transformer models for identifying trolling, aggression and cyberbullying at trac 2020. In Ritesh Kumar, et al., editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*.

Mojica, L. G. (2016). Modeling trolling in social media conversations.

Mubarak, H., Rashed, A., Darwish, K., Samih, Y., and Abdelali, A. (2020). Arabic Offensive Language on Twitter: Analysis and Experiments. *arXiv preprint arXiv:2004.02192*.

Nitin, Bansal, A., Sharma, S. M., Kumar, K., Aggarwal, A., Goyal, S., Choudhary, K., Chawla, K., Jain, K., and Bhasinar, M. (2012). Classification of flames in computer mediated communications.

Pascucci, A., Manna, R., Masucci, V., and Monti, J. (2020). The role of computational stylometry in identifying (misogynistic) aggression in english social media texts. In Ritesh Kumar, et al., editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*.

Pitenis, Z., Zampieri, M., and Ranasinghe, T. (2020). Offensive Language Identification in Greek. In *Proceedings of LREC*.

Ramiandrisoa, F. and Mothe, J. (2020). Irit at trac 2020. In Ritesh Kumar, et al., editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*.

Risch, J. and Krestel, R. (2020). Bagging bert models for robust aggression identification. In Ritesh Kumar, et al., editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*.

Samghabadi, N. S., Patwa, P., PYKL, S., Mukherjee, P., Das, A., and Solorio, T. (2020). Aggression and misogyny detection using bert: A multi-task approach. In Ritesh Kumar, et al., editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*.

Sax, S. (2016). Flame Wars: Automatic Insult Detection. Technical report, Stanford University.

Sharifirad, S. and Matwin, S. (2019). When a tweet is actually sexist. A more comprehensive classification of different online harassment categories and the challenges in NLP. *CoRR*, abs/1902.10584.

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 KONVENS*.

Tawalbeh, S., Hammad, M., and AL-Smadi, M. (2020). Saja at trac 2020 shared task: Transfer learning for aggressive identification with xgboost. In Ritesh Kumar, et al., editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding Abuse: A Typology of Abusive Language Detection Subtasks. *Proceedings of ALW*.

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

Xu, J.-M., Jun, K.-S., Zhu, X., and Bellmore, A. (2012). Learning from Bullying Traces in Social Media. In *Proceedings of NAACL*.

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

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019b). SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of SemEval*.