Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pages 99–105
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

SAJA at TRAC 2020 Shared Task: Transfer Learning for Aggressive Identification with XGBoost

Saja Khaled Tawalbeh, Mahmoud Hammad, Mohammad AL-Smadi
Jordan University of Science and Technology
Irbid, Jordan
sajatawalbeh91@gmail.com, {m-hammad, masmadi}@just.edu.jo

Abstract

This paper describes the participation of the SAJA team to the TRAC 2020 shared task on aggressive identification in the English text. we have developed a system based on transfer learning technique depending on universal sentence encoder (USE) embedding that will be trained in our developed model using xgboost classifier to identify the aggressive text data from English content. A reference dataset has been provided from TRAC 2020 to evaluate the developed approach. The developed approach achieved in sub-task EN-A 60.75% F1 (weighted) which ranked fourteenth out of sixteen teams and achieved 85.66% F1 (weighted) in sub-task EN-B which ranked six out of fifteen teams.

Keywords: aggression identification, social media, NLP, USE, transfer learning, XGBoost

1. Introduction

In today's time, the advances in the web and the communication technologies is one of the main reasons to increase the impact of the nasty content on social media, blogs, and other websites. Detecting aggressive and insulting content is gained recent attention according to the negative effects on its users. For instance, demeaning comments, incidents of aggression, trolling, cyberbullies, hate speech, insulting, and toxic utterance have negative impact of users. Unfortunately, during the recent years, the percentage of using toxic utterance has been increased. Consequently, led to problems affecting real societies.

In 2018 the first shared task on aggression identification has been announced (Kumar et al., 2018). (Davidson et al., 2017) presented work for aggression classification by performing the logistic regression classifier depending on several hand-crafted features. (Djuric et al., 2015) focused on the embedding that has been learnt from an input text using paragraph2vec (Le and Mikolov, 2014) to train the logistic regression classifier. In 2013, (Kwok and Wang, 2013) developed a Naive Bayes classifier based on unigram features. (Bhattacharya et al., 2020) the second shared task on aggression identification will behold on Trolling, Aggression, and Cyberbullying (TRAC 2020) focusing on three languages as a Multilingual shared task. It aims to classify social media posts into one of three labels (Overtly aggressive 'OAG', Covertly aggressive 'CAG', Non-aggressive 'NAG'). Moreover, to classify social media posts as binary classifications into (gendered 'GEN' or non-gendered 'NGEN').

The major contribution of this paper is to describe our participation of the SAJA team to the TRAC 2020 shared task on aggressive identification and more precisely we participate in English language. We have developed a system based on transfer learning technique depending on universal sentence encoder (USE) embedding that will be trained in our developed model using XGBoost classifier to identify the aggressive text data from English content. Several approaches have been performed to solve the provided task. We mentioned the best-reported results according to the evaluation step. A reference dataset has been provided from TRAC 2020 to evaluate the developed approach. The developed approach achieved in sub-task EN-A 60.75% F1 (weighted) and achieved 85.66% F1 (weighted) in sub-task EN-B.

We discuss the problem statement in section 2. Section 3 contains details about our methodology and the used dataset. In Section 4, we discuss the results and Section 5 concludes our work.

2. Related Work

Micro-blogging is considered as one of the most popular social network applications. In recent years, the rapid of using social media to express the users feeling and share their ideas. On the other hand, the uses of aggressive, hate speech, and offensive language obviously has been increased gradually.

Present comprehensive studies for hate speech detection by (Schmidt and Wiegand, 2017) and (Fortuna and Nunes, 2018), (Davidson et al., 2017) presenting the Hate Speech Detection dataset. Additionally, (Spertus, 1997) consider as the earliest efforts in hate speech detection, had been presented a decision tree-based classifier with 88.2% accuracy. Moreover, Offensive identification for sentences have been tried for several languages behind the English such that, Arabic (Mubarak et al., 2017) and (Al-Hassan and Al-Dossari, 2019), German (Ross et al., 2017), (Fišer et al., 2017), and (Su et al., 2017).

In particular, Zampieri et al. (2019a) OLID dataset presented last year for offensive language detection (Zampieri et al., 2019b). (Mohaouchane et al., 2019) presents several neural networks namely: (i) CNN, (ii) Bi-LSTM, (iii) Bi-LSTM with attention mechanism, (iv) Combined CNN and LSTM on Arabic language. Moreover, the dataset has been

used is Arabic YouTube comments. The best performing model was CNN with 84.05% F1-score. In (Liu et al., 2019) Proposed a fine-tuned technique for the Bidirectional Encoder Representation from Transformer (BERT) to solve the shared task of identifying and categorizing offensive language in social media at SemEval 2019. Several features were used such as word unigrams and word2vec. (Pelicon et al., 2019) Adds LSTM neural network architecture to perform fine-tuned a BERT. Several automatically and manually crafted features were used namely: word embeddings, TFIDF, POS sequences, BOW, the length of the longest punctuation sequence, and the sentiment of the tweets. (Mahata et al., 2019) Proposed an ensemble technique consist of CNN, Bidirectional LSTM with attention, and Bidirectional LSTM + Bidirectional GRU to tackle the shared SemEval 2019 - Task 6 Identifying and Categorizing Offensive Language. The train data used to obtain a set of heuristics as features. (Han et al., 2019) Presented two approaches namely: bidirectional with GRU and probabilistic model modified sentence offensiveness calculation (MSOC) for the same Task Identifying and Categorizing Offensive Language. Word2vec embedding used as a feature. (Swamy et al., 2019) Introduced an ensemble approach consist of L1-regularised Logistic Regression, L2-regularised Logistic Regression, Linear SVC, and LSTM neural network. Several features were used, for instance, GloVe embedding, word/character n-grams by TF-IDF, POS tags, sentiment Score and count features for URLs, mentions, hashtags, punctuation marks.

In 2018 shared task TRAC 1 has been released, (Ramiandrisoa and Mothe, 2018) have been developed an approach to detect aggressive language for English language. Three approaches have been developed based on machine learning and deep learning models. Several features have been used (i.e Part-Of-Speech, emoticonssentiment frequency and logistic regression built with document vectorization using Doc2vec). (Samghabadi et al., 2018) discussing the lexical and semantic features for English and Hindi languages. (Roy et al., 2018) presented an ensemble solution depending on CNN and SVM for English and Hindi languages. Word embedding, n-grams, and Tf-Idf vectors have been discussed. (Nikhil et al., 2018) demonstrate LSTM approach with an attention unit according to the remarkable results for this approach in NLP tasks. It performs for English and Hindi language as well. Moreover, (Aroyehun and Gelbukh, 2018) presents an investigation between deep learning and traditional machine learning (i.e NB and SVM) to achieve the best efficient model. The remarkable point in this paper to improve the overall performance, the augmented data and pseudo labeled method have been used during the training step.

3. Data and Methodology

Shared task on Aggression Identification focused on the English language which provided to identifying the aggressive language thought the social media content.

3.1. Task Description

The shared task TRAC 2020 (Ritesh Kumar and Zampieri, 2020; Bhattacharya et al., 2020), is a multilingual task, which provides two subtasks namely: A- aggression identification shared task, it represents a classification task to classify a given text into three classes between (1) Overtly Aggressive where it represents the human behavior meant to hurt a community through the verbal, physical and psychological attitude. (2) Covertly Aggressive where it represents the hidden aggressive attack consist of the negative ironic emotions and (3) Non-aggressive. Table 1 represents the aggressive type cases. B- misogynistic aggression identification shared task, it represents a binary classification that aims to classify a given text to gendered or non-gendered.

Type	Cases
Overtly Aggressive (OAG)	verbal attack directly pointed towards any group or individuals, abusive words or comparing in a derogatory manner, supporting false attack
Covertly Aggressive (CAG)	foucus on figurative words aims to attack(individual, nation, religion), Praising someone by criticizing group irrespective of being right or wrong.
Non Aggressive (NAG)	In this case, the absence of the intention to be aggressive.

Table 1: The classes of the Aggressive including their cases for sub-task EN-A

3.2. Dataset

This shared task represents a multilingual dataset Bhattacharya et al. (2020) which contains three languages namely: English, Bangla, and Hindi. In this paper, we participate in the English language for both subtasks A and B. The shared task provides three files train, validation, and test file which consists of 5000 annotated rows from social media that have been represented for both subtasks. Tables 2 and 3 provide more details about the distribution of the provided dataset.
Table 4 represents examples of the provided dataset for both subtasks.

3.3. Data Pre-processing

The pre-processing step on a text is crucial processes, especially social network datasets such that, Facebook and Twitter where posts and tweets are noisy and contain a lot of slang language. In order to have a clean version of the provided dataset to remove the unnecessary noise, for instance, special character, punctuation marks (*,@#-(—),

Dataset File	Total	Count of Labels sub-task EN-A
Train Set	4263	OAG= 435 CAG= 453 NAG= 3375
Validation Set	1066	OAG= 113 CAG= 117 NAG= 836
Test Set	1200	-

Table 2: Represents the distribution of the provided dataset for the English language for suntask A

Dataset File	Total	Count of Labels sub-task EN-B
Train Set	4263	NGEN= 3954 GEN= 309
Validation Set	1066	NGEN= 993 GEN= 73
Test Set	1200	-

Table 3: Represents the distribution of the provided dataset for the English language for sub-task EN-B

URLs, and user mentions. Whereas, pre-processing step is required to improve the analysis process applied to the raw tweets. We have been done various pre-processing to achieve a clean version of the provided dataset, such that, each tweet was normalized. and then tokenized. The normalization is a necessary process since some words are written on shortcut format (i.e. dont returned to (do not)). Finally, numbers and non-English characters were also removed. The following are examples of pre-processing step have been shown in table 5 for the provided dataset.

3.4. Embeddings

Recently, word embeddings widely used in NLP applications and their research, where word embedding aims to obtain the vector representation of the input of textual data to input numeric for deep neural networks. Word embeddings tend to capture the semantic features for each word and the linguistic relationship among them, whereas these embeddings help to improve system performance in several NLP domains (e.g text mining). Since 2003 (Bengio et al., 2003) has been started to generate word embedding using neural probabilistic language model, then Word2Vec by (Mikolov et al., 2013), Glove (Pennington et al., 2014), AraVec (Soliman et al., 2017) and the recent model ElMo Embedding by (Peters et al., 2018), BERT contextual embedding (Devlin et al., 2018), and Universal Sentence Encoder USE (Cer et al., 2018). The distributional linguistic hypothesis it's the main intuition of word embedding idea, whereas each model has its own way to capture the semantic meaning or the idea of how they trained. Consequently, each model can capture different semantic attributes com-

ID	Original Text	Label sub-task EN-A	Label sub-task EN-B
C68.872	Nice video..	NAG	NGEN
C10.689	She is a traitor of India	OAG	NGEN
C32.128	”Wrong message for youth. Fight, dont be a coward”	CAG	NGEN
C65.70	Hot	NAG	GEN

Table 4: Examples that represents the dataset for both subtasks

ID	Original Text	Processed Text	Label sub-task EN-A	Label sub-task EN-B
C68.872	Nice video..	nice video	NAG	NGEN
C10.689	She is a traitor of India	she is a traitor of india	OAG	NGEN
C32.128	”Wrong message for youth. Fight, dont be a coward”	wrong message for youth fight do not be a coward	CAG	NGEN
C65.70	Hot	hot	NAG	GEN

Table 5: Data pre-processing performed on the available dataset for both subtasks

pared to other models. In this research, we depend on pre-trained sentence USE embedding to trained the developed model. It is a language representation model and sentence embedding provided by Google which aims to extract the sentence embeddings from the provided dataset. Moreover, it will become one of the state of art model for most of NLP research.

3.5. Proposed Model-(XGB-USE)

The transformer and contextual embedding added much progress in the NLP research area. In addition, it outperforms the deep learning approaches according to the promising results achieved. The transformer considered an encoder-decoder architecture applied to attention mechanisms tasks. More particularly, Google has been released Universal Sentence Encoder (USE) Cer et al. (2018) which aims to map an input sentence to vector representations, this kind of representation aims to capture

the meaning of the sentence. Moreover, Google has been released a pre-trained USE embedding using TensorFlow Hub Module [1] to extract the embedding directly and find the semantic similarities for the provided sentences.

The proposed model based on transfer learning architecture that has used in common especially in image classification and computer vision (Litjens et al., 2017). Moreover, as we mentioned earlier, the applied transformers show significant results compared to deep learning approaches. For instance, USE developers created several versions of the pre-trained models such as multilingual USE to represent the semantic relationships among text as well as it could be applied as an independent classifier in different NLP domains (i.e. aggressive identification). Moreover, the extracted embedding dimension for USE is 512. In this research, we used USE2 pre-trained model to extract the sentence embedding based on transfer learning architecture to tackle the shared task problem. The XGboost, distributed gradient boosting library (XGB) classifier (Chen and Guestrin, 2016) have been built to be highly efficient. The XGB has been used as a text transfer learning model powered by the USE embedding whereas XGB considered as a powerful classifier compared to other machine learning classifiers as well as compared to deep learning. It becomes a popular method to solve NLP tasks. The reason why XGB has been used as follows: a) XGB considered as a regularized boosting technique prepared to prevent overfitting, b) it has a structure to handle the missing values, c) it is fast compared to others gradient boosting.

As we mentioned above for the sake of this research, the XGB classifier approach has been developed based on transfer learning with Universal Sentence Encoder (XGB-USE). This developed approach performed to solve the aggressive identification for the English language. USE embedding has been extracted from the pre-trained model with 512 dimensions for each input sentence before they prepared to train step using XGB. Table 6 provides more details about the value of each parameter have been used during the training step, which represents the best parameters are used. The XGB-USE architecture shown as depicted in Figure 1.

For BERT-GRU training procedure, we fine-tuned the BERT by excluding the last 3 layers as well as adding the Gaussian Noise layer followed by GRU (Chung et al., 2014) layer consist of 300 neurons, and global average pooling aims to extract the discriminative features from the past layer aims to pass them to the next layer. L2 regularization and Dropout have been used to prevent overfitting. The last layer used to predict the output predictions with a dense layer of 1 neuron, sigmoid activation function, and TruncatedNormal kernal initializers. we trained TRAC 2020 dataset without any external resources, however, in the future we will try an external dataset for the experimentation step. The the best parameter as follows: batch size= 16, optimizer= Adam, learning rate= 2e-5, and finally BERT max length= 40.

[1] https://www.tensorflow.org/hub/api_docs/python/hub/Module

parameter	Value
Embedding dimension (USE)	512
# of Estimators	3000
Sub-sample	0.3
max_depth	5
gamma	0.2
objective	(multi:softmax/ for sub-task EN-A) (binary:logistic/ for sub-task EN-B)
booster	gbtree
num_class	3 for sub-task EN-A

Table 6: The XGB classifier parameters used by grid search in the training phase

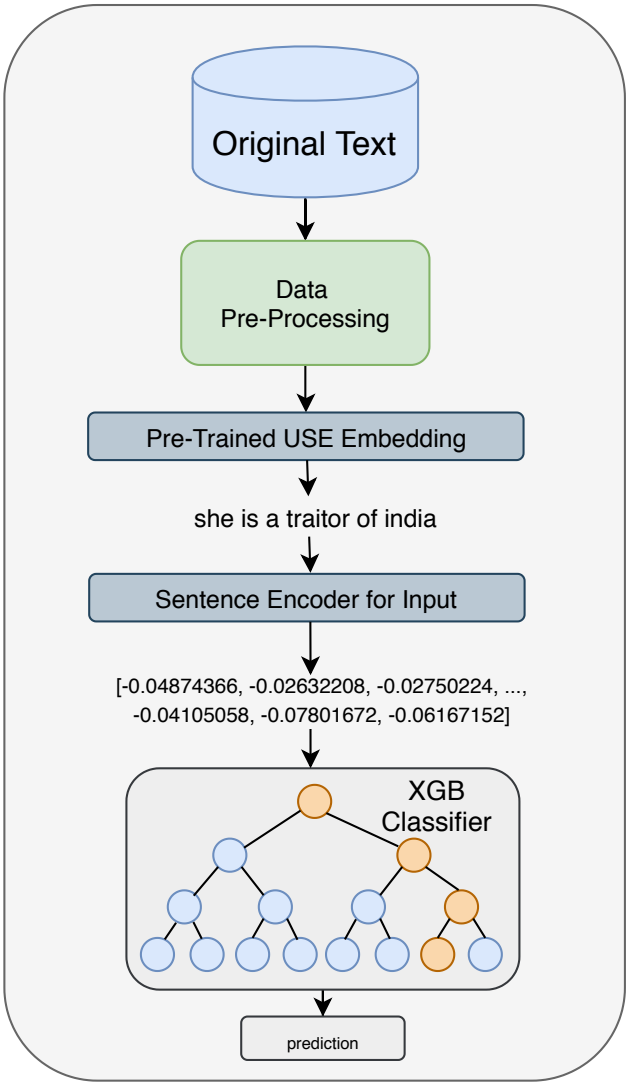

Figure 1: The architecture of our system (XGB-USE)

4. Results

4.1. Evaluation Measures

In order to evaluate the implemented approach, weighted F1 has been used according to the provided shared task. Moreover, the accuracy has been included to used for the comparison as well.

4.2. Discussion

Focusing on both sub-task EN-A and B for English language to tackle the problem of aggressive identification, table 7 presents the reported results for our proposed approaches for sub-task EN-A aggression identification shared task. The best results achieved with XGB-USE approach including the hyper-parameter that discussed above, where it achieves 0.6075 F1 (weighted) and 0.6833 accuracy. Moreover, the second approach has been used for the same sub-task achieves 0.5965 F1 (weighted) and 0.6758 accuracy where XGB-USE-PLL approach the same main approach including the pseudo label testing. The last approach using the fine-tuning BERT embedding with GRU where it achieves 0.5461 F1 (weighted) and 0.6392 accuracy. It's obvious that the XGB-USE had the best results according to F1 (weighted) and accuracy

System	F1 (weighted)	Accuracy
XGB-USE	**0.6075**	**0.6833**
XGB-USE-PLL	0.5965	0.6758
BERT-GRU	0.5461	0.6392

Table 7: Results for Sub-task EN-A (where PLL pseudo label testing, USE universal sentence encoder, and XGB XGBoot classifier)

For sub-task EN-B misogynistic aggression identification shared task, table 8 presents the reported results for our proposed approaches. The best results achieved with XGB-USE approach including the hyper-parameter that discussed above, where it achieves 0.8567 F1 (weighted) and 0.8758 accuracy. Moreover, the second approach has been used for the same sub-task achieves 0.8547 F1 (weighted) and 0.8825 accuracy where XGB-USE-PLL approach the same main approach including the pseudo label testing as a feature. The last approach using the fine-tuning BERT embedding with GRU where it achieves 0.8180 F1 (weighted) and 0.8433 accuracy. It's obvious that the XGB-USE had the best results according to F1 (weighted) and accuracy

System	F1 (weighted)	Accuracy
XGB-USE	**0.8567**	**0.8758**
XGB-USE-PLL	0.8547	0.8825
BERT-GRU	0.8180	0.8433

Table 8: Results for Sub-task EN-B (where PLL pseudo label testing, USE universal sentence encoder, and XGB XGBoot classifier)

4.3. Results and Findings

In order to show the reported results for focusing on sub-task A table 9 shows the reported results for the top teams. The best results achieve with 0.8029 F1 (weighted) presented by (Julian) team compared to our team (SAJA) achieved 0.6075 F1 (weighted).

Team	F1 (weighted)
Julian	0.8029
sdhanshu	0.7592
Ms8qQxMbnjJMgYcw	0.7568
zhixuan	0.7393
SAJA	0.6075

Table 9: Results for Sub-task EN-A compared to other teams.

For sub-task EN-B, table 10 shows the reported results for the top teams as well. The best results achieve with 0.8715 F1 (weighted) presented by (Ms8qQxMbnjJMgYcw) team compared to our team (SAJA) achieved 0.8566 F1 (weighted). We can see all the results are close to each other. We are ranking number six in this sub-task.

Team	F1 (weighted)
Ms8qQxMbnjJMgYcw	0.8715
abaruah	0.8701
na14	0.8579
sdhanshu	0.8578
SAJA	0.8566

Table 10: Results for Sub-task EN-B compared to other teams.

The figure 2 shows the confusion matrix of our best model of sub-task EN-A all for the three classes, it's clear that the XGB-USE model is performing well at classifying the non-aggressive (NAG) inputs compared to other classes. However, figure3 represents the confusion matrix for sub-task EN-B obviously the XGB-USE model is performing better for detecting the non-gendered 'NGEN' class compared to gendered'GEN' class.

5. Conclusion

In this paper, we presented our participation to TRAC 2020 shared task on aggression identification in the English language for both sub-task EN-A and EN-B. Combination of transformers have been developed to solve the provided problem, XGB-USE has been used as the main approach for this paper which extracts the USE embeddings to performs transfer learning using XGB classifier. We have been ranked fourteenth out of sixteen teams for sub-task EN-A. For sub-task EN-B, we have been ranked six out of fifteen teams which are encouraging results especially the difference between our results and the top ranked teams are very close.

This paper shows that the developed model produced great results compared to deep learning approaches and transfer learning with BERT transformers. We have used a reference dataset that provided for the TRAC 2020 shared task on aggression identification multilingual languages. The

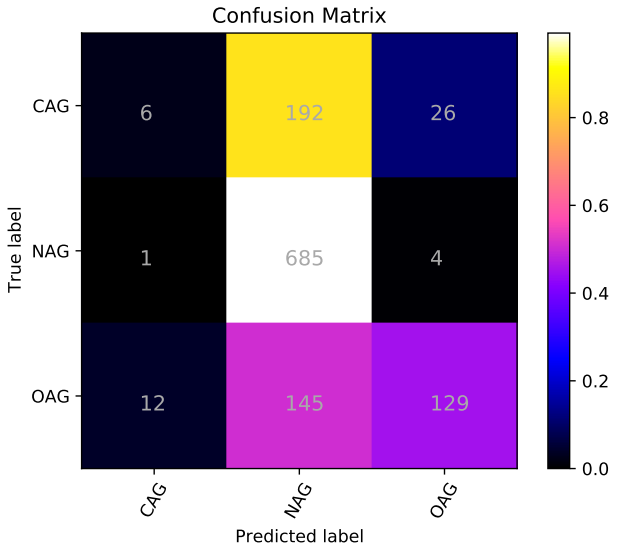

Figure 2: Sub-task EN-A, the confusion matrix for XGB-USE approach

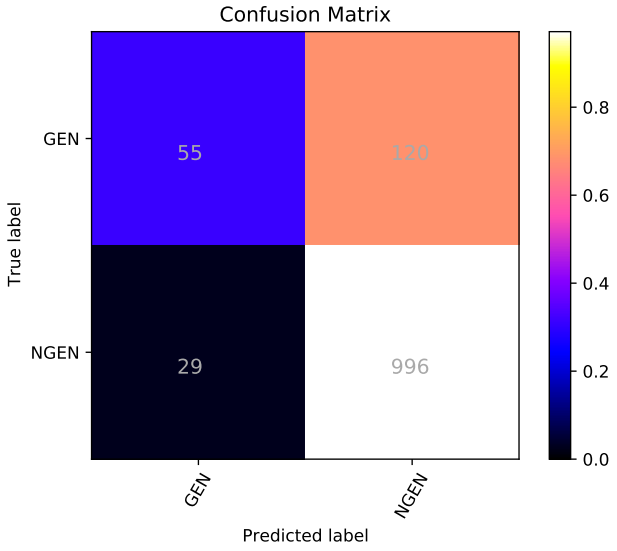

Figure 3: Sub-task EN-B, the confusion matrix for XGB-USE approach

best-reported results for sub-task EN-A achieved 0.6075 F1 (weighted) and 0.8567 F1 (weighted) for sub-task EN-B.

In the future, we will use several features and analyze them to get the best features for aggression detection. Moreover, we will study the impact of data augmentation types on the performance of various ML models.

Acknowledgements

This research is partially funded by Jordan University of Science and Technology.

6. Bibliographical References

Al-Hassan, A. and Al-Dossari, H. (2019). Detection of hate speech in social networks: a survey on multilingual corpus. In *6th International Conference on Computer Science and Information Technology*.

Aroyehun, S. T. and Gelbukh, A. (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-2018)*, pages 90–97.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

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.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., and Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.

Fišer, D., Erjavec, T., and Ljubešić, N. (2017). Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in slovene. In *Proceedings of the first workshop on abusive language online*, pages 46–51.

Fortuna, P. and Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Han, J., Wu, S., and Liu, X. (2019). jhan014 at SemEval-2019 task 6: Identifying and categorizing offensive language in social media. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 652–656, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

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 Cyberbullying (TRAC-2018), pages 1–11.

Kwok, I. and Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

Liu, P., Li, W., and Zou, L. (2019). NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Mahata, D., Zhang, H., Uppal, K., Kumar, Y., Shah, R. R., Shahid, S., Mehnaz, L., and Anand, S. (2019). MIDAS at SemEval-2019 task 6: Identifying offensive posts and targeted offense from twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 683–690, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mohaouchane, H., Mourhir, A., and Nikolov, N. S. (2019). Detecting offensive language on arabic social media using deep learning. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 466–471. IEEE.

Mubarak, H., Darwish, K., and Magdy, W. (2017). Abusive language detection on arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56.

Nikhil, N., Pahwa, R., Nirala, M. K., and Khilnani, R. (2018). Lstms with attention for aggression detection. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 52–57.

Pelicon, A., Martinc, M., and Kralj Novak, P. (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*, pages 604–610, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Ramiandrisoa, F. and Mothe, J. (2018). Irit at trac 2018. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 19–27, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Ritesh Kumar, Atul Kr. Ojha, S. M. 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)*, Paris, France, may. European Language Resources Association (ELRA).

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Roy, A., Kapil, P., Basak, K., and Ekbal, A. (2018). An ensemble approach for aggression identification in english and hindi text. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 66–73.

Samghabadi, N. S., Mave, D., Kar, S., and Solorio, T. (2018). Ritual-uh at trac 2018 shared task: aggression identification. *arXiv preprint arXiv:1807.11712*.

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. Association for Computational Linguistics*, pages 1–10, Valencia, Spain.

Soliman, A. B., Eissa, K., and El-Beltagy, S. R. (2017). Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science*, 117:256–265.

Spertus, E. (1997). Smokey: Automatic recognition of hostile messages. In *Aaai/iaai*, pages 1058–1065.

Su, H.-P., Huang, Z.-J., Chang, H.-T., and Lin, C.-J. (2017). Rephrasing profanity in chinese text. In *Proceedings of the First Workshop on Abusive Language Online*, pages 18–24.

Swamy, S. D., Jamatia, A., Gambäck, B., and Das, A. (2019). NIT_Agartala_NLP_Team at SemEval-2019 task 6: An ensemble approach to identifying and categorizing offensive language in twitter social media corpora. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 696–703, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

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 The 13th International Workshop on Semantic Evaluation (SemEval)*.