IRIT at TRAC 2018

Faneva Ramiandrisoa
IRIT, UMR5505, CNRS
Université de Toulouse, France
faneva.ramiandrisoa@irit.fr

Josiane Mothe
IRIT, UMR5505, CNRS
Université de Toulouse, France
ESPE, UT2J
Josiane.Mothe@irit.fr

Abstract

This paper describes the participation of the IRIT team to the TRAC 2018 shared task on Aggression Identification and more precisely to the shared task in English language. The three following methods have been used: a) a combination of machine learning techniques that relies on a set of features and document/text vectorization, b) Convolutional Neural Network (CNN) and c) a combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). Best results were obtained when using the method (a) on the English test data from Facebook which ranked our method sixteenth out of thirty teams, and the method (c) on the English test data from other social media, where we obtained the fifteenth rank out of thirty.

1 Introduction

In recent years, the emergence of social media platforms like Facebook, Twitter, etc. changes the way people communicate (Sticca and Perren, 2013).

Although these platforms give many benefits to their users, they can also have several negative impacts where people can be hurt for example by some aggressive texts (Dadvar et al., 2012). The aggression on social media platforms is actually more harmful than traditional bullying for many reasons such as allowing people to hide behind an alias (Sticca and Perren, 2013). Unfortunately, such phenomena of on-line aggression and bullying not only have created psychological and mental health issues for on-line users, but it can end by forcing some of them to change several things in their lives and can even conduct them to suicide according to (Kumar et al., 2018b). The TRAC 2018 workshop has been created in order to study these problems (Kumar et al., 2018a).

The TRAC 2018 workshop aims at providing the framework for evaluatig systems that aim at detecting/identifying aggression, trolling, cyberbullying and other related phenomena in both speech and text from social media (Kumar et al., 2018a). The shared task challenge consists in distinguishing three levels of aggression from text: overtly aggressive, covertly aggressive and non-aggressive. Overtly aggressive means that there is an expression of aggression directly with specific words or keywords. Covertly aggressive expresses aggression subtly such as indirect attack or with more polite expressions. Two different languages are studied independently: English and Hindi.

The shared task was organized into two different stages: training and testing. During the training stage, two datasets of $15,000$ aggression-annotated Facebook posts and comments, one in English and the other in Hindi, were provided. During testing stage, the organizers also included a data set from a second social media platform. Overall, the test data set contains four sub-collections: two from Facebook and two from another social media (not named by the organizers); for each source, one sub-collection is in English while the other is in Hindi.Participants could participate to one or several of the four subtasks which are: (1) English (Facebook) task, (2) Hindi (Facebook) task, (3) English (another social media) task, and (4) Hindi (another social media).

More details about the task can be found in (Kumar et al., 2018a). Our team, IRIT, participated to shared task in English language for both Facebook and the other social media platform: subtask (1) and (3).

In this paper, we report the methods we proposed when participating to the subtasks; we have developed three approaches that we compare in this paper on the problem of aggression identification in texts.

The first method is a combination of two classifiers : *random forest* that relies on features ranging from surface features to more linguistic features and *logistic regression* based on text vectorization. We named this method Trac-RF_LR. The second approach is a deep learning technique widely used in image area and adapted for text classification. It uses a Convolutional Neural Network and we named it Trac-CNN in the rest of the paper. Finally, the third approach is a combination of two deep learning techniques: Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). We named this method Trac-CNN_LSTM.

The remaining of this paper is organized as follows: Section 2 is an overview of state-of-the-art approaches for aggression detection. Section 3 details the methods we developed. Section 4 reports the results as well as the data sets and Section 5 concludes this paper.

2 Related Work

Several series of evaluation forums for applications related to social media have been developed in the recent years, such as tweet contextualisation (Ermakova et al., 2017) or e-risk (Losada et al., 2017) in CLEF. TRAC is the first that focus on detecting aggressive text. In the last few years, several studies have been published in which the problem of detecting abusive language in social media has been tackled. Although researchers focused on different aspects of abusive language such as cyberbullying (Dadvar et al., 2013), hate speech (Warner and Hirschberg, 2012), profanity (Sood et al., 2012) and abusive language in general (Chen et al., 2012), most of the approaches are based on supervised approaches (Schmidt and Wiegand, 2017). Generally, the detection of abusive language is considered as a classification problem, mainly as a binary text classification problem (Bosque and Garza, 2014) detecting whether the text contains abusive parts or not.

For text classification problems, approaches based on features are widely used.

Methods are based on simple *surface features* such as unigrams and/or n-grams (Xu et al., 2012; Chen et al., 2012; Waseem and Hovy, 2016; Abdou Malam et al., 2017), key-phrases (Mothe et al., 2018), the frequency of punctuation, capitalized words, Htag, image, URL mentions (Hoang and Mothe, 2017), average length of words, or frequency of words that do not exist in English dictionaries (Chen et al., 2012; Nobata et al., 2016). Other features are: *word generalization* (for example using Latent Dirichlet Allocation (LDA) (Zhong et al., 2016) and word/paragraph embedding (Nobata et al., 2016)), *sentiment analysis* (Xu et al., 2012; Dinakar et al., 2012; Hee et al., 2015); *lexical resources* (Burnap and Williams, 2015; Burnap and Williams, 2016; Gitari et al., 2015; Hoang and Mothe, 2018); *linguistic features* (Xu et al., 2012; Gitari et al., 2015; Burnap and Williams, 2016; Nobata et al., 2016; Abdou Malam et al., 2017); *knowledge-based features* (Dinakar et al., 2012); and *meta-information* (Dadvar et al., 2012; Dadvar et al., 2013; Waseem and Hovy, 2016).

Using these features, supervised classifiers such as Support Vector Machines (SVM) are used in order to classify a text as containing aggressiveness or not (Schmidt and Wiegand, 2017).

In recent years, deep learning has been also employed. For example, (Mehdad and Tetreault, 2016) used Recurrent Neural Network Language Model which is based on Recurrent Neural Networks for the task of abusive language detection. The choice of this type of models is based on the fact that with few training data, they can achieve good results for language models.

In addition of detecting if a text contains something harmful and/or aggressive, another challenge is to distinguish the different aspects of the abusive language. (Malmasi and Zampieri, 2018) try to distinguish general profanity from hate speech, but their results show that it is difficult to differentiate one from the other. Their study also shows that all hate speech, bullying text, cyberbullying and abusive language are not explicit, there are aggressive text written with subtle language and it can be hard to distinguish them. This observation motivated the shared task not only to classify the text as aggressive or not, but also to distinguish two kinds of aggression: overt and covert aggression.

3 Developped methods for the IRIT participation

In this section, we present the three supervised approaches we developed to automatically detect aggression in texts and that we submitted to TRAC 2018 shared task for English.

3.1 Trac-RF_LR: combination of two classifiers

In this model we combine two classifiers namely random forest and logistic regression where the first one is based on different set of features from surface features to more linguistic features and the second one is based on document vectorization. In the following, we first describe each classifier, then we detail the combination method used.

Classifiers description

i) Features - RF Classifier:

This model uses different features, adapted from features used for depression detection in (Abdou Malam et al., 2017), and are computed from the texts to predict the aggression. We represent texts with a vector composed of the features as presented in Table 1.

Name	Hypothesis or tool/resource used
Part-Of-Speech frequency	Normalized frequencies of each tag: adjectives, verbs, nouns and adverbs (four features). The idea behind is to check offensive words used as nouns, verbs, adjectives, or adverbs.
Negation	Normalized frequencies of negative words like: *no, not, didn't, can't*, ... The idea behind is to detect non direct aggressiveness.
Capitalized	The idea behind is that aggressive texts tend to put emphasis on the target they mention. It can indicate feelings or speaking volume.
Punctuation marks	! or ? or any combination of both can emphasize offensiveness of texts.
Emoticons	Another way to express sentiment or feeling.
Sentiment	Use of NRC-Sentiment-Emotion-Lexicons[1] to trace the polarity in text.
Emotions	Frequency of emotions from specific categories: anger, fear, surprise, sadness and disgust. The idea behind is to check the categories related to aggressiveness.
Gunning Fog Index	Estimate of the years of education that a person needs to understand the text at first reading.
Flesch Reading Ease	Measure how difficult to understand a text is.
Linsear Write Formula	Developed for the U.S. Air Force to calculate the readability of their technical manuals[2].
New Dale-Chall Readability	Measure the difficulty of comprehension that persons encounter when reading a text. It is inspired from Flesch Reading Ease measure.
Swear words	The intuition behind is that the texts containing insults are often aggressive.
Lexical analysis with python library *empath*	Empath is a tool for analyzing text across lexical categories. By default, it has 194 lexical categories and each category is considered as feature.

Table 1: List of features used in RF.

Some of these features are used for abusive language detection, hate speech, cyberbullying and the others are used for sentiment or personality analysis that we judged useful for aggression detection.

A random forest classifier was trained on train and validation sets by representing each text with a vector composed by the features we mentioned above. The following parameters were used during the training: class_weight="balanced", max_features="sqrt", n_estimators=60, min_weight_fraction_leaf=0.0, criterion='entropy', random_state=2.

At prediction time, a text from the test set is represented with features and then run the trained model. The output is the estimated probabilities for the three classes (overtly aggressive, covertly aggressive and non-aggressive).

ii) Document vector - LR Classifier:

This model is based on document vectorization using *Doc2vec* (Le and Mikolov, 2014). *Doc2vec* is used to represent sentences, paragraphs, or whole documents as vectors and it can be trained on small corpora, which is case of the task datasets.

Before building the model (LR Classifier), we first trained two separate *Doc2vec* models: a Distributed Bag of Words and a Distributed Memory model (Le and Mikolov, 2014). For the training, we used the same configuration as in (Trotzek et al., 2017) for representing user's text in order to detect if he or she is depressed. The two *Doc2vec models are trained* on English text from the train and validation sets and we used the Python package *gensim*[3](Rehurek and Sojka, 2010). We also concatenated the output vectors of these two models, as done in (Trotzek et al., 2017), resulting in a representation by a 200-dimension vector per text.

Then a logistic regression classifier was trained on the vectors for both the train and validation sets with the following parameters : class_weight="balanced", random_state=1, max_iter=100, solver="liblinear".

At prediction time, the texts from the test set were vectorized by using the two *Doc2vec* models and the 200-dimension vectors were given as input of trained classifier. The output is also a set of class probabilities.

Combination of two classifiers

After the two classifiers (as described in section 3.1) provided their results, the calculated class probabilities obtained from RF classifier and LR Classifier were averaged and finally the class with the highest probability was considered as the class the text belongs to.

3.2 Trac-CNN: Model based on CNN

This model is based on the deep learning technique known as Convolutional Neural Networks (CNN)(LeCun et al., 1998). CNN which is widely used in image analysis (Bhandare et al., 2016) and has been adapted for text classification such as sentiment analysis (Chen et al., 2017). Figure 1 shows the model architecture that we used for TRAC 2018 shared task.

We can see in this figure that before sending the text to the CNN, it is transformed as a matrix where each row is a vector representation[4] of words that compose the text[5] . To perform convolutions on the sentence matrix, we decided to use three sizes of filters[6]: bigrams (height = 2), trigrams (height = 3) and fourgrams (height = 4) filters, each of which has 100 filters (in total we had 300 filters). The result of convolutions is called feature map, a vector with variable-length according to the filter size and we had 300 features map. The 1-max pooling (Boureau et al., 2010), a common strategy, is performed over each

[1] `http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`, Accessed on 2017-02-23

[2] `http://www.streetdirectory.com/travel_guide/15675/writing/how_to_choose_the_best_readability_formula_for_your_document.html`, Accessed on 2018-02-25

[3] `https://radimrehurek.com/gensim/index.html`

[4] We used *word2vec* trained on the training and validation sets. Two models were trained, namely CBOW and Skip-gram (Mikolov et al., 2013), and the final vector is a 200-dimensional vector which is a concatenation of the results from these models.

[5] All the texts need to be the same length in order to have the same matrix size. We used Keras zero-pads at the beginning if a text length is shorter than the maximum length (= length of longer text/document in the datasets).

[6] The width of filters are the same as word vectors (i.e 200) but the heights are different.

feature map. More precisely, the largest number from each feature map is kept and then concatenated to form a concatenated vector, where the dimension is equal to 300. Then we added one fully connected hidden layer to reduce the dimension of the concatenated vector , the dropout is performed on the hidden layer because dropout helps to reduce over fitting (Srivastava et al., 2014). Finally, the latest feature vector is fed through a sigmoid function to generate the final classification.

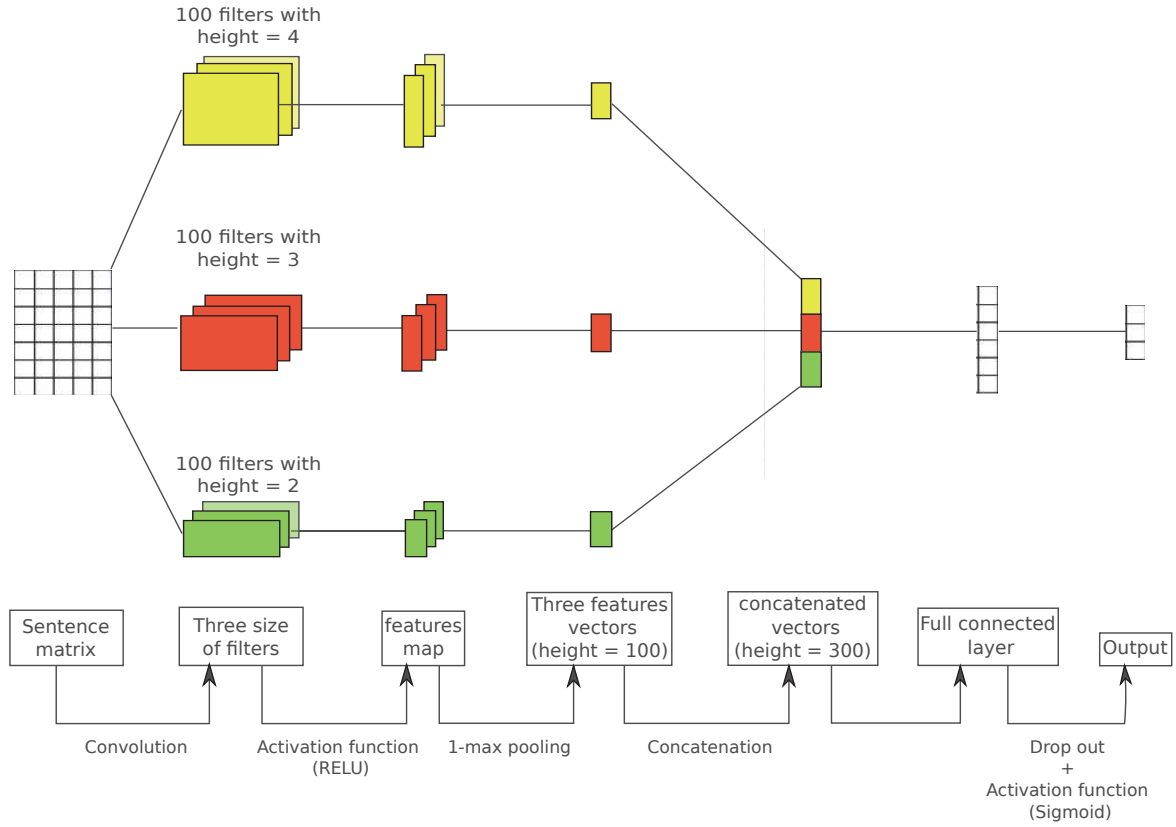

Figure 1: Illustration of a CNN architecture for aggression detection inspired from (Zhang and Wallace, 2015).

3.3 Trac-CNN_LSTM: Combination of CNN and LSTM

This model combines two deep learning techniques: CNN and LSTM. The main idea is to modify the CNN architecture presented in Figure 1 by adding a LSTM layer and change the activation function to softmax for the last classification. The LSTM is inserted after the poling layer and before the fully connected hidden layer. It takes as input the 300-dimension concatenated feature vector and gives as output a 300-dimension vector. Figure 2 presents the architecture of the combination.

4 Results

In this section, we detail the data used for TRAC 2018 shared task and the result we obtained using our different models.

4.1 Data

The way the dataset used in the TRAC 2018 shared task was built is described in (Kumar et al., 2018b).

The dataset in the shared task was divided in three sets: training, validation and test. The training and validation sets were released during the training stage to allow participants to train and built their systems. These two sets are composed of 15,000 aggression-annotated Facebook posts and comments written in English for participants who want to participate in this language and 15,001 aggression-annotated Facebook writen in Hindi for those who prefer to work in this language.

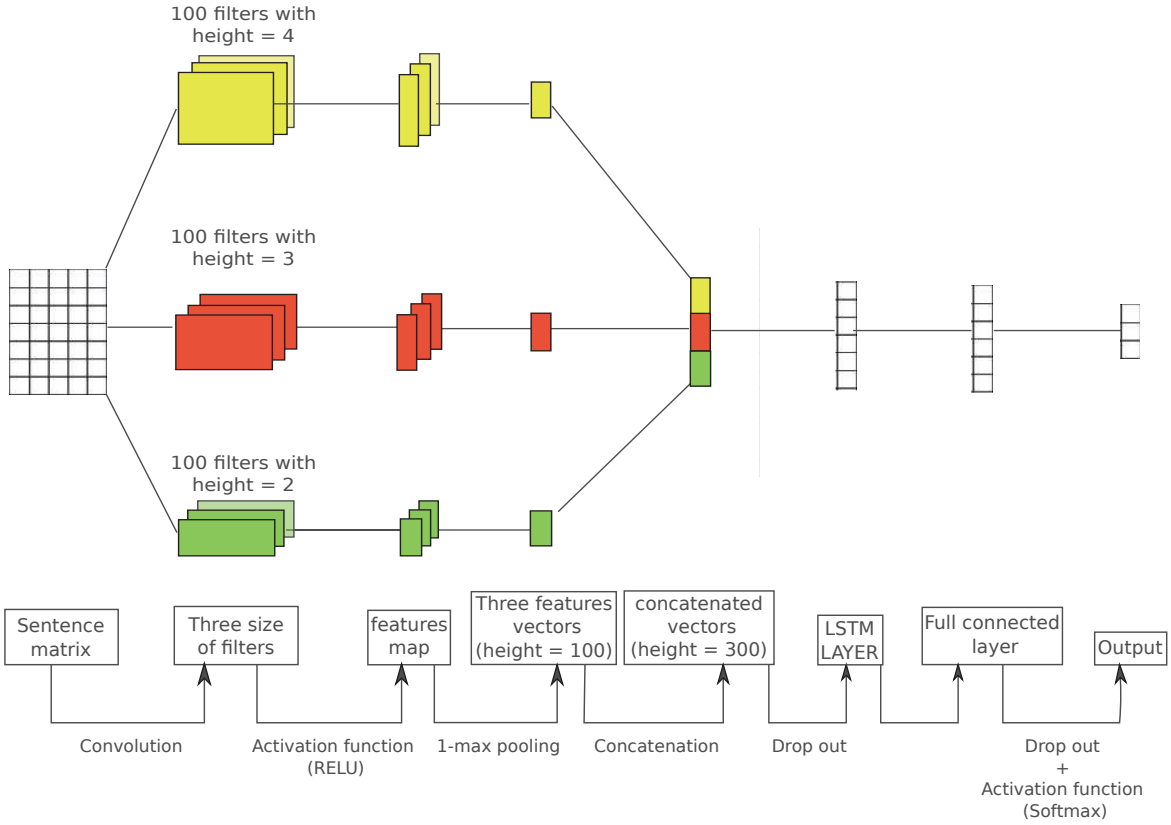

Figure 2: Illustration of a CNN + LSTM architecture for aggression detection inspired from (Zhang and Wallace, 2015).

For the test set, the collection crawled from Facebook is also divided into two languages, 916 posts and comments for the English dataset and 970 for the Hindi dataset. The organizers also included what they called the surprise collection crawled from another social media, also divided into two languages (1,257 posts and comments for English and 1,194 for Hindi). Table 2 describes the datasets used in the shared task, it reports the total number of posts and comments (texts) in each set and the number of texts with overt and covert aggression and texts without aggression. We can see from the Table 2 that the proportion of aggressive text is more important in the Hindi dataset (13,803 texts which correspond to 80.4% of total aggressive text (17,173)) than in the English dataset (9,776 texts which correspond to 56.9% of total aggressive text (17,165)).

Language	Number of	Train	Validation	Test	
				Facebook	other social media
English	texts (=posts+comments)	11,999	3,001	916	1,257
	Overt aggression	2,708	711	144	361
	Covert aggression	4,240	1,057	142	413
	No aggression	5,051	1,233	630	483
Hindi	texts (=posts+comments)	12,000	3,001	970	1,194
	Overt aggression	4,856	1,217	362	459
	Covert aggression	4,869	1,246	413	381
	No aggression	2,275	538	195	354

Table 2: Distribution of training, validation and testing data on TRAC 2018 data collection.

4.2 Models

All submitted models was evaluated using weighted macro-averaged F-scores. This measure is defined by organizers as follow: for each class, the individual F-score was weighted by the proportion of the concerned class in the test set and then the average of these individual weighted F-scores is the final F-score (See (Kumar et al., 2018a)). In order to have a baseline for comparison, the organizers give a random baseline generated by assigning random labels.

Table 3 reports the results we obtained with the three models (see Section 3) we submitted for the shared task and the baseline given by organizers. We can see that all of our models outperform the baseline on both Facebook and the other social media platform. More precisely Trac-RF gives the best results on Facebook and Trac-CNN_LSTM on the other social media.

We can also see that our three models give better results on Facebook, this could be due to the train dataset which is only composed of texts crawled from Facebook.

System	**F1 (weighted)**	
	Facebook	**other social media**
Random Baseline	0.354	0.348
Trac-RF_LR	**0.576**	0.405
Trac-CNN	0.562	0.494
Trac-CNN_LSTM	0.559	**0.511**

Table 3: Results for the English (Facebook and other social media) task. Bold value is the best performance.

When compared to other participants, our best model on both social media platforms give encouraging results. We are ranked sixteenth out of thirty teams on Facebook and fifteenth for other social media (see Table 4).

Rank	**Team**	**Facebook**
1	saroyehun	0.642
2	EBSI-LIA-UNAM	0.632
3	DA-LD-Hildesheim	0.618
4	TakeLab	0.616
5	sreeIN	0.604
...	...	...
16	**IRIT** (Trac-RF_LR)	0.576

Rank	**Team**	**other social media**
1	vista.ue	0.601
2	Julian	0.599
3	saroyehun	0.920
4	EBSI-LIA-UNAM	0.572
5	uOttawa	0.569
...	...	...
15	**IRIT** (Trac-CNN_LSTM)	0.511

Table 4: Results of top 5 teams vs our results which are in Bold.

5 Conclusion and Future Work

In this paper, we presented our participation to TRAC 2018 shared task on aggression identification in English language for both platforms: Facebook and the other social media. We proposed three kinds different methods: (a) a combination of machine learning techniques (classifiers) that relies on a set of features and document vectorization, (b) Convolutional Neural Network (CNN); and (c) combination of CNN and Long Short-Term Memory (LSTM). We obtained encouraging results: sixteenth out of thirty teams with method (a) for Facebook and fifteenth out of thirty for the other social media with method (c).

We can conclude that the model based on features gives the best results if the train dataset is built from the same platform as the test dataset. However a deeper analysis of features is to be done to get stronger conclusions. We can also see that deep learning models perform well even if the train and test datasets are built from different platforms (unlike combined RF and LR models). So these techniques are useful to build systems that can identify aggression on different social media.

For future work, we will analyze what the best features are for aggression detection by analyzing those we already used and/or adding new ones. For example, more linguistic-oriented features such as

specific writing or structure of sentences/paragraphs containing aggression could be added and feature like ”proportion of aggressive text posted last year”. Finally, we would like to improve our model based on deep learning by using different configurations and architectures, by combining it with traditional machine learning algorithms (RF and LR models), and also training a model on large datasets built from different social media platforms.

References

Idriss Abdou Malam, Mohamed Arziki, Mohammed Nezar Bellazrak, Farah Benamara, Assafa El Kaidi, Bouchra Es-Saghir, Zhaolong He, Mouad Housni, Véronique Moriceau, Josiane Mothe, and Faneva Ramiandrisoa. 2017. IRIT at e-Risk (regular paper). In *International Conference of the CLEF Association, CLEF 2017 Labs Working Notes*, volume 1866 of *ISSN 1613-0073*, http://CEUR-WS.org. CEUR Workshop Proceedings.

Ashwin Bhandare, Maithili Bhide, Pranav Gokhale, and Rohan Chandavarkar. 2016. Applications of convolutional neural networks. *International Journal of Computer Science and Information Technologies*, pages 2206–2215.

Laura P. Del Bosque and Sara Elena Garza. 2014. Aggressive text detection for cyberbullying. In *Human-Inspired Computing and Its Applications - 13th Mexican International Conference on Artificial Intelligence, MICAI 2014, Tuxtla Gutiérrez, Mexico, November 16-22, 2014. Proceedings, Part I*, pages 221–232.

Y-Lan Boureau, Jean Ponce, and Yann LeCun. 2010. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 111–118.

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, 5(1):11.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and 2012 International Confernece on Social Computing, SocialCom 2012, Amsterdam, Netherlands, September 3-5, 2012*, pages 71–80.

Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. 2017. Improving sentiment analysis via sentence type classification using bilstm-crf and CNN. *Expert Syst. Appl.*, 72:221–230.

Maral Dadvar, de FMG Jong, Roeland Ordelman, and Dolf Trieschnigg. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *Advances in Information Retrieval*, pages 693–696. Springer.

Karthik Dinakar, Birago Jones, Henry Lieberman, Rosalind W. Picard, Carolyn Penstein Rosé, Matthew Thoman, and Roi Reichart. 2012. You too?! mixed-initiative LDA story matching to help teens in distress. In *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012*.

Liana Ermakova, Lorraine Goeuriot, Josiane Mothe, Philippe Mulhem, Jian-Yun Nie, and Eric SanJuan. 2017. Clef 2017 microblog cultural contextualization lab overview. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 304–314. Springer.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Recent Advances in Natural Language Processing, RANLP 2015, 7-9 September, 2015, Hissar, Bulgaria*, pages 672–680.

Thi Bich Ngoc Hoang and Josiane Mothe. 2017. Predicting Information Diffusion on Twitter - Analysis of predictive features. *Journal of Computational Science*, 22, octobre.

Thi Bich Ngoc Hoang and Josiane Mothe. 2018. Location extraction from tweets. *Information Processing & Management*, 54(2):129–144.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018a. Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling (TRAC)*, Santa Fe, USA.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018b. Aggression-annotated Corpus of Hindi-English Code-mixed Data. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1188–1196.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

David E Losada, Fabio Crestani, and Javier Parapar. 2017. erisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 346–360. Springer.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in Discriminating Profanity from Hate Speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30:1–16.

Yashar Mehdad and Joel R. Tetreault. 2016. Do characters abuse more than words? In *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 13-15 September 2016, Los Angeles, CA, USA*, pages 299–303.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Josiane Mothe, Faneva Ramiandrisoa, and Michael Rasolomanana. 2018. Automatic Keyphrase Extraction using Graph-based Methods (regular paper). In *ACM Symposium on Applied Computing (SAC), Pau, France, 09/04/2018-13/04/2018*. ACM.

Chikashi Nobata, Joel R. Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 145–153.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

Anna Schmidt and Michael Wiegand. 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.

Sara Owsley Sood, Judd Antin, and Elizabeth F. Churchill. 2012. Using crowdsourcing to improve profanity detection. In *Wisdom of the Crowd, Papers from the 2012 AAAI Spring Symposium, Palo Alto, California, USA, March 26-28, 2012*.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Fabio Sticca and Sonja Perren. 2013. Is cyberbullying worse than traditional bullying? examining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying. *Journal of Youth and Adolescence*, 42(5):739–750, May.

Marcel Trotzek, Sven Koitka, and Christoph M. Friedrich. 2017. Linguistic metadata augmented classifiers at the CLEF 2017 task for early detection of depression. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 88–93.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada*, pages 656–666.

Ye Zhang and Byron C. Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *CoRR*, abs/1510.03820.

Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J. Miller, and Cornelia Caragea. 2016. Content-driven detection of cyberbullying on the instagram social network. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 3952–3958.