NIT_Agartala_NLP_Team at SemEval-2019 Task 6: An Ensemble Approach to Identifying and Categorizing Offensive Language in Twitter Social Media Corpora

Steve Durairaj Swamy[1], Anupam Jamatia[1], Björn Gambäck[2] and Amitava Das[3]

[1]National Institute of Technology, Agartala, India
[2]Norwegian University of Science and Technology, Trondheim, Norway
[3]Mahindra École Centrale,Hyderabad, Telangana, India
{steve050798,anupamjamatia}@gmail.com,gamback@ntnu.no,amitava.das@mechyd.ac.in

Abstract

The paper describes the systems submitted to OffensEval (SemEval 2019, Task 6) on 'Identifying and Categorizing Offensive Language in Social Media' by the 'NIT_Agartala_NLP_Team'. A Twitter annotated dataset of 13,240 English tweets was provided by the task organizers to train the individual models, with the best results obtained using an ensemble model composed of six different classifiers. The ensemble model produced macro-averaged F_1-scores of 0.7434, 0.7078 and 0.4853 on Subtasks A, B, and C, respectively. The paper highlights the overall low predictive nature of various linguistic features and surface level count features, as well as the limitations of a traditional machine learning approach when compared to a Deep Learning counterpart.

1 Introduction

Offensive language has been the scourge of the internet since the rise of social media. Social media provides a platform for everyone and anyone to voice their opinion. This has empowered people to make their voices heard and to speak out on global issues. The downside to this, however, is the misuse of such platforms to attack an individual or a minority group, and to spread hateful opinions. Pairing this with the perceived anonymity the internet provides, there has been a massive upswing in the use of social media for cyberbullying and hate speech, with technology giants coming under increased pressure to address the issue.

Most of what we may be interested in detecting can be broadly labelled as hate speech, cyberbullying or abusive use of swearing. The union of these three subsets form what can be identified as 'Offensive Language on Social Media'. However, what we consider offensive is often a grey area, as is evident by the low inter-annotator agreement rates when labelling data for offensive language (Waseem et al., 2017b).

Detecting offensive language has proven to be difficult, due to the broad spectrum in which language can be used to convey an insult. The nature of the abuse can be implicit — drawing from sarcasm and humour rather than offensive terms — as well as explicit, by making extensive use of traditional offensive terms and profanity. It does not help that the reverse is also entertained, with profanity often being used to imply informality in speech or for emphasis. Coincidentally, these are also the reasons why lexical detection methods have been unfruitful in classifying text as offensive or non-offensive.

The OffensEval 2019 shared task (Zampieri et al., 2019b) is one of several endeavours to further the state-of-the-art in addressing the offensive language problem. The paper describes the insights obtained when tackling the shared task using an ensemble of traditional machine learning classification models and a Long Short-Term Memory (LSTM) deep learning model. Section 2 first discusses other related approaches to detecting hate speech and offensive language. Then Section 3 describes the dataset and Section 4 the ideas and methodology behind our approach. Section 5 reports the results obtained, while Section 6 discusses those results with a particular eye towards the errors committed by the models. Finally, Section 7 sums up the key results and points to ways the work can be extended.

2 Related Work

Most datasets for offensive language detection represent multiclass classification problems (Davidson et al., 2017; Founta et al., 2018; Waseem and Hovy, 2016), with the annotations often obtained via crowd-sourcing portals, with

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 696–703
Minneapolis, Minnesota, USA, June 6–7, 2019.

varying degrees of success. Waseem et al. (2017b) state that annotation via crowd-sourcing tends to work best when the abuse is explicit (Waseem and Hovy, 2016), but is considerably less reliable when considering implicit abuse (Dadvar et al., 2013; Justo et al., 2014; Dinakar et al., 2011). They propose a typology that can synthesise different offensive language detection subtasks. Zampieri et al. (2019a) expand on these ideas and propose a hierarchical three-level-annotation model, which is used in the OffensEval 2019 shared task. Another issue is whether the datasets should be balanced or not (Waseem and Hovy, 2016), since there are much fewer offensive comments than benign comments in randomly sampled real-life data (Schmidt and Wiegland, 2017).

Classical Machine learning algorithms have been wielded to some success in automated offensive language detection, mainly Logistic Regression (Davidson et al., 2017; Waseem and Hovy, 2016; Burnap and Williams, 2015) and Support Vector Machines (Xu et al., 2012; Dadvar et al., 2013). Recently, however, deep learning models have outperformed their traditional machine learning counterparts, with both Recurrent Neural Networks (RNN) — such as LSTM (Pitsilis et al., 2018) and Bi-LSTM (Gao and Huang, 2017) — and Convolutional Neural Networks (CNN) having been used. Gambäck and Sikdar (2017) utilised a CNN model with word2vec embeddings to obtain higher F_1-score and precision than a previous logistic regression model (Waseem and Hovy, 2016), while Zhang et al. (2018) combined a CNN model with a Gated Recurrent Unit (GRU) layer. Malmasi and Zampieri (2018) used an ensemble system much like ours to separate profanity from hate speech, but reported no significant improvement over a single classifier system.

In terms of features, simple bag of words models have proven to be highly predictive (Waseem and Hovy, 2016; Davidson et al., 2017; Nobata et al., 2016; Burnap and Williams, 2015). Mehdad and Tetreault (2016) endorsed the use of character n-grams over token n-grams citing their ability to glaze over the spelling errors that are frequent in online texts. Nobata et al. (2016); Chen et al. (2012) showed small improvements by including features capturing the frequency of different entities such as URLs and mentions, with other features such as part-of-speech (POS) tags (Xu et al., 2012; Davidson et al., 2017) and sentiment scores (Van Hee et al., 2015; Davidson et al., 2017) also having been used (Schmidt and Wiegland, 2017). More recently, meta information about the users have been suggested as features, but no consistent correlation between user information and tendency for offensive behaviour online has been shown, with Waseem and Hovy (2016) claiming gender information leading to improvements in classifier performance, but with Unsvåg and Gambäck (2018) challenging this and reporting user-network data to be more important instead. Wulczyn et al. (2017) concluded that anonymity leads to an increase in the likelihood of a comment being an attack.

3 Data

The training dataset used for the shared task, the Offensive Language Identification Dataset (Zampieri et al., 2019a), contains 13,240 tweets, with each tweet having been annotated on the basis of a hierarchical three-level model. An additional 860 tweets were used as the test set for the shared task. The three levels/subtasks are as follows:

A – Whether the tweet is offensive (OFF) or non-offensive (NOT).

B – Whether the tweet is targeted (TIN) or untargeted (UNT).

C – If the target is an individual (IND), group (GRP) or other (OTH; e.g., an issue or an organisation).

The dataset does not have an equal number of offensive and non-offensive tweets. Only about one-third of the tweets are marked offensive, to partially account for the fact that most online discourse mainly is non-offensive. The corpus exhibits a larger number of male (∼3000) than female pronouns (∼2500), but is reasonably balanced.

Noticeably, the annotators were very conservative in their classification of tweets as non-offensive. It is unclear whether this was due to a more strict definition provided by the task organisers. For example, it is not immediately clear why tweets such as:[1] *"@USER Ouch!" (23159)*, *"@USER He is a beast" (50771)*, and *"@USER That shit weird! Lol" (31404)* were annotated as offensive.

The annotators furthermore seemed to disagree over the cathartic and emphatic use of swearing, as in *"@USER Oh my Carmen. He is SO FRICK-*

[1] In the examples, tweet IDs are given in parenthesis.

ING CUTE" (39021), *"@USER GIVE ME A FUCKING MIC" (60566)*, and *"@USER why are you so fucking good." (80097)*. These tweets do not really seem to be offensive except for them containing varying degrees of profanity. However, this is inconsistent, with some other tweets annotated not offensive, as expected: *"@USER No fucking way he said this!" (47427)*, and *"@USER IT'S FUCKING TIME!!" (59465)*, although most tweets that contained profanity were included in the offensive class.

Another thing to note is a large amount of political criticism within the tweets in the corpus. Whether it be left wing or right wing, extreme cases seem to be correctly annotated as offensive, while a healthy amount of criticism and political discourse correctly is annotated as non-offensive. The dataset also exhibits a dearth of racist tweets.

4 Methodology

Initially, a suite of features was composed based on those used successfully in previous work such as Waseem and Hovy (2016), Davidson et al. (2017), Nobata et al. (2016) and Burnap and Williams (2015): surface-level token unigrams, bigrams, and trigrams, weighted by TF-IDF; POS tags obtained through the `CMU tagger`[2] (Gimpel et al., 2011), which was specifically developed for the language used on Twitter; sentiment score assigned using a pre-trained model included in `TextBlob`[3]; and count features for URLs, mentions, hashtags, punctuation marks, words, syllables, and sentences.

`Scikit-learn`[4] (Pedregosa et al., 2011) was used as the primary library for modelling and training. L1-regularised Logistic Regression and a Linear Support Vector Classifier stood out initially as the best models. Further experimentation displayed that while those two models exhibited the highest accuracy, their recall of offensive tweets in subtask A and of untargeted insults in subtask B were lower than other classifiers provided in the Scikit-learn library, such as the Passive-Aggressive (PA) classifier (Crammer et al., 2006) and stochastic gradient descent (SGD).

Further exploration showed that the classifiers were not in agreement on certain tweets. This led to the idea of a vote-based ensemble model built on the following five classifiers combined by plurality voting (Kuncheva, 2004): L1-regularised Logistic Regression, L2-regularised Logistic Regression, Linear SVC, SGD, and PA. The ensemble model exhibited the best results in subtasks A and B. In subtask C, the multi-class classification problem and a severe reduction of the size of the training set led to much lower macro-averaged F_1-scores, with the ensemble model performing badly. A deep learning approach, based on an LSTM architecture (Hochreiter and Schmidhuber, 1997), was adopted specifically for this subtask. The model used a 200 dimensional `GloVe embedding`[5] pre-trained on 2 billion tweets (Pennington et al., 2014), with trainability set to False. The embedding layer was followed by a 1D convolution layer with 64 output filters and a Rectified Linear Unit (ReLU) activation function. The output of this layer was down-sampled using a max pooling layer of size 4. These inputs were fed into an LSTM layer of 200 units and subsequently a dense layer of 3 units with a softmax activation function. The model used the 'Adam' optimiser and the categorical cross entropy loss function. Due to the less amount of data, overfitting was quite common on as few as 3 epochs. Therefore, the model benefited from larger dropout values (up to 0.5). This model exhibited a better result than the ensemble model in subtask C, although only by a small margin.

5 Results

The experiments were run in three stages. First, before choosing the models, a mini ablation study was carried out on how various features affected the accuracy and F_1-score metrics of different models. The selected models were then optimised on the training set, before being evaluated on the test dataset.

5.1 Feature Engineering

The initial ablation study was carried out on a small sample space of models: the Linear SVC and L1/L2-penalised Logistic Regression. The results are represented in Table 1.

The ablation analysis revealed that surface-level token/character n-grams are by far the most predictive of the features. An interesting observation is the significantly improved recall of offensive tweets when character n-grams are included.

[2] `www.cs.cmu.edu/~ark/TweetNLP/`
[3] `textblob.readthedocs.io/en/dev/`
[4] `scikit-learn.org/stable/`
[5] `nlp.stanford.edu/projects/glove/`

Features	Linear SVC			Logistic Regression L1			Logistic Regression L2		
	Acc	F_1	Rec	Acc	F_1	Rec	Acc	F_1	Rec
(1,3) word n-gram	.7694	.7257	.4884	.7671	**.7309**	.5938	.7583	.7193	.5684
+ POS tags	.7647	.7192	.4650	.7584	.7155	.5659	.7482	.7072	.5502
+ Sentiment Score	.7635	.7188	.4920	.7399	.7057	.5997	.7261	.6894	.5752
+ Sentiment Score - POS Tags	.7694	.7265	.5068	.7558	.7226	.6161	.7384	.7033	.5929
+ Count Features	.7673	.7234	.4925	.7422	.7096	.6125	.7327	.7004	.6075
+ Count Features - POS Tags	**.7710**	.7286	.5115	.7587	.7260	.6209	.7478	.7149	.6138
(2,5) char n-gram	.7532	.7185	.6032	.7376	.7080	.6288	.7487	.7203	.6459
+ Sentiment Score	.7539	.7189	.6015	.7315	.7060	.6490	.7503	.7240	.6604
+ Count Features	.7534	.7191	.6068	.7323	.7077	.6557	.7523	.7262	**.6636**

Table 1: Ablation analysis on subtask A, with the training set.

	Subtask A				Subtask B			Subtask C				
	All NOT	All OFF	Linear SVC	Ensemble model	All TIN	All UNT	Ensemble model	All GRP	All IND	All OTH	Ensemble model	LSTM network
F_1	.4189	.2182	.7369	**.7434**	.4702	.1011	**.7079**	.1787	.2130	.0941	.4854	**.5056**
Acc.	.7209	.2790	.8012	**.8023**	**.8875**	.1125	.8833	.3662	.4695	.1643	.6291	**.6385**

Table 2: Test set results (macro-F_1 and accuracy) for all subtasks, with class baselines ("All X").

However, the best F_1-score/accuracy was never achieved with the character n-gram model, and hence only token n-grams were included on the final feature list. Other features provided only small improvements in accordance with previous observations (Wiegand et al., 2018). The addition of POS information seems to cause a reduction in performance, so this feature was dropped, except for in subtask C, where a small positive effect could be observed. Furthermore, artificially balancing the classes by modifying class weights helped alleviate the low recall issue to some extent.

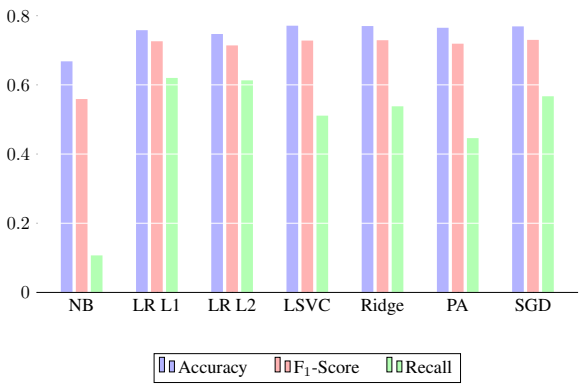


Figure 1: Performance of individual classifiers

5.2 Training Set

A 10-fold cross-validation was performed on each model used in the ensemble, with the metrics obtained in each fold averaged to obtain a median for each model's performance on the dataset. These initial results were obtained only for subtask A, to decide which models would be a part of the ensemble. Most models used in the ensemble exhibited similar accuracy, but varied in the recall of offensive tweets. It was also observed that models with the higher recall of offensive tweets exhibited equivalently lower recall of non-Offensive tweets. These observations are graphically represented in Figure 1. Small improvements in F_1-score and accuracy were achieved while using the **ensemble model (F_1-score: .7338 and Accuracy: .7720)** over any other single classifier model.

5.3 Test Set

After the models were trained, their performance was measured on a separate set of 860 unseen tweets. All F_1-scores provided by the OffensEval organising team were macro-averaged. Baselines for each metric were also provided.

Subtask A: The best single model, Linear SVC came in at .7369 F_1-score and .8012 accuracy, while the ensemble model achieved a slightly improved .7434 F_1 and .8023 accuracy, as highlighted in Table 2. Most models used in the ensemble exhibited similar F_1 and accuracy, but recall of offensive tweets varying in the 0.4–0.7 range,

with models with high offensive recall exhibiting equivalent decrease in recall of non-offensive tweets. On the unseen test data, the ensemble model reached a .5792 recall on offensive tweets and .8887 on non-offensive tweets.

Subtask B: This subtask represented a highly imbalanced dataset, with the number of targeted instances (213) dwarfing the number of untargeted instances (27). Here the ensemble model performed the best by far, while the different individual models exhibited high disparity on separation into the two classes. Though the ensemble at .7079 exhibited the highest F_1-score, its accuracy still trailed behind the baseline targeted (TIN) accuracy by a small margin (.8833 vs .8875). The recall of the targeted and untargeted (UNT) tweets were .9343 and .4815, respectively.

Subtask C: Subtask C entailed multi-class classification over the target type of insult. This was the only subtask which exhibited improvement through the inclusion of POS data. As seen in Table 2, the ensemble model achieved .4854 F_1 and .6291 accuracy. The LSTM network provided better results, coming in at .5056 F_1-score and .6385 accuracy, when using a 200-dimensional GloVe embedding. In this subtask, as expected, classification of the minority class, OTH, proved to be the most troublesome. Both the ensemble model and the LSTM exhibited very low recall on that class: .0571 and .0857, respectively. The recall of the IND and GRP classes were .7800 and .6923, respectively.

6 Error Analysis

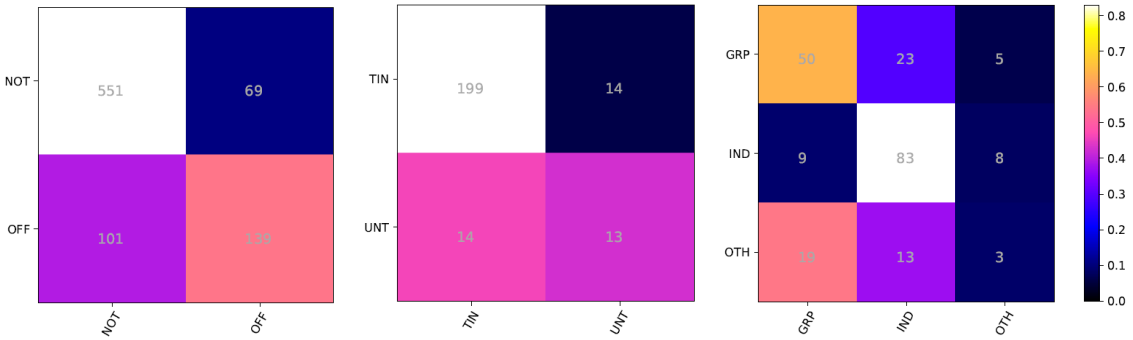
This section gives a short qualitative analysis of the misclassifications in each subtask and hypothesises potential reasons for the errors.

Subtask A: As seen in Figure 2a, the ensemble model had more difficulty identifying offensive tweets than the non-offensive ones. As also noted in previous work by Davidson et al. (2017) and others, we see that the classifier finds it difficult to identify offensive tweets that lack profanity such as *"@USER Get back on your peanut farm old man" (24726)* and *"@USER She is such a witch. All she needs is a broom" (49813)*. The classifier also faced issues in classifying political discourse, as it may have learned trends of words such as 'MAGA', 'Trump', 'Liberals' and 'Conservatives' being appearing relatively often under the OFF (offensive) label. This leads to misclassification of tweets such as *"@USER Up next: liberals calling us out for calling him guilty." (59807)* and *"@USER there is a point where even liberals must question motives" (15788)* as offensive.

Subtask B: Due to the highly imbalanced data set, the minority class (UNT) as expected accounted for most of the misclassifications, as seen in Figure 2b. The simple trend deduced was that tweets with pronouns such as 'she', 'your', 'he', and 'I' were biased to be classified as targeted (TIN). This leads to misclassification of untargeted insults such as, *"@USER @USER Still no excuse... Where TF are her parents??? They are using him & he is using her" (10641)* and *"@USER If someone is being too nice to you at happy hour and asking probing questions about what you do at Pub Citizen....make sure to troll them and say you're with Antifa or something." (58699)*, *"@USER I hate him im so fucking sorry" (91969)*. The opposite is also true, with targeted insults that contain no pronouns being misclassified as untargeted: *"@USER Google go to hell!" (52798), "@USER and bale is shit" (47806)*.

Subtask C: Most misclassification in this subtask occurred on the Other (OTH) label; see Figure 2c. Here most tweets labelled OTH were classified under the Group (GRP) class, due to close similarity between the two labels. Consider the following examples: *"@USER @USER Because 45% of Americans are too lazy to vote. Non-voters skew liberal. And too many liberals who do vote throw their vote away on 3rd party losers. Next question?" (82171)* and *"@USER @USER @USER @USER Connections are vital with all of the crap Twitter forces on conservatives." (89193)*. Both these tweets are classified as GRP insults, probably due to the presence of terms such as 'Americans', 'liberals', and 'conservatives' that tend to relate to groups, while actually being annotated as OTH as they address an issue rather than a group.

There were also a considerable number of misclassifications of OTH class tweets as IND. These misclassifications are justified on similar grounds. Examples include *"@USER Google go to hell!" (52798)*, and *"@USER get your shit together" (18315)*.



(a) Sub-task A: Ensemble model (b) Sub-task B: Ensemble model (c) Sub-task C: LSTM network

Figure 2: Confusion matrices (X-axis = predicted label; Y-axis = true label)

7 Conclusion

The idea of a hierarchical classification of offensive language is a step in the right direction in reducing the ambiguity existing between various similar subtasks. It is yet to be seen, however, how effective this method would be in synthesising more specific subsets of offensive language. For example, the cyberbullying subtask instances may yield either OFF, TIN or IND labels at each level of classification, but we are unaware of how effectively models developed for the OffensEval subtask perform on cyberbullying data sets. Some issues that have plagued offensive language detection — such as the problem of ambiguity and overlap between various subtasks — could effectively be solved if the idea of hierarchical classification achieves what it sets out to do.

Consistent with previous work, we find that it is difficult to classify non-offensive tweets containing profanity and offensive tweets lacking profanity. We also found that a similar issue persists with tweets that are politically motivated and valid criticism incorrectly classified as offensive and similarly, political hate incorrectly classified as non-offensive.

On the topic of selecting a classification model, it is noteworthy that even a simple and crude deep learning model such as the one used here can obtain better results than a more polished ensemble model. Except for surface level n-grams, most features are not as predictive as we would like them to be.

The data analysis showed that even though the annotators of the OLID data set were experienced with the platform, there still exist quite a few cases of erroneous classification by the annotators, just as noted for other datasets (Waseem and Hovy, 2016; Davidson et al., 2017; Nobata et al., 2016), for which amateur annotators were found unreliable.

Offensive language detection has proven to be a more layered issue than was initially expected, but with various developments in research the task seems surmountable. Future work must focus on building upon previous endeavours, to reduce the redundancy between subtasks and publications. The OffensEval shared task is a significant step forward in achieving this goal and we look forward to seeing how future research will be affected by the work that has been done here.

References

Peter Burnap and Matthew Leighton Williams. 2015. Hate speech, machine classification and statistical modelling of information flows on Twitter: Interpretation and communication for policy decision making. *Policy and Internet*, 7(2):223–242.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, pages 71–80, Amsterdam, Netherlands. IEEE.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

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.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In

Proceedings of the 11th International Conference on Web and Social Media, pages 512–516, Montréal, Québec, Canada. AAAI Press.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *The Social Mobile Web*, pages 11–17.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of Twitter abusive behavior. *CoRR*, abs/1802.00393.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In (Waseem et al., 2017a), pages 85–90.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. *CoRR*, abs/1710.07395.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 2, short papers, pages 42–47, Portland, Oregon, USA. ACL.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Raquel Justo, Thomas Corcoran, Stephanie M. Lukin, Marilyn Walker, and M. Inés Torres. 2014. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*, 69(1):124–133.

Ludmila I. Kuncheva. 2004. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, New York, New York, USA.

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 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303, Los Angeles, California, USA. ACL/SIGDIAL.

Chikashi Nobata, Joel 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*, pages 145–153, Montréal, Québec, Canada. IW3C2.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar. ACL.

Georgios Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48(12):47304742.

Anna Schmidt and Michael Wiegland. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. ACL.

Elise Fehn Unsvåg and Björn Gambäck. 2018. The effects of user features on Twitter hate speech detection. In *Proceedings of the 2nd Workshop on Abusive Language Online*, pages 75–86, Brussels, Belgium. ACL.

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

Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault, editors. 2017a. *Proceedings of the First Workshop on Abusive Language Online*. ACL, Vancouver, Canada.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017b. Understanding abuse: A typology of abusive language detection subtasks. In (Waseem et al., 2017a), pages 78–84.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the North American Chapter of the Association for Computational Linguistics, Student Research Workshop*, pages 88–93, San Diego, California, USA. ACL.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of the GermEval 2018 Workshop*, Austrian Academy of Sciences, Vienna, Austria.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399, Perth, Australia. IW3C2.

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

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, USA. ACL.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 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)*. ACL.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In *Proceedings of the 15th International Semantic Web Conference*, pages 745–760, Heraklion, Greece. Springer Verlag.