Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 198 (2022) 360–366

Procedia Computer Science

www.elsevier.com/locate/procedia

The 2nd International Workshop on Artificial Intelligence for Natural Language Processing (IA&NLP 2021)
November 1-4, 2021, Leuven, Belgium

Predicting Different Types of Subtle Toxicity in Unhealthy Online Conversations

Shlok Gilda[a,*], Luiz Giovanini[b], Mirela Silva[b], Daniela Oliveira[b]

[a] *Department of Computer and Information Science and Engineering, University of Florida, Gainesville, USA*
[b] *Department of Electrical and Computer Engineering, University of Florida, Gainesville, USA*

Abstract

This paper investigates the use of machine learning models to classify unhealthy online conversations containing one or more forms of subtler abuse, such as hostility, sarcasm, and generalization. We leveraged a public dataset of 44K online comments containing healthy and unhealthy comments labeled with seven forms of subtle toxicity. We were able to distinguish between these comments with a micro F1-score, macro F1-score, and ROC-AUC of 88.76%, 67.98%, and 0.71, respectively. Hostile comments were easier to detect than other types of unhealthy comments. We also conducted a sentiment analysis that revealed that most unhealthy comments were associated with a slight negative sentiment, with hostile comments being the most negative.

Keywords: Sentiment analysis; Artificial neural networks; Convolutional neural networks

1. Introduction

Healthy online conversations occur when posts or comments are made in good faith, are not blatantly abusive or hostile, typically focus on substance and ideas, and generally invite engagement [30]. Conversely, toxic comments are a harmful type of conversation widely found online that are insulting and violent in nature [30]. This kind of toxic conversation has been the primary focus of several previous studies (e.g., [11, 38]); however, many comments that deter people from engaging in online conversations are not necessarily outright abusive but contain subtle forms of abuse. These comments are written not only to engage people but also hurt, antagonize, or humiliate others and are thus referred to as *unhealthy conversations* [30].

Behaviors such as condescension, "benevolent" stereotyping, and microaggressions are frequently targeted to members of minority social groups [39, 12]. Nadal et al. [24] indicated that such subtle abuse can be as emotionally harmful as outright toxic to some individuals.

[*] Corresponding author.
E-mail address: shlokgilda@ufl.edu

10.1016/j.procs.2021.12.254

In this paper, we sought to answer two research questions in the context of unhealthy conversations:

- **RQ1:** What is the general sentiment associated with unhealthy conversations compared to healthy conversations?
- **RQ2:** Can we differentiate between unhealthy and healthy conversations? If so, which type of unhealthy conversation is the most detectable?

Towards these goals, we leveraged a public dataset of 44K online comments, finding that most unhealthy comments contained negative sentiment, and healthy and unhealthy comments were distinguishable from each other with micro and macro F1-scores of nearly 89% and 68%, respectively.

This paper is organized as follows. Section 2 summarizes related work. Section 3 describes the dataset we leveraged in our experiments, as well as our preprocessing procedures. Section 4 details our study's methodology. Section 5 analyzes our study's results and how they answer our research questions. Section 6 analyzes our study's limitations and proposes directions for future work. Section 7 concludes the paper.

2. Related Work

Sentiment classification of social media posts relative to toxicity has been researched extensively over the past years [5, 32, 35, 38]. Most work have primarily focused on algorithmic moderation of toxic comments, which are derogatory and threatening. The importance of community norms in detecting and classifying these subtler forms of abuse has been noted [4, 13, 22, 36], but has not received the same attention in the NLP community.

Although recognized in the larger NLP abuse typology [44], there have been only a few attempts at solving the problems associated with subtle abuse detection, such as a study on the classification of ambivalent sexism using Twitter data [17]. Detecting subtler forms of toxicity requires idiosyncratic knowledge, familiarity with the conversation context, or familiarity with the cultural tropes [2, 26]. It also requires reasoning about the implications of the propositions. Dinakar et al. [9] extract implicit assumptions in statements and use common sense reasoning to identify social norm violations that would be considered an insult. Identification of subtle indicators of unhealthy conversations in online comments is a challenging task due to three main reasons [30]: (i) comments are less extreme and thus have lesser explicit vocabulary; (ii) a remark may be perceived differently based on context or expectations of the reader; and (iii) greater risk of false positives or false negatives. Cultural diversity also plays a vital role in how a comment/remark may be perceived differently [31], thus making the identification of subtle toxicity online more challenging.

From an ML perspective, abusive comments classification research initially began with the application of combining TF-IDF with sentiment and contextual features by Yin et al. [45]. Since then, there have been many advances in the field of toxicity classification. Safi Samghabadi et al. [33] applied a linear SVM to detect invective posts on Ask.fm, a social networking site for people to ask questions. The authors also utilized additional features such as Linguistic Inquiry and Word Count (LIWC; [27]), word2vec [23], paragraph2vec [21], and topic modeling. The authors reported an F1-score of 59% and AUC-ROC of 0.785 using a specific subset of features. Yu et al. [46] proposed a word vector refinement model that could be applied to pre-trained word vectors (e.g., Word2Vec and Glove) to improve the efficiency of sentiment analysis.

3. Data Preparation

The dataset used in this study was made publicly available[1] by Price et al. [30] in October 2020. It contains 44,355 unique comments of 250 characters or less from the Globe and Mail opinion articles sampled from the Simon Fraser University Opinion and Comments Corpus dataset by Kolhatkar et al. [19]. Each comment was coded by at least three annotators with at least one of the following class labels: *antagonize, condescending, dismissive, generalization, generalization unfair, healthy, hostile,* and *sarcastic*. The comments were presented in isolation to annotators, without the surrounding context of the news article and other comments, thus possibly reducing bias.

[1] https://github.com/conversationai/unhealthy-conversations

3.1. Preprocessing

First, we removed one empty comment from the dataset and 1,106 other comments which were not assigned to any class label. All comments were then preprocessed (e.g., convert all characters to lower-case, remove HTML tags, etc.) and lemmatized for the feature extraction step. After preprocessing, three comments were deleted as they contained just numbers or special characters. Thus, the total number of comments was **43,245**. Most were assigned to a single class label ($n = 38,661$, 89.4%), while 10.6% ($n = 4,584$) were associated with two or more labels. The final distribution of the classes is as follows: *antagonize* (2,066), *condescending* (2,434), *dismissive* (1,364), *generalization* (944), *generalization unfair* (890), *healthy* (41,040), *hostile* (1,130), and *sarcastic* (1,897).

4. Methodology and Analysis

This section describes our machine learning and deep learning analyses, followed by a description of sentiment analysis of the comments.

4.1. Machine Learning Analysis

In our machine learning experiments of multi-label classification, we considered the following well-known models:

Logistic Regression: We used the Logistic Regression model with TF-IDF vectorized comment texts using only words for tokens (limited to 10K features).

Support Vector Machine (SVM): SVM focuses on a small subset of examples critical to differentiating between class members and non-class members, throwing out the remaining examples [42]. This is a crucial property when analyzing large data sets containing many ambiguous patterns. We used a linear kernel since it is robust to overfitting.

4.2. Deep Learning Analysis

For our experiment, we implemented a CNN-LSTM with pre-trained GloVe [28] word embeddings. Since the comments had a variable length ($[3, 250]$ characters), we fixed the comment length at 250 characters with zero padding. We used Tensorflow [1] to implement our CNN-LSTM deep model (architecture illustrated in Fig. 1). Furthermore, we used binary cross-entropy as loss function because it handles each class as an independent vector (instead of as an 8-dimensional vector). The input to the model was a random number of samples (represented as "?" in Fig. 1), all having a fixed length of 250 characters. Our evaluation metrics were *micro and macro F1-scores* and *AUC-ROC*. F1-score is well-suited to handle imbalanced datasets [14] (as in our case). The model was evaluated using 5-fold cross-validation with iterative stratification [37, 41], via the *IterativeStratification* method from Scikit-Mulitlearn [40], which gives a well-balanced distribution of evidence of label relations up to a given order.

4.3. Sentiment Analysis

We used NLTK VADER [16] to analyze the polarity of comments. We used VADER's *compound* value from the result for analyzing the polarity of the sentiments. For every input text, VADER normalizes the overall sentiment score to fall within -1 (very negative) and $+1$ (very positive), where scores between $(-0.05, 0.05)$ are labeled as neutral polarity.

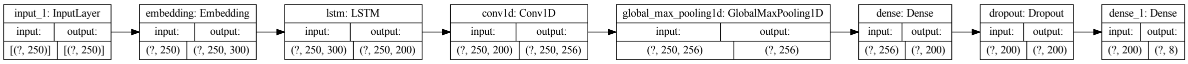

Fig. 1. CNN-LSTM network architecture.

5. Results & Discussion

In this paper, we analyzed the granularities of subtle toxic online comments. This section presents our experimental results in detecting the general sentiments associated with healthy and unhealthy online comments (**RQ1**) as well as recognizing such comments via our deep learning classifier (**RQ2**). Lastly, we summarize our main takeaways.

5.1. Results

Table 1 exhibits the average micro F1-score, macro F1-score, and ROC-AUC obtained with all tested classifiers. As can be observed, the best classification results were achieved with the CNN-LSTM model, followed by SVM and Logistic Regression.

Table 1. Classification results.

Model	Average Micro F1	Average Macro F1	AUC-ROC
Logistic Regression	57.54%	48.31%	0.51
SVM	69.15%	61.29%	0.62
CNN LSTM Network	**88.76%**	**67.98%**	**0.71**

Our CNN-LSTM model achieved an F1-micro of 88.76%, F1-macro of 67.98%, and ROC-AUC of 0.71 (Table 2). The best result was observed for the class *healthy* (AUC = 0.9524), followed by *hostile* (AUC = 0.8141) and *antagonize* (AUC = 0.7362). *Sarcasm* yielded the poorest result (AUC = 0.5707). Based on the VADER compound scores, we observed that all types of unhealthy comments except *sarcastic* and *condescending* resulted in slight negative scores. The most negative result was observed for the class *hostile*, while the most positive was *sarcastic*. *Condescending* and *healthy* produced the most neutral sentiment scores.

A similar study of unhealthy conversations [30] employed a pre-trained Bidirectional Encoder Representations from Transformers (BERT) [8] model to classify these conversations and reported a mean AUC-ROC of 0.74. Multiple studies have shown that BERT usually outperforms traditional classifiers in NLP-related tasks, but it also has a high training time. CNN and LSTM based models achieve comparable results while requiring less training time [6, 10, 18, 20]. Due to the ever-changing nature of online conversations, classifiers for unhealthy conversations would have to be updated and trained regularly with newer examples. Thus, the use of resource-intensive transformer-based models would not be appropriate in such situations.

Table 2. VADER compound score and AUC-ROC results from the CNN-LSTM model.

Class label	Vader compound score	AUC-ROC
Antagonize	−0.104498	0.7362
Condescending	−0.035128	0.6702
Dismissive	−0.081356	0.6311
Generalization	−0.085851	0.6633
Generalization Unfair	−0.091320	0.6590
Healthy	0.035745	0.9524
Hostile	−0.175975	0.8141
Sarcastic	0.101285	0.5707
Mean	**N/A**	**0.7121**

5.2. Take-Aways

RQ1: What is the general sentiment associated with unhealthy conversations compared to healthy conversations?

Our sentiment analysis revealed that none of the types of comments (i.e., classes) have extremely polarizing sentiment values (Table 2). However, all the classes except *healthy*, *sarcastic*, and *condescending* had an overall slightly negative sentiment, as expected for unhealthy conversations. Hostile comments were the most negative, likely due to the use of blatantly vulgar and vile language. This could possibly indicate that hostile comments were less subtle in their hateful content. Antagonizing, dismissive, and generalizing comments all had similar negative sentiment scores in the range of $[-0.1045, -0.0814]$. Interestingly, condescending comments scored neutral sentiment (-0.03513). Detecting patronizing and condescending language is still an open research problem because, amongst many reasons, condescension is often shrouded under “flowery words” and can itself fall into seven different categories [3]. Meanwhile, sarcastic comments were notably labeled as positive sentiment (0.1013), even higher than healthy comments (0.0357); this may be because sarcasm tends to be an ironic remark, veiled in a potentially distracting positive tone.

RQ2: Can we differentiate unhealthy and healthy conversations? If so, which type of unhealthy conversation is the most detectable one?

It was possible to differentiate between unhealthy comments with a micro F1-score, macro F1-score, and ROC-AUC of 88.76%, 67.98%, and 0.7121, respectively, using the CNN-LSTM model. Our model reported a lower average macro F1-score compared to the micro F1-score results, which is intuitive given the highly imbalanced dataset—note that macro F1-score gives the same importance to each class, i.e., this value is low for models that perform poorly on rare classes, which was the case for CNN-LSTM when analyzing unhealthy conversation classes.

Additionally, Table 2 shows that healthy comments can be differentiated from the remaining classes relatively well with generally high predictive accuracy ($AUC_{healthy} = 0.9524$), likely due to the high number of healthy samples in the dataset. In contrast, despite only 1,130 samples associated with the *hostile* class, $AUC_{hostile} = 0.8141$, was the second-highest AUC achieved by our top performer classifier. Most of the hostile comments have explicit language, which might make it easier for the classifier to recognize this type of conversation properly. The AUC for antagonizing comments was the third-highest achieved, at 0.7362, potentially because $n_{antagonize} = 2,066$ was the second-largest sample size out of the remaining unhealthy conversations categories; *antagonize* also achieved the second most negative mean sentiment score (-0.1045), further indicating that there may be characteristics of this class that facilitate detection from machine learning algorithms.

Detection of *sarcasm* was the most difficult ($AUC_{sarcasm} = 0.5707$). There have been numerous studies with similar issues with the detection of sarcasm, given the difficulty of understanding the nuances and context surrounding sarcastic texts [29, 47]. The AUC scores for condescending, dismissive, and generalizing comments were only slightly better (range: $[0.6590, 0.6702]$), again highlighting the difficulty in detecting such nuanced and subtle language.

6. Limitations & Future Work

Although our analyses yielded promising results, our dataset nonetheless targets a highly specific context (data from a single Canadian newspaper website), which likely decreases the generalizability of our results. The dataset was also notably imbalanced, with primarily *healthy* comments. In future work, we thus aim to expand our experiments on a more diverse dataset by replicating Price et al.’s [30] coding process using various comments from different news websites.

One limitation regarding our machine learning analysis is our use of a deep learning architecture (CNN-LSTM), trained using a relatively small dataset, mainly in terms of unhealthy categories of conversations. In future work, we plan to increase the number of unhealthy comments to be considered by using data augmentation techniques such as Generative Adversarial Networks (GAN), which can synthetically generate new comments. NLP data augmentation techniques could also be used to increase the dataset without requiring significant human supervision, such as simulating keyboard distance error, substituting words according to their synonyms/antonyms, replacing words with their common spelling mistakes etc. [34]. This may help increase the performance of machine learning classifiers in general. Lastly, given the challenging nature of unhealthy comments classification tasks, future work should look at specialized corpora (e.g., [43, 25]) and machine learning models trained at differentiating particular types of conversations, for example, solely sarcastic comments from non-sarcastic ones (e.g., [7, 15]).

7. Conclusion

This paper analyzed the granularities of subtle toxic online conversations. We leveraged a public dataset containing healthy and unhealthy comments labeled with seven forms of subtle toxicity: antagonize, condescending, dismissive, generalization, generalization unfair, hostile and sarcastic. Our machine learning models distinguished between these comments with a micro F1-score, macro F1-score, and AUC-ROC of 88.76%, 67.98%, and 0.71, respectively, using a CNN-LSTM network with pre-trained word embeddings. Our conclusions are two-fold: (i) hostile comments are the most negative (i.e., less subtly toxic) and detectable form of unhealthy online conversation; and (ii) most types of unhealthy comments are associated with a slight negative sentiment. Findings from this work have the potential to inform and advance future work on online moderation tools, which pave the way for safer online environments.

Acknowledgements

This work was supported by the National Science Foundation under Grant No. 2028734, by the University of Florida Seed Fund award P0175721, and by the Embry-Riddle Aeronautical University award 61632-01/PO# 262143. This material is based upon work supported by (while serving at) the National Science Foundation.

References

[1] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. Tensorflow: A system for large-scale machine learning, in: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), USENIX Association, Savannah, GA. pp. 265–283. URL: https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi.

[2] van Aken, B., Risch, J., Krestel, R., Löser, A., 2018. Challenges for toxic comment classification: An in-depth error analysis. arXiv:1809.07572 [cs] URL: http://arxiv.org/abs/1809.07572. arXiv: 1809.07572.

[3] Almendros, C.P., Anke, L.E., Schockaert, S., 2020. Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities, in: Proceedings of the 28th International Conference on Computational Linguistics, pp. 5891–5902.

[4] Blackwell, L., Dimond, J., Schoenebeck, S., Lampe, C., 2017. Classification and its consequences for online harassment: Design insights from heartmob. Proceedings of the ACM on Human-Computer Interaction 1, 24:1–24:19. doi:10.1145/3134659.

[5] Chen, Z., Qian, T., 2019. Transfer capsule network for aspect level sentiment classification, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics. pp. 547–556. URL: https://www.aclweb.org/anthology/P19-1052, doi:10.18653/v1/P19-1052.

[6] Colón-Ruiz, C., Segura-Bedmar, I., 2020. Comparing deep learning architectures for sentiment analysis on drug reviews. Journal of Biomedical Informatics 110, 103539. URL: https://www.sciencedirect.com/science/article/pii/S1532046420301672, doi:https://doi.org/10.1016/j.jbi.2020.103539.

[7] Dadu, T., Pant, K., 2020. Sarcasm detection using context separators in online discourse, in: Proceedings of the Second Workshop on Figurative Language Processing, pp. 51–55.

[8] Devlin, J., Chang, M., Lee, K., Toutanova, K., 2018. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805. URL: http://arxiv.org/abs/1810.04805, arXiv:1810.04805.

[9] Dinakar, K., Jones, B., Havasi, C., Lieberman, H., Picard, R., 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Transactions on Interactive Intelligent Systems 2, 1–30. doi:10.1145/2362394.2362400.

[10] Ezen-Can, A., 2020. A comparison of LSTM and BERT for small corpus. CoRR abs/2009.05451. URL: https://arxiv.org/abs/2009.05451, arXiv:2009.05451.

[11] Georgakopoulos, S.V., Tasoulis, S.K., Vrahatis, A.G., Plagianakos, V.P., 2018. Convolutional neural networks for toxic comment classification. arXiv:1802.09957 [cs] URL: http://arxiv.org/abs/1802.09957. arXiv: 1802.09957.

[12] Glick, P., Fiske, S.T., 2001. An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. American psychologist 56, 109–118. doi:10.1037/0003-066x.56.2.109.

[13] Guberman, J., Hemphill, L., 2017. Challenges in modifying existing scales for detecting harassment in individual tweets, in: Proceedings of 50th Annual Hawaii International Conference on System Sciences (HICSS). doi:10.24251/hicss.2017.267.

[14] Han, J., Kamber, M., Pei, J., 2011. Data mining concepts and techniques, third edition. The Morgan Kaufmann Series in Data Management Systems 5, 83–124. URL: http://www.amazon.de/Data-Mining-Concepts-Techniques-Management/dp/0123814790/ref=tmm_hrd_title_0?ie=UTF8&qid=1366039033&sr=1-1.

[15] Hazarika, D., Poria, S., Gorantla, S., Cambria, E., Zimmermann, R., Mihalcea, R., 2018. CASCADE: Contextual sarcasm detection in online discussion forums, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA. pp. 1837–1848. URL: https://www.aclweb.org/anthology/C18-1156.

[16] Hutto, C.J., Gilbert, E., 2015. Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.

[17] Jha, A., Mamidi, R., 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data, in: Proceedings of the Second Workshop on NLP and Computational Social Science, Association for Computational Linguistics. p. 7–16. URL: https://www.aclweb.org/anthology/W17-2902, doi:10.18653/v1/W17-2902.

[18] Kaliyar, R.K., Goswami, A., Narang, P., 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. Multimedia Tools and Applications 80, 11765–11788.

[19] Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., Taboada, M., 2020. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. Corpus Pragmatics 4, 155–190. doi:10.1007/s41701-019-00065-w.

[20] Korpusik, M., Liu, Z., Glass, J.R., 2019. A comparison of deep learning methods for language understanding., in: Interspeech, pp. 849–853.

[21] Le, Q., Mikolov, T., 2014. Distributed representations of sentences and documents, in: International conference on machine learning, PMLR. pp. 1188–1196.

[22] Liu, P., Guberman, J., Hemphill, L., Culotta, A., 2018. Forecasting the presence and intensity of hostility on instagram using linguistic and social features, in: Twelfth international aaai conference on web and social media.

[23] Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781 [cs] URL: http://arxiv.org/abs/1301.3781. arXiv: 1301.3781.

[24] Nadal, K.L., Griffin, K.E., Wong, Y., Hamit, S., Rasmus, M., 2014. The impact of racial microaggressions on mental health: Counseling implications for clients of color. Journal of Counseling & Development 92, 57–66. doi:https://doi.org/10.1002/j.1556-6676.2014.00130.x.

[25] Oraby, S., Harrison, V., Reed, L., Hernandez, E., Riloff, E., Walker, M., 2017. Creating and characterizing a diverse corpus of sarcasm in dialogue. arXiv preprint arXiv:1709.05404 arXiv:1709.05404.

[26] Parekh, P., Patel, H., 2017. Toxic comment tools: A case study. International Journal of Advanced Research in Computer Science 8.

[27] Pennebaker, J.W., Francis, M.E., Booth, R.J., 2001. Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates 71, 2001.

[28] Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation, in: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

[29] Poria, S., Cambria, E., Hazarika, D., Vij, P., 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. arXiv preprint arXiv:1610.08815 arXiv:1610.08815.

[30] Price, I., Gifford-Moore, J., Flemming, J., Musker, S., Roichman, M., Sylvain, G., Thain, N., Dixon, L., Sorensen, J., 2020. Six attributes of unhealthy conversation. arXiv:2010.07410 [cs] URL: http://arxiv.org/abs/2010.07410. arXiv: 2010.07410.

[31] Qiufen, Y., 2014. Understanding the impact of culture on interpretation: A relevance theoretic perspective. Intercultural Communication Studies 23.

[32] Saeed, H.H., Shahzad, K., Kamiran, F., 2018. Overlapping toxic sentiment classification using deep neural architectures, in: 2018 IEEE International Conference on Data Mining Workshops (ICDMW), p. 1361–1366. doi:10.1109/ICDMW.2018.00193.

[33] Safi Samghabadi, N., Maharjan, S., Sprague, A., Diaz-Sprague, R., Solorio, T., 2017. Detecting nastiness in social media, in: Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics. p. 63–72. URL: https://www.aclweb.org/anthology/W17-3010, doi:10.18653/v1/W17-3010.

[34] Şahin, G.G., Steedman, M., 2019. Data augmentation via dependency tree morphing for low-resource languages. arXiv preprint arXiv:1903.09460 .

[35] Saif, M.A., Medvedev, A.N., Medvedev, M.A., Atanasova, T., 2018. Classification of online toxic comments using the logistic regression and neural networks models, in: AIP conference proceedings, AIP Publishing LLC. p. 060011.

[36] Salminen, J., Almerekhi, H., Milenkovic, M., Jung, S.G., An, J., Kwak, H., Jansen, J., 2018. Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media, in: Proceedings of the International AAAI Conference on Web and Social Media.

[37] Sechidis, K., Tsoumakas, G., Vlahavas, I., 2011. On the stratification of multi-label data. Machine Learning and Knowledge Discovery in Databases , 145–158.

[38] Srivastava, S., Khurana, P., Tewari, V., 2018. Identifying aggression and toxicity in comments using capsule network, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Association for Computational Linguistics, Santa Fe, New Mexico, USA. pp. 98–105. URL: https://www.aclweb.org/anthology/W18-4412.

[39] Sue, D.W., Capodilupo, C.M., Torino, G.C., Bucceri, J.M., Holder, A.M.B., Nadal, K.L., Esquilin, M., 2007. Racial microaggressions in everyday life: Implications for clinical practice. American psychologist 62, 271–286. doi:10.1037/0003-066x.62.4.271.

[40] Szymański, P., Kajdanowicz, T., 2017. A scikit-based Python environment for performing multi-label classification. ArXiv e-prints .

[41] Szymański, P., Kajdanowicz, T., 2017. A network perspective on stratification of multi-label data, in: Torgo, L., Krawczyk, B., Branco, P., Moniz, N. (Eds.), Proceedings of the First International Workshop on Learning with Imbalanced Domains:Theory and Applications, PMLR, ECML-PKDD, Skopje, Macedonia. pp. 22–35.

[42] Vapnik, V.N., 1998. Statistical Learning Theory. Wiley-Interscience.

[43] Wang, Z., Potts, C., 2019. Talkdown: A corpus for condescension detection in context. arXiv preprint arXiv:1909.11272 arXiv:1909.11272.

[44] Waseem, Z., Davidson, T., Warmsley, D., Weber, I., 2017. Understanding abuse: A typology of abusive language detection subtasks, in: Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics. p. 78–84. URL: https://www.aclweb.org/anthology/W17-3012, doi:10.18653/v1/W17-3012.

[45] Yin, D., Xue, Z., Hong, L., Davison, B., Edwards, A., Edwards, L., 2009. Detection of harassment on web 2.0. Proceedings of the Content Analysis in the WEB 2, 1–7.

[46] Yu, L.C., Wang, J., Lai, K.R., Zhang, X., 2017. Refining word embeddings for sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics. p. 534–539. URL: https://www.aclweb.org/anthology/D17-1056, doi:10.18653/v1/D17-1056.

[47] Zhang, M., Zhang, Y., Fu, G., 2016. Tweet sarcasm detection using deep neural network, in: COLING.