**REFERENCES**

1. F. Del Vigna, A. Cimino, F. Dell’Orletta, M. Petrocchi, and

M. Tesconi, “Hate me, hate me not: Hate speech

detection on Facebook,” in Proc. 1st Italian Conf.

Cybersecurity, 2017, pp. 86–95.

2. T. Davidson, D. Warmsley, M. Macy, and

I.Weber,“Automated hate speech detection and the

problem of offensive language,” in Proc. 11th Int. AAAI

Conf. Web Social Media, 2017, pp. 512–515.

3. W. Warner and J. Hirschberg, “Detecting hate speech

on the world wide web,” in Proc. 2nd Workshop Lang.

Social Media, 2012, pp. 19–26.

4. Y. Chen, Y. Zhou, S. Zhu, and H. Xu, “Detecting offensive

language in social media to protect adolescent online

safety,” in Proc. IEEE Int. Conf. Privacy, Secur., Risk Trust

Int. Conf. Soc. Comput., 2012,pp. 71–80.

5. H. Hosseinmardi, S. A.Mattson, R. I. Rafiq, R. Han,

Q. Lv, and S. Mishra, “Analyzing labeled cyberbullying

incidents on the instagram social network,” Social Inform.,

T. Y. Liu, C. N. Scollon, andW. Zhu, Eds., 2015, pp. 49–66.

6. Z. Waseem and D. Hovy, “Hateful symbols or hateful

people? predictive features for hate speech detection

on Twitter,” in Proc. NAACL Student Res. Workshop,

2016, pp. 88–93.

7. C. VanHee et al., “Detection and fine-grained classification

of cyberbullying events,” in Proc. Int. Conf. Recent Adv.

Natural Lang. Process., 2015,pp. 672–680.

8. J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, “Learning

from bullying traces in social media,” in Proc. Conf.

North Amer. Chapter Assoc. Comput. Linguistics:

Human Lang. Technol., 2012, pp. 656–666. 9. G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose,

“Detecting offensive tweets via topical feature

discovery over a large scale Twitter corpus,” in Proc.

21st ACM Int. Conf. Inf. Knowl. Manage., 2012,

pp. 1980–1984.

10. H. Zhong et al., “Content-driven detection of

cyberbullying on the Instagram social network,” in Proc.

Int. Joint Conf. Artif. Intell., 2016, pp. 3952–3958.

11. N. Djuric, J. Zhou, R. Morris, M. Grbovic,

V. Radosavljevic, and N. Bhamidipati, “Hate speech

detection with comment embeddings,” in Proc. 24th Int.

Conf. World Wide Web, 2015, pp. 29–30.

12. P. Burnap and M. L. Williams, “Cyber hate speech on

Twitter: An application of machine classification and

statistical modeling for policy and decision making,”

Policy Internet, vol. 7, no. 2, pp. 223–242, 2015.

13. C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and

Y. Chang, “Abusive language detection in online user

content,” in Proc. 25th Int. Conf. World Wide Web, 2016,

pp. 145–153.

14. Y. Mehdad and J. Tetreault, “Do characters abuse more

than words?” in Proc. 17th Annu. Meeting Special

Interest Group Discourse Dialogue, 2016,

pp. 299–303.

15. Y. Kim, “Convolutional neural networks for sentence

classification,” in Proc. Conf. Empirical Methods

Natural Lang. Process., 2014, pp. 1746–1751.

16. W. Yin and H. Sch€utze, “Attentive convolution:

Equipping CNNs with RNN-style attention

mechanisms,” Trans. Assoc. Comput. Linguistics,

vol. 6, pp. 687–702, 2018.

17. M. S. Sorower, “A literature survey on algorithms for

multi-label learning,” Oregon State Univ., Corvallis,

vol. 18, pp. 1–25, 2010