Toxic Comment Classification using Online Georgian Forum Discussions

Nineli Lashkarashvili

e-mail: ninelilashkarashvili78@gmail.com

Department of Computer Science

Faculty of Exact and Natural Sciences

Iv. Javakhishvili Tbilisi State University

University street. N13, Tbilisi

Supervisor: Magda Tsintsadze

Candidate of Physical-Mathematical Sciences

Department of Computer Science

In the rise of technologies, textual information is transferred abundantly. Correspondingly, the analysis of written information is one of the foremost topics in researches and applications. Sentiment analysis (text classification) is a common method to analyze subjective opinions and attitudes that are conveyed via sentences or passages. This technique can be used to improve marketing strategies, make recommendations, increase customer satisfaction (González-Fierro, 2017). Apart from that, social media content is studied in order to circumvent cyber-attacks, detect harassment, toxic comments, and cyberbullying. Textual data on Twitter and Facebook was largely investigated and there was a clear shift in 2014-2016 from review to social content analysis (Mäntylä et al, 2018).

In this work, the comments extracted from Tbilisi Forum (an online platform for public discussions in Georgia) were classified using deep learning (DL) models. The data, specifically from the politics section was used. Then, 10000 comments were manually labeled as toxic (label 1) or non-toxic (label 0). Any information that could identify the user was excluded. As a result, 4639 comments were toxic and 5361 non-toxic.

In this project, attention- (Vaswani et al, 2017), and biRNN-based (Schuster & Paliwal, 1997) DL algorithms were used.

For toxic comment classification following architectures were implemented: transformer encoder only, biRNN, biLSTM, and biGRU. Per architecture, 2 separate models were developed: with fastText pre-trained Georgian word embeddings (Grave et al, 2018) and without it. In the latter one, custom-developed functions for Georgian language were used and the output was fed to the embedding layer.

Stratified 5-fold cross-validation was used to maintain the same portions of toxic and non-toxic comments in train/validation/test sets. Additionally, this method runs each model multiple times for different random partitions. The transformer model without pre-trained word embeddings demonstrated superior performance (0.869 +/- 0.009 ACC, 0.931 +/- 0.012 AUC).

References

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

González-Fierro, M. (2017). A Gentle Introduction To Text Classification And Sentiment Analysis. *Miguelgfierro.com* https://miguelgfierro.com/blog/2017/a-gentle-introduction-to-text-classification-and-sentiment-analysis/.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. arXiv preprint arXiv:1802.06893.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep Learning--based Text Classification: A Comprehensive Review. ACM Computing Surveys (CSUR), 54(3), 1-40.

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Computer Science Review, 27, 16-32.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE transactions on Signal Processing, 45(11), 2673-2681.

Tbilisi forum. forum.ge

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.