Alexis Jennings

CS 4395.001

ACL Paper Summary

Title, Author, Affiliations

This summary is on the ACL long paper, “Hate Speech Detection based on Sentiment Knowledge Sharing” by authors Xianbing Zhou, Yong Yang, Xiaochao Fan, Ge Ren, Yunfeng Song, Yufeng Diao, Liang Yang, and Hongfei Lin, in affiliation with the School of Computer Science and Technology, Xinjiang Normal University, China, and the Department of Computer Science and Technology, Dalian University of Technology, China.

Summary of the Problem Addressed

The paper presents the modern-day issue of widespread hate speech on social media and the Internet in general. Hate speech can spread virally, which leads to cyberbullying and the spread of harmful information. The inherent complexity of natural language constructs makes this a challenging issue in particular. The two main solutions released so far, one rules-based and one revolving around manual feature extraction, each have their own drawbacks that fail to semantically understand the text. Neural networks have also performed unsatisfactorily because they ignore the sentiment features of the target sentences and external sentiment resources. The team proposes their solution of a hate speech detection framework based on sentiment knowledge sharing (SKS).

Summary of Prior Work

The paper goes on to describe prior work on hate speech detection. Various machine learning-based methods based on feature engineering, deep learning-based methods, and multi-task learning methods of hate speech detection are highlighted for their considerable success. The authors conclude that sentiment features play an important role in hate speech detection, as well as deep learning models that can extract the latent semantic features of text, which can provide the most direct clues for detecting hate speech. In addition, multi-task learning can improve the performance and generalization ability of models in hate speech detection by using the correlation between the task of sentiment analysis and hate speech detection.

Unique Contributions

The SKS model is unique in that it considers both target sentence sentiment and external sentiment knowledge in detecting hate speech. It uses a multi-task learning framework to model task relationships and learn task-specific features to take advantage of shared sentiment knowledge. In the multi-head attention layer, the self-attention mechanism connects any two words in a given sentence by calculating the semantic similarity and features of each word in the sentence and other words so as to better obtain the long distance dependency. In the pooling layer, they used maximum pooling and average pooling simultaneously, as opposed to a single pooling strategy.

Evaluation of Work

The authors evaluated their results by comparing their SKS model to several strong baseline models using two datasets, SemEval2019 task5 and Davidson dataset. The accuracy of the models and the F1 scores were computed, and SKS outperforms all of them. The influence of different parts of the model were also evaluated by removing sentiment knowledge sharing and the category embedding. Similarly, they also tested when sentiment data is not used as input for the model, and it only uses category embedding. SKS performed the best in these experiments as well. The influence of gated attention was also evaluated, with SKS performing the best yet again.

Google Scholar Citations and Conclusion

Hongfei Lin had the most citations: 7,701 total, while the other authors had at least 7, as that was how many times the paper had been cited, and at most around 100 to 150. The authors’ work on this paper is becoming more and more important as the Internet and social media become more prevalent ways of how people communicate with one another. It is important that these online spaces be a positive influence on people, as they use these applications and websites every day of their lives.