Qeios

Peer Review

Review of: "Leveraging Large Language Models and Topic Modeling for Toxicity Classification"

Karin Becker[1]

1. Informatic Institute, Universidade Federal do Rio Grande do Sul, Brazil

The work examines whether partitioning a dataset using topic modeling prior to hate speech supervised model training would yield better performance. The premise of the work is exciting as we move towards accepting that the consensus premise of data annotation may be acceptable if we want to consider diversity.

However, this research requires further development to contribute to the state-of-the-art in hate speech detection.

The following issues, in my opinion, should be addressed:

a) The authors must provide more details on the dataset in general and the ones derived for training (including its partitions) to test their hypotheses.

b) The authors mention that as the number of topics increases, their meaning becomes more specific. Table I illustrates this specificity, but nevertheless, they perform all their experiments with the more general topics (k=3) due to dataset size. Have the authors verified whether the number of instances represents this problem, and at which threshold?

c) The authors did not achieve significant results with their approach. A specific topic (topic 0) improved results, but otherwise, the gains – if any – are marginal. Their method needs further work unless the authors relate this result to some positionality property.

d) The authors justify that many design decisions were guided by the existence of this particular dataset with positionality (the research, the number of topics, the performance, etc.). Nevertheless, positionality is addressed only in the annex.

doi.org/10.32388/FPNRKT

Declarations

Potential competing interests: No potential competing interests to declare.