Contribution ID: **5** Type: **Poster**

Attacks on LSTM-Based Recurrent Neural Networks for Sentiment Analysis

Thursday 24 April 2025 11:30 (30 minutes)

Recurrent Neural Networks (RNN) and their other variants such as Long-Short Term Memory (LSTM) networks have become a widely used tool for natural langague processing (NLP). Thanks to their ability to effectively capture sequential dependencies in textual data they are well-suited for determining sentiment expressed in user-generated data such as media posts or reviews. However despite their effectiveness we cannot forget about their vulnerability for intentional perturbations in the input data possibly affecting their classification capabilities. This work is aimed to analyze impact of such manipulations on LSTM and non-LSTM based networks. The analysis was conducted on custom-made RNNs trained on database containing game reviews on Twitter. Research will involve specificaly two types of input data manipulations - synonymization and token replacement with the most semantically similar vectors. This work presents the effects of experiments comparing the aforementioned networks' resilience to perturbations and the way they affect model's classification abilities. Final findings suggest that LSTM models exhibit greater resistance to subtle input changes (using synonyms) than standard RNNs but remain susceptible to more advanced attacks (vector replacement). This study highlights the importance of research in domain of neural networks' security and opens new paths for future study in this direction.

Author: Mr ZAN, Rafał (Politechnika Warszawska)

Presenter: Mr ZAN, Rafał (Politechnika Warszawska)

Session Classification: Session C (Poster)