**Abstract**

Since the textual contents on online social media are highly unstructured, informal, and often misspelled, existing research on message-level offensive language detection cannot accurately detect offensive content. Meanwhile, user-level offensiveness detection seems a more feasible approach but it is an under researched area.

Hence, automatic detection and filtering of such inappropriate language has become an important problem for improving the quality of conversations with users as well as virtual agents. In this paper, we propose a novel deep learning-based technique for automatically identifying such inappropriate language. We especially focus on solving this problem in two application scenarios (b) Users conversations in messengers. Detecting inappropriate language is challenging due to various natural language phenomenon such as spelling mistakes and variations, polysemy, contextual ambiguity and semantic variations. filtering inappropriate conversations, we use LSTM and Bi-directional LSTM (BLSTM) sequential models. The proposed models do not rely on hand-crafted features, are trained end-end as a single model, and effectively capture both local features as well as their global semantics. Evaluating C-BiLSTM, LSTM and BLSTM models on real-world search queries and conversations reveals that they significantly outperform both pattern-based and other hand-crafted feature-based baselines.