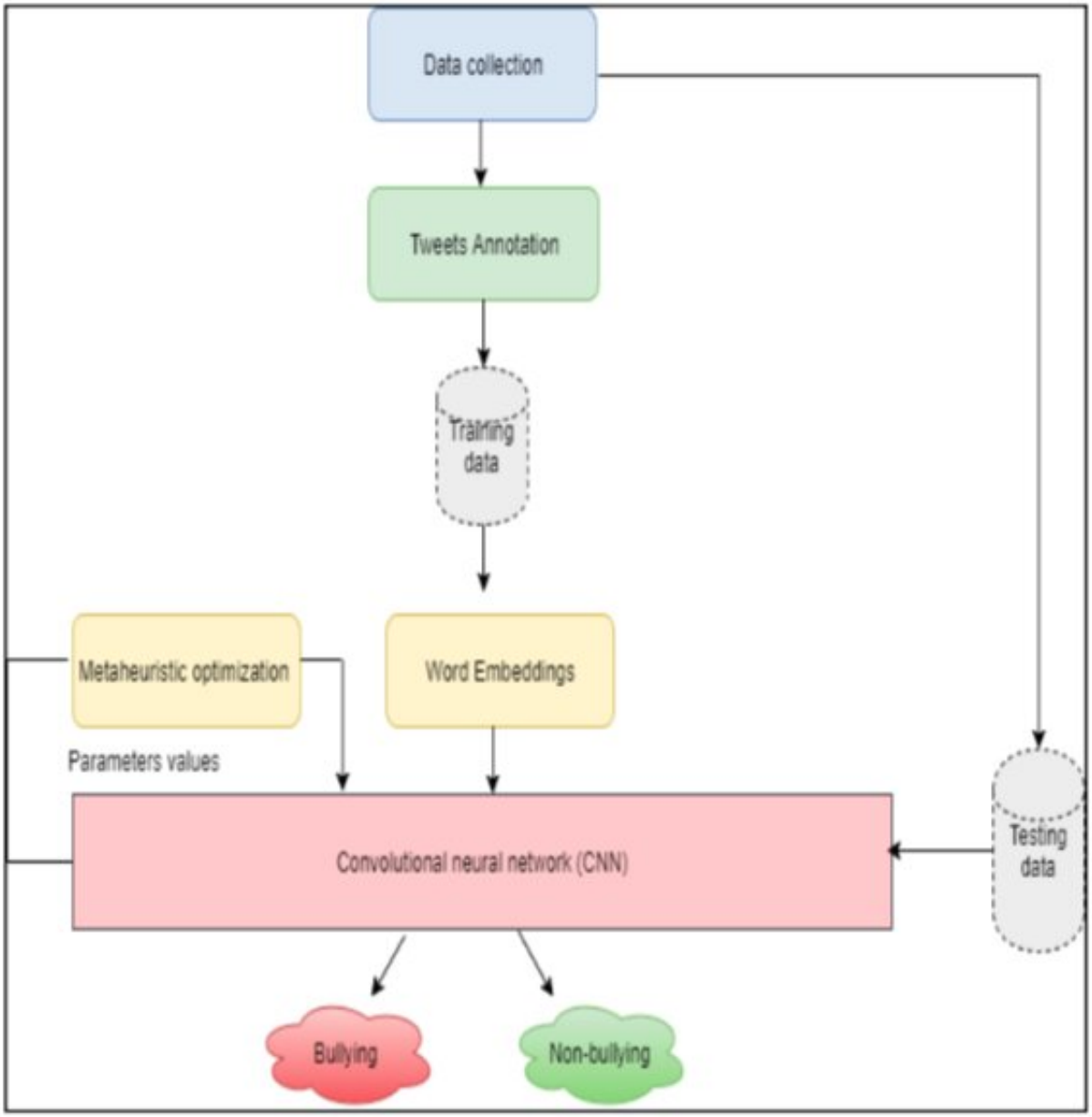
A deep learning-based method to detect cyberbullying on Twitter

January 16 2019, by Ingrid Fadelli



The system's architecture. Credit: Al-Ajlan & Ykhlef.

Researchers at King Saud University, in Saudi Arabia, have developed a new approach to detect cyberbullying on Twitter using deep learning called OCDD. In contrast with other deep-learning approaches, which extract features from tweets and feed them to a classifier, their method represents a tweet as a set of word vectors.

In recent years, cyberbullying on social media has become a huge and widely discussed issue. Cyberbullying entails the use of online communication channels to bully other users by sending intimidating, threatening or abusive messages. This can have psychological and sometimes life-threatening consequences for the victims.

Researchers worldwide have been trying to develop new ways to detect cyberbullying, manage it and reduce its prevalence on social media. Many deep learning approaches to identify cyberbullying work by analyzing textual and user features. However, these techniques come with several limitations, which can significantly reduce their performance.

For instance, some of these approaches try to improve detection by introducing new features. Yet increasing the number of features can complicate the feature extraction and selection phases. Moreover, these approaches do not consider that some user data, such as age and date of birth, can be easily fabricated. To address the limitations of existing cyberbullying detection methods, Monirah A. Al-Ajlan and Mourad Ykhlef, two researchers at King Saud University, proposed a new approach called optimized Twitter cyberbullying detection (OCDD).

"Unlike prior work in this field, OCDD does not extract features from

tweets and feed them to a classifier: Rather, it represents a tweet as a set of word vectors," the researchers explain in their paper, published on IEEE Explore and presented at the 21st Saudi Computer Society National Computer Conference (NCC). "In this way, the semantics of words are preserved, and the feature extraction and selection phases can be eliminated."

Al-Ajlan and Ykhlef built their approach on labeled training data and generated word embeddings for individual words using GloVe, an unsupervised learning algorithm that can obtain vector representations for words. These word embeddings are then fed to a convolutional neural network (CNN) to detect whether they could be associated with cyberbullying.

CNN algorithms typically consist of an input and output layer, as well as several other layers. Manually setting parameters for each of these layers can be a time-consuming and challenging task. The researchers thus decided to incorporate a metaheuristic optimization algorithm into their model, which can facilitate this process by identifying optimal or near optimal values to be used for classification.

"OCDD advances the current state of cyberbullying detection by eliminating the hard task of feature extraction/selection and replacing it with word vectors which capture the semantic of words and CNN which classifies tweets in a more intelligent way than traditional classification algorithms," the researchers write in their paper.

When tested on text mining tasks, OCDD attained very promising results. However, it is yet to be implemented and evaluated within cyberbullying detection contexts. The researchers are now planning to adapt their approach so that it can also analyze text in Arabic.

More information: Optimized Twitter cyberbullying detection based

on deep learning. DOI: 10.1109/NCG.2018.8593146.
ieeexplore.ieee.org/abstract/document/8593146