Can large language models detect sarcasm?

December 28 2023, by Ingrid Fadelli

Credit: Pixabay/CC0 Public Domain

Large language models (LLMs) are advanced deep learning algorithms that can analyze prompts in various human languages, subsequently generating realistic and exhaustive answers. This promising class of natural language processing (NLP) models has become increasingly popular after the release of Open AI's ChatGPT platform, which can rapidly answer a wide range of user queries and generate convincing

written texts for different uses.

As these models become increasingly widespread, assessing their capabilities and limitations is of utmost importance. These evaluations can ultimately help to understand the situations in which LLMs are most or least useful, while also identifying ways in which they could be improved.

Juliann Zhou, a researcher at New York University, recently carried out a study aimed at assessing the performance of two LLMs trained to detect human sarcasm, which entails conveying ideas by ironically stating the exact opposite of what one is trying to say. Her findings, posted on the preprint server *arXiv*, helped her to delineate features and algorithmic components that could enhance the sarcasm detection capabilities of both AI agents and robots.

"In the field of sentimental analysis of Natural Language Processing, the ability to correctly identify sarcasm is necessary for understanding people's true opinions," Zhou wrote in her paper. "Because the use of sarcasm is often context-based, previous research has used language representation models, such as Support Vector Machine (SVM) and Long Short-Term Memory (LSTM), to identify sarcasm with contextual-based information. Recent innovations in NLP have provided more possibilities for detecting sarcasm."

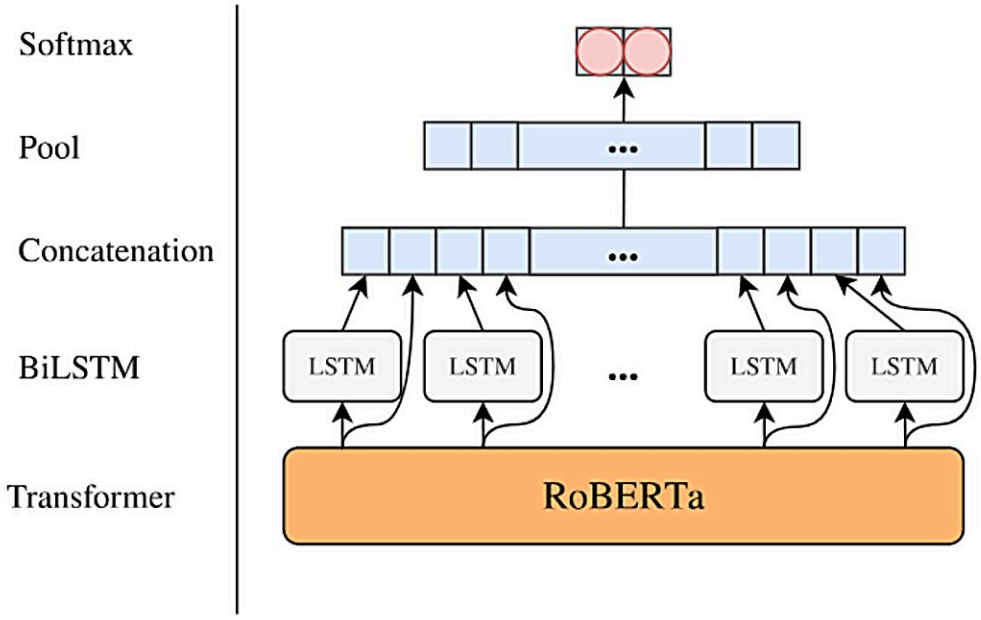

FIGURE 1. RCNN-RoBERTa transformer-based architecture proposed by Potamias et al. (2020)

Credit: Juliann Zhou.

Sentiment analysis is a field of research that entails analyzing texts typically posted on social medial platforms or other websites to gain insight on how people feel about a particular topic or product. Today, many companies are investing in this area, as it can help them to understand how they can improve their services and meet the needs of their customers.

There are now several NLP models that can process texts and predict their underlying emotional tone, or in other words if they are expressing positive, negative or neutral emotions. Many reviews and comments posted online, however, contain irony and sarcasm, which could trick

models into classifying them as "positive" when they are in fact expressing a negative emotion, or vice versa.

Some computer scientists have thus been trying to develop models that can detect sarcasm in written texts. Two of the most promising among these models, called CASCADE and RCNN-RoBERTa, were presented in 2018 by distinct research groups.

"In BERT: Pre-training of deep bidirectional transformers for language understanding, Jacob Devlin et al (2018) introduced a new language representation model and demonstrated higher precision in interpreting contextualized language," Zhou wrote. "As proposed by Hazarika et al (2018), CASCADE is a context-driven model that produces good results for detecting sarcasm. This study analyzes a Reddit corpus using these two state-of-the-art models and evaluates their performance against baseline models to find the ideal approach to sarcasm detection."

Essentially, Zhou carried out a series of tests aimed at evaluating the ability of the CASCADE and RCNN-RoBERTa model to detect sarcasm in comments posted on Reddit, the renowned online platform typically used to rate content and discuss various topics. The ability of these two models to detect sarcasm in the sample texts was also compared to the average human performance on this same task (reported in a previous work) and to the performance of a few baseline models for analyzing texts.

"We found that contextual information, such as user personality embeddings, could significantly improve performance, as well as the incorporation of a transformer RoBERTa, compared with a more traditional CNN approach," Zhou concluded in her paper. "Given the success of both contextual- and transformer-based approaches, as shown in our results, augmenting a transformer with additional contextual information features may be an avenue for future experiments."

The results gathered as part of this recent study could soon guide further studies in this area, ultimately contributing to the development of LLMs that are better at detecting sarcasm and irony in human language. These models may eventually prove to be extremely valuable tools to quickly perform sentiment analyses of online reviews, posts, and other user-generated content.

More information: Juliann Zhou, An Evaluation of State-of-the-Art Large Language Models for Sarcasm Detection, *arXiv* (2023). DOI: 10.48550/arxiv.2312.03706