WELMSD – word embedding and language model based sarcasm detection

Pradeep Kumar
IT and Systems, Indian Institute of Management Lucknow, Lucknow, India, and
Gaurav Sarin
VIPS-TC, Delhi School of Business, Delhi, India

Abstract

Purpose – Sarcasm is a sentiment in which human beings convey messages with the opposite meanings to hurt someone emotionally or condemn something in a witty manner. The difference between the text's literal and its intended meaning makes it tough to identify. Mostly, researchers and practitioners only consider explicit information for text classification; however, considering implicit with explicit information will enhance the classifier's accuracy. Several sarcasm detection studies focus on syntactic, lexical or pragmatic features that are uttered using words, emoticons and exclamation marks. Discrete models, which are utilized by many existing works, require manual features that are costly to uncover.

Design/methodology/approach – In this research, word embeddings used for feature extraction are combined with context-aware language models to provide automatic feature engineering capabilities as well superior classification performance as compared to baseline models. Performance of the proposed models has been shown on three benchmark datasets over different evaluation metrics namely misclassification rate, receiver operating characteristic (ROC) curve and area under curve (AUC).

Findings – Experimental results suggest that FastText word embedding technique with BERT language model gives higher accuracy and helps to identify the sarcastic textual element correctly.

Originality/value – Sarcasm detection is a sub-task of sentiment analysis. To help in appropriate data-driven decision-making, the sentiment of the text that gets reversed due to sarcasm needs to be detected properly. In online social environments, it is critical for businesses and individuals to detect the correct sentiment polarity. This will aid in the right selling and buying of products and/or services, leading to higher sales and better market share for businesses, and meeting the quality requirements of customers.

Keywords Sarcasm identification, Sentiment analysis, Language models, Word embeddings

Paper type Research paper

1. Introduction

Opinion analysis is a big concern that requires handling numerous natural language processing (NLP) undertakings, including subjectivity recognition, idea extraction, aspect detection, sarcasm discovery and so on. Verbal correspondence is certifiably not a trifling procedure. It suggests sharing a typical code just like having the option to surmise data past the semantic significance. A ton of open acts suggest data not linguistically communicated to have the option to unravel the entire sense: if the listener is not equipped for deriving that data, the informative procedure is deficient. As per the Oxford dictionary, sarcasm is, "The use of irony to mock or convey contempt." From another point of view, sarcasm is characterized as "a sharp, unpleasant or cutting articulation or comment; a harsh sneer or insult". As per Cambria *et al.* (2016), it is a significant concern to recognize wry, unexpected and allegorical articulations. Sarcasm, particularly, is key for estimation investigation as it can totally flip the extremity of suppositions. It is a difficult task to automatically identify figurative language for computational and linguistic communities Reyes *et al.* (2014). There are two forms of irony, namely verbal and situational. In verbal irony, an announcer wants to convey the opposite sense of what is being stated. Situational irony refers to comparison, co-existence, contradictory, untrue, deceit etc. It appears that the exemplary 'bag-of-words'