Abstract No: SI-11

Sinhala language-based social media analysis to detect fake news

W. M. S. N. P. Wijayarathna* and S. Jayalal

Department of Industrial Management, Faculty of Science, University of Kelaniya, Sri Lanka
Wijayara_im15046@stu.kln.ac.lk*

In a rapidly evolving digital age, societies rely heavily upon social media to express opinions and to share the news, publicly. With billions of users, this fast mode of information exchange takes only a few minutes to take polarized opinions, oftentimes malicious or misleading, to go viral. The objective of this research is to propose a detection technique that can be used to identify fake news published in the Sinhala language to evade public unrest. Approaches to detect fake news generally rely upon features intrinsic to either the user/source or features based on the content in the text or any hybrid set of above features. The hybrid methodology which was applied in this study, mainly focused on the verifiability of the news text content against credible sources and the credibility of the source was used to obtain the news content. Ordinary user tweets and credible sources' tweets (from 08 sources) were extracted from Twitter. The selected data set consisted of about 6000 credible sources' tweets. Then, ordinary user tweets were labelled as fake (120) and non-fake (250) using the domain knowledge about the news published in the particular month. Both types of tweets were converted into a numerical format. The text encoding was done using FastText, which derives a word as the vector summation of character *n-grams* and converts words into a 300-dimensional vector. The average of word vectors in a sentence was taken as the overall sentence numeric representation. Then, the vector representation of each user tweet was compared against credible news tweet vectors to check whether semantically similar contents appeared on credible sources within a given period. Out of the list of similarity scores obtained by each ordinary user-tweet, the maximum similarity score was used for further analysis. Moreover, a point scheme was introduced for features of a user-account by considering their contribution to the overall credibility of the user- account (e.g.: for each of the 10 followers → 1 point). The summation of the points was taken as the user-account credibility score. Then, the formula T_{val} (UC) + (1− T_{val}) TS [i.e. $T_{val} \in (0.5,1]$], where UC is the account credibility score, and TS is the text verification score was generated. Here, $T_{val} > 0$ decides the relative contributions of content verification and user-account credibility to the overall tweet's credibility assessment. In the initial implementation, for $\mathrm{T}val = 0.7$, results indicated a maximum accuracy of 70% with credibility detection of tweets, after comparison with human-annotated tags. While source credibility plays an important role in overall content's credibility, the study demonstrates that the use of the verification-based method is more effective in Sinhala fake news detection.

Keywords: Fake News, Hybrid Methodology, Social Media