Fake news language analysis and detection via a text mining approach

Matteo Farnè, Giulia Benelli

Università di Bologna – matteo.farne@unibo.it; giulia.benelli2@studio.unibo.it

Abstract

Fake news has become a growing threat to society because of the speed at which they spread and the impact they have on the shaping of people's opinions and decisions. In recent years, misinformation on the Internet and social media has found a perfect breeding ground for people to comment, elaborate and share fake news without external control. Finding a way to detect and prevent the spread of fake information has indeed become a pressing issue in the literature, to understand the characteristics of the Fake News phenomenon in the 2016 US presidential election in light of the upcoming 2024 election. In this research paper, a general statistical framework combining machine learning and natural language processing techniques is proposed. The statistical model is trained on the ISOT Fake News dataset, which contains labelled fake and real articles from 2016 to 2017 and uses Latent Dirichlet Allocation (LDA) topic-document distributions as input to various classifiers, to obtain a list of topics and their predictive impact on news manipulation. The aim of this research is also to identify a super-structure of identified topics by using cluster analysis, to derive clusters of topics that can describe the macro-subjects associated with Fake News. The experimental results confirm the effectiveness of the proposed framework, which achieves high accuracy, precision and recall rates in identifying fake news by using as predictors both LDA topic inputs and identified cluster labels for classification. This research successfully contributes to the development of a more consistent Fake news detection system by providing additional insights into this alarming phenomenon.