Biodiversity data in twitter network: Understanding the latent information by text mining approach

H.M.T.N. Bandara[1]

Abstract

The increasing popularity of social media networks significantly generated a large amount of data. Mining of these data provides insight into various topics discussed among the social networking community. The scope of application of Twitter data in the field of biodiversity research includes large-scale citizen science programs, species identifications and ecological monitoring. Present study aims to understand the latent information of twitter messages/tweets related to the biodiversity using a text mining approach. Orange™ data mining toolkit and R programming language were employed to extract and analyze the tweets related to the term biodiversity. To follow up linguistic rules, cleaning and organization of extracted tweets were achieved by *tm* R package. Filtering of common English stopwords was achieved by Natural language processing toolkit and custom stop-words list. Document term matrix was created. Processed tweets were analyzed for descriptive statistics. Topic modeling approach followed by Latent dirichlet allocation was employed to identify the abstracted topics among the tweets related to biodiversity. A total of 5798 tweets and 4216 twitter profiles were extracted. Geographical distribution of the tweets indicated that the USA, India, Canada and the United Kingdom were top countries that tweeted on biodiversity. Term frequency analysis of tweets indicated that biodiversity, conservation, species, nature and wildlife were the most frequent keywords. Loss, ecology, conservation and modeling terms were correlated with the term biodiversity. Correlated terms indicated that biodiversity loss and conservation efforts are one of the dominant topics discussed among the twitter community. Term analysis of topic modeling indicated that five biodiversity-related themes such as climate change effect on biodiversity, meat consumption effect on greenhouse emission, biodiversity action plans, US President's environmental policy and environment pollution exist among the tweets. There was a considerable discussion among the twitter community on US President's deregulation policy on fossil fuel industry. Further analysis on a large set of twitter messages may reveal additional topics related to biodiversity. Since more tweets were focused on climate change and associated greenhouse gas emissions, detailed analysis of these keywords in social media networks may also reveal attitudes of people toward climate change.

Keywords: *Biodiversity, Social media, Tweets*

---

[1] *Department of Aquaculture and Aquatic Resources Management, University College of Anuradhapura, University of Vocational Technology, Sri Lanka. Corresponding author's email: tharinducacademia@hotmail.com*