ISSN: 2473-4901
v9 n5912

Using Textual Analytics to Process Information Overload of Cyber Security Subreddits

Stephanie Omakwu
so05640@georgiasouthern.edu
Department of Information Technology
Georgia Southern University
Statesboro, GA 30460, USA

Hayden Wimmer
hwimmer@georgiasouthern.edu
Department of Information Technology
Georgia Southern University
Statesboro, GA 30460, USA

Carl M Rebman, Jr.
carlr@sandiego.edu
Knauss School of Business
Department of Supply Chain, Operations, and Information Systems
University of San Diego
San Diego, CA 92110, USA

Abstract

Increases in digitalization have made it possible to track and measure every click, every payment, every message, and almost everyone's daily thoughts. Companies are extremely interested in the robustness of this data, specifically regarding understanding the sentiment of consumers. Yet the amount of information being produced and processed is quite staggering causing information overload. As such, companies tend to fall into analysis paralysis which can result in missing important insights that could help their business. The goal of this study is to analyze and categorize the top posts on multiple hacking subreddits to determine the most discussed topics and to examine the sentiment of these posts expressed by the users. We began by scraping data, specifically the title, ID, score, comments, and URL for each top post from multiple hacking subreddit communities. We then used the Natural Language Toolkit (NLTK) to perform the data preprocessing techniques for an effective analytic process and bias-free results. The results of the testing allowed us to filter through the posts and determine whether sentiment was positive, negative, or neutral. In the case of the hacking subreddits, many of the posts were of a neutral opinion. This study aims to provide a contribution by utilizing Natural Language Processing methods Topic Modeling such as Term Frequency Inverse Document Frequency, Latent Semantic Analysis (LSA) algorithm, and Sentiment Analysis to gather and synthesize cybersecurity data.

Keywords: information overload, text analytics, sentiment analysis, LSA, term frequency, NLP

An updated and full version of this abstract may be found at https://jisar.org