Evaluation of Sampling Methods for Content Analysis of Facebook Data

Xavier Javines Bilon
School of Statistics, University of the Philippines Diliman

Jose Antonio R. Clemente
Department of Psychology, University of the Philippines Diliman

A methodological challenge for researchers performing content analysis on social media data involves deciding on a sampling procedure for obtaining content to be analyzed with least sampling error. The study used and recommended two different kinds of elementary unit—post and day—that allow probability sampling of Facebook data, regardless of whether the sampling frame of all posts within the time period of interest is obtainable. Four sampling designs for post as elementary unit and five for day as elementary unit—including three commonly used sampling options for content analysis: simple random sampling without replacement (SRSWOR), constructed week sampling, and consecutive day sampling—were employed on Facebook data mined from Mocha Uson Blog from 2010 to 2018. Estimates for parameters, such as measures of user engagement and proportions of topic-related posts, were obtained at increasing sample sizes. Sampling designs for each elementary unit were evaluated by comparing the normalized area under the coefficient of variation curve (NAUCV) over the different sample sizes. For post as elementary unit, with content type as the stratification variable, stratified random sampling (StRS) using Neyman allocation based on total user engagement is recommended (average NAUCV = 31.28%). For day as elementary unit, SRSWOR is recommended (average NAUCV = 42.31%).

Keywords: content analysis, sampling method, social media analytics

1. Introduction

The twenty-first century saw the emergence of social media platforms as the Internet and other computing technologies became more accessible to the public. These computer-mediated technologies have afforded people from across the