ABSTRACT

Text mining have gained great momentum in recent years, with user-generated content becoming widely available. One key use is comment mining, with much attention being given to sentiment analysis and opinion mining. An essential step in the process of comment mining is text pre-processing; a step in which each linguistic term is assigned with a weight that commonly increases with its appearance in the studied text, yet is offset by the frequency of the term in the domain of interest. A common practice is to use the well-known TF-IDF formula to compute these weights.

This paper reveals the bias introduced by between-participants' discourse to the study of comments in social media, and proposes an adjustment. We find that content extracted from discourse is often highly correlated, resulting in dependency structures between observations in the study, thus introducing a statistical bias. Ignoring this bias can manifest in a non-robust analysis at best and can lead to an entirely wrong conclusion at worst. We propose an adjustment to TF-IDF that accounts for this bias. We illustrate the effects of both the bias and correction with seven Facebook fan pages data, covering different domains, including news, finance, politics, sport, shopping, and entertainment.