R-Opitools – An Opinion Analytical Tool for Big Social Media Data

Monsuru Adepeju
Crime and Well-being Big Data Centre, Manchester Metropolitan University

Statement of Need

The recent advent of social media systems, such as the Twitter and Facebook, has not only heralded enormous data opportunities, but also new advances in the opinion mining of natural language texts (Liu, 2012; Tsapatsoulis & Djouvas, 2019). However, in spite of the availability of wide range of opinion mining tools (Kiomourtzis et al., 2014; Munson et al., 2019; Tai et al., 2015; Wawer, 2016), there has been a lack of capability for cross-impact analysis of multiple subjects on the opinions expressed within a text document. For instance, it may be desirable to determine whether the overall opinion score concerning a given subject A has been influenced by the presence of text records relating to another subject B within the text document. The new ‘opitools’ package is designed in R language to address the aforementioned limitation. We demonstrated the utility of ‘opitools’ in Adepeju & Jimoh (2021), in which we assess the impacts of COVID-19 pandemic (subject B) on the public opinion concerning neighbourhood policing (subject A). The outcome of the demonstration shows that ‘opitools’ can be applied to many public services for unravelling important issues (subjects) that may underlie public confidence and trust in the organizations that provide such services.

Implementation

Given a text document comprising individual record texts, such as tweets or Facebook posts, the goal is to computationally classify the sentiments expressed in each text record into positive, negative or a neutral sentiments, using a lexicon-based classification approach (Adepeju & Jimoh, 2021; Nielsen, 2011). The resulting sentiment scores across each text record are combined in order to derive the overall opinion score of the subject in question. Further, in order to assess the impacts of another inherent subject within the same document on the estimated opinion score, we employ a non-parametric randomization testing strategy (Fisher, 1935; Good, 2006). We simply ask the question, “If expected opinion scores were generated under the null hypothesis, how likely would we be to find a score higher than the observed (computed) score?” The randomization testing involves generating an expectation distribution of the observed score through random re-assignment of sentiment labels of original text records. The re-assignment is repeated a large number of time to obtain an expectation distribution. Lastly, the observed score is compared with the expectations in order to derive the statistical significance (p-value) of the impacts.

Key Functionalities

In order to perform the impact analysis using the opitools package, a user needs to utilize the opi_impact() function. The function draws from two other supporting functions, opi_
score() and opi_sim(), to compute the observed score and its expectations, respectively.

In addition to four different opinion score functions defined in the package, I provided a fun parameter in the opi_impact() function to allow a user to integrate their own user-defined function of opinion score. This is to allow opitools package to be applicable to a wide range of application domains. Lastly, the key output from the opi_impact() function includes a summary statistic describing the level of impacts that a defined subject B has exerted on the observed opinion score concerning the main subject A.

**Acknowledgment**

We gratefully acknowledge the Economic and Social Research Council (ESRC), who funded the Understanding Inequalities project (Grant Reference ES/P009301/1) through which this research was conducted.

**References**


