Challenges in developing English text analysis software for Indonesian

Howard Manns, Muhammad Iqbal & Gede Primahadi Wijaya Rajeg
Monash University, Melbourne, Australia

This talk reviews the linguistic challenges encountered in developing an English text analysis system for the Indonesian language.

The Linguistic Inquiry Word Count (LIWC) system was designed by psychologists to analyze large amounts of textual data in terms of implicit attitudes, cognitive complexity and emotion (Pennebaker, Booth & Francis 2007). The original LIWC dictionary contained 4500 English words classified according to 70 linguistic (e.g. pronouns) and psychological (e.g. emotion, cognition) categories.

Subsequent dictionaries have been created for a number of languages, including Spanish, French and Mandarin. In 2015, the Australia-Indonesia Centre (AIC) funded the creation of an Indonesian LIWC dictionary to monitor ebb and flow of relations between the two nations.

In this paper, we review some of the challenges encountered in developing this dictionary. We start by outlining some of the nuances of LIWC and why accurate classification of linguistic/psychological categories impacts on the analytical power of the tool. For instance, in Islamic fundamentalist blogs, a sudden increase in inclusive personal pronouns and prepositions and a drop in language related to causation can predict a violent attack (Pennebaker 2011).

We next discuss specific challenges posed by the Indonesian language, including translation and classification issues. For example, like the Flesch-Kinkaid readability test, English LIWC measures cognitive complexity in part by word length. However, word length vis-à-vis cognitive complexity must be approached differently in Indonesian. Further, English LIWC relies on accurate and definitive classification of words as either ‘verb’ or ‘adjective’ where in Indonesian such categories can be less distinct or more contextually bound.

This paper concludes with a discussion of how decisions made for the Indonesian LIWC dictionary were informed by careful consideration of the original English dictionary, subsequent non-English dictionaries and the idiosyncrasies of the Indonesian language.

References

Pennebaker, J., Booth, R. & Francis, M. (2007). Linguistic Inquiry and Word Count: LIWC [Computer software]. Austin, TX: LIWC.net.

Pennebaker, J. (2011). Using computer analyses to identify language style and aggressive intent: The secret life of function words. *Dynamics of Asymmetric Conflict*, 4(2): 92-102.