Speech Pattern Analysis in Social Spheres
Proposal

Daniel Padé

May 6, 2019

Abstract

We propose a method to analyze the influence of high impact individuals through social media. High impact individuals, a.k.a *influencers*, are a significant source for marketing and brand awareness across social media. Here we give an alternative metric (pending further analysis) that provides a new means of ranking the effect on such individuals' followers.

Introduction

High impact individuals, a.k.a. *influencers*, have become a significant part of new marketing strategies with the advent of social media. Their presence on micro-blogging services in particular (such as Twitter) has gained widespread recognition across many media outlets as a secondary sphere of influence. Despite this, current techniques to measure the effectiveness of an individual remain frustratingly rudimentary, mostly limited to a follower count. Such metrics are particularly flawed in the face of stale or malicious users which dilutes the information about the actual effectiveness.

Despite this, a recent wave of interest has begun into part-of-speech tagging in such social media services. Burnap & Williams (2015) and Davidson *et al.* (2017), for example, both represent recent forays into the array as a result of increasing interest in more effective spam countering techniques, building on the results of earlier attempts such as Gimpel *et al.* (2010). Others have attempted similar analysis using different techniques — Nur'aini *et al.* (2015) in particular used a method of word grouping by k-means to determine trending topics.

Project

Here we seek to combine the recent results of Suresh *et al.* (2016) and Davidson *et al.* (2017) to develop a method of categorizing subgroups of users both by varying metrics, including the following:

- Sentiment analysis, from Suresh *et al.* (2016)
- Topic Detection, from Nur'aini *et al.* (2015)
- k-means clustering of followers
- shortest-path clustering of followers

Simultaneously, we seek to analyze the vernacular and content of users' discussions through TFIDF and part of speech tagging. By comparing the mutual information between large subsets of user text, we seek to find a relationship between *user distance* (that is, a users distance from an influencer) and that influencer's effect on the user's speech.

Timeline

Due to privacy regulations, most services disallow public sharing of actual post data. Therefore it is estimated that the first two weeks will consist of data collection in order to build a sizeable subset of user posts that can be mapped to followers. Approximately one month of data analysis will follow, as the structure of the network will determine whether exact results are feasible or what approximations must be made.

Acknowledgements

The author would like to thank Allison Rogers for introducing the idea of speech analysis on social networks. The author would also like to thank Duncan Buell, whose text processing curriculum inspired much of the work. The author extends extra thanks to both for their lively discussions on this and other topics.

References

Burnap, Pete, & Williams, Matthew L. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, **7**(2), 223–242.

Davidson, Thomas, Warmsley, Dana, Macy, Michael, & Weber, Ingmar. 2017. Automated hate speech detection and the problem of offensive language. *In: Eleventh International AAAI Conference on Web and Social Media*.

Gimpel, Kevin, Schneider, Nathan, O'Connor, Brendan, Das, Dipanjan, Mills, Daniel, Eisenstein, Jacob, Heilman, Michael, Yogatama, Dani, Flanigan, Jeffrey, & Smith, Noah A. 2010. *Part-of-speech tagging for twitter: Annotation, features, and experiments*. Tech. rept. Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science.

Nur'aini, Khumaisa, Najahaty, Ibtisami, Hidayati, Lina, Murfi, Hendri, & Nurrohmah, Siti. 2015. Combination of singular value decomposition and K-means clustering methods for topic detection on Twitter. *Pages 123–128 of: 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE.

Suresh, Hima, *et al.* . 2016. An unsupervised fuzzy clustering method for twitter sentiment analysis. *Pages 80–85 of: 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*. IEEE.