Tweet Sentiment Analysis with Latent Dirichlet Allocation

Masahiro Ohmura, Kwansei Gakuin University, Sanda-shi, Japan
Koh Kakusho, Kwansei Gakuin University, Sanda-shi, Japan
Takeshi Okadome, Kwansei Gakuin University, Sanda-shi, Japan

ABSTRACT

The method proposed here analyzes the social sentiments from collected tweets that have at least 1 of 800 sentimental or emotional adjectives. By dealing with tweets posted in a half a day as an input document, the method uses Latent Dirichlet Allocation (LDA) to extract social sentiments, some of which coincide with our daily sentiments. The extracted sentiments, however, indicate lowered sensitivity to changes in time, which suggests that they are not suitable for predicting daily social or economic events. Using LDA for the representative 72 adjectives to which each of the 800 adjectives maps while preserving word frequencies permits us to obtain social sentiments that show improved sensitivity to changes in time. A regression model with autocorrelated errors in which the inputs are social sentiments obtained by analyzing the contracted adjectives predicts Dow Jones Industrial Average (DJIA) more precisely than autoregressive moving-average models.

Keywords: Latent Dirichlet Allocation, Sentiment analysis, Stock prediction, Tweets

1. INTRODUCTION

Recent research reveals that Twitter feeds permit us to capture trends in the real world in real time. In particular, progress in sentiment-tracking techniques enables us to extract indicators of public mood directly from large-scale tweets. Golder and Macy (2011) for example, identified individual-level diurnal and seasonal mood rhythms in cultures across the globe, using data from millions of public Twitter messages. They found that individuals awaken in a good mood that deteriorates as the day progresses (which is consistent with the effects of sleep and circadian rhythm) and that seasonal change in baseline positive affect varies with changes in day length. This may conceivably also be the case for the stock market. Bollen et al. (2011) investigated whether public sentiment, as expressed in large-scale collections of daily Twitter posts, can be used to predict the stock market. They used GPOMS (Google-POMS) to which they extended the Profile of Mood States Bipolar (POMS-bi) (Lorr et al., 2011) and analyzed the text content of tweets to generate a six-dimensional daily time series of public mood (“calm,” “alert,” “sure,” “vital,” “kind,” and “happy”) to provide a more detailed view of changes in the public along a variety of different mood dimensions. They found that the result-
ing public mood time series were correlated to the Dow Jones Industrial Average (DJIA) to assess their ability to predict changes in the DJIA over time. Their results indicated that the prediction accuracy of standard stock market prediction models was significantly improved when certain mood dimensions are included, but not others. In particular, variations along the public mood dimensions of “calm” and “happiness” seem to have a predictive effect, but not general happiness.

Because public mood is the aggregate of personal feelings that reflect locality, racial characteristics, the economic situation, etc.; it can be complicated. Bollen et al. (2011) used the six-dimensional factors of POMS-bi to represent public mood. The six factors do not, however, seem to be able to represent such complex public mood sufficiently because they are highly correlated and strongly dependent. In fact, our analysis reveals no correlation between the daily closing prices of the DJIA and the POSM-bi “calm” factor extracted from tweets from the previous day, although Bollen et al. (2011) reported a high correlation between them1.

This study aims to extract public mood from tweets including adjectives that express feelings such as those in Bollen et al. (2011). In this article, we describe a study that determines measures that represent public mood more appropriately by using a topic model that permits us to identify topics in documents. We propose a new method that analyzes tweets with Latent Dirichlet Allocation (LDA) using a set of a day’s or half-a-day’s worth of tweets as a document. An evaluation experiment reveals that daily public moods obtained by using our method capture social feelings on days when crucial events occur, although it extracts some mood measures that cannot be interpreted intuitively except those representing common social moods.

The evaluation experiment also shows that the measures of the method are insensitive to daily variation of the public mood. This is a disadvantage for predicting social events such as a stock market. The lexicon-reduction method, which is also described in this paper, copes with the disadvantage. That is, by focusing on adjectives that represent feelings in tweets, the method first corresponds each of the adjectives to 1 of 72 representative adjectives, and then it extracts public mood with LDA by using, as a document, a collection of the representative adjectives in daily tweets. The measures obtained by the lexicon-reduction method are more sensitive to the daily variation of public mood than those obtained by the first method. Furthermore, a time-series analysis technique together with daily public mood extracted using the lexicon-reduction (LR) method predicts the DJIA more precisely than that without the public mood.

The rest of the paper is organized as follows. Section 2 summarizes related work. Section 3 describes a method of public mood estimation from tweets using LDA. Section 4 provides a tweet-sentiment analysis method by lexicon reduction and its application to stock prediction. Section 5 gives a general discussion and Section 6 summarizes the paper. Sections 2 and 3 of this paper is partly presented in Ohmura et al. (2014a) and Section 4 is also partly presented in Ohmura et al. (2014b).

2. RELATED WORK

In addition to Golder and Macy (2011) and Bollen et al. (2011) many other researchers infer or use public mood by analyzing Twitter messages. This section summarizes the studies that estimate public mood and those that predict real-world trends by analyzing tweets.

Diakopoulos and Shamma (2010) demonstrated an analytic methodology that includes visual representations and metrics that aid in making sense of the sentiment of Twitter messages during a televised 2008 U.S. political debate. They demonstrated that the overall sentiment of the debate was negative and that tweeters tended to favor Obama over McCain. They also showed that interesting events can be detected by looking at anomalies in the pulse of the sentiment signal and that controversial
SAR: An Algorithm for Selecting a Partition Attribute in Categorical-Valued Information System Using Soft Set Theory
www.igi-global.com/article/sar-algorithm-selecting-partition-attribute/68375?camid=4v1a