Twitter Sentiment Analysis

CS365 : Artificial Intelligence
Jayant Sharma(Y9259)
Aniruddh Vyas(Y9086)

Mentor:
Dr. Amitabha Mukerjee
Department of Computer Science and Engineering

Abstract

We perform a sentiment analysis on a twitter tweet corpus collected during the period January 2009 to March 2010. Using an extended version of the Profile of Mood States (bipolar) questionnaire, we extract the public mood along six bipolar dimensions(Composed, Agreeable, Elated, Confident, Tired, Confused). The results are then compared with some major social, political and economic events during the same period. Its observed that important events have an immediate bearing on the public mood as gleaned from twitter. We also perform a Granger causality analysis on the mood series thus obtained and the Dow Jones Industrial Average (DJIA) closing values during the same period, to check if the two series are correlated and the mood series might contain predictive information about the DJIA series. We find that one of the mood series is strongly correlated with the DJIA time series.

1 Introduction

A huge number of people express themselves on the Web in a number of ways through various platforms like blogs and social networks. Of these since the launch of twitter, microblogging has become increasingly popular. Twitter allows users to publish small updates (140 character limit) to their profiles.
There are a number of polls conducted to gauge public mood; a prominent example would be election polls. These polls often require a lot of manual work and resources like time, money and the like. Is it possible to come up with an automated way of analysing public mood that is cost-effecitve and at the same time reliable? We would like to perform a sentiment analysis on twitter that does just that: give reliable indicators of public mood. The above work was initiated and published by Johan Bollen, Huina Mao and Alberto Pepe [1]. We employ a similar methodology here.
We explore the hypothesis that public mood is indicative and predictive of stock market shifts. Here we draw our methodology from Johan Bollen, Huina Mao and Xiaojun Zeng's "Twitter mood predicts the stock market" [2].

2 Method

2.1 Data and instrument

2.1.1 Data sources

1. a timeline of important political, cultural, social and economic phenomena during the time period January 2009 - March 2010

2. a collection of 5,156,047 tweets tweeted during the same period as above

2.1.2 Instrument

We employ an extended version of a well known psychometric instrument Profile of Mood States(POMS) [3]. The POMS questionnaire consists of 65 mood adjectives; the questionnaire is self scored and asks the patient to rate each mood adjective on a scale of 0-4. Each mood adjective can be mapped back to one of six mood states(tension, anger, fatigue, depression, vigor, confused), using the POMS scoring key, and hence, a six dimensional mood vector for a patient can be computed, as well as a mood disturbance score. We use the POMS bipolar instead of POMS standard, since the POMS bipolar may be used in non-clinical settings and can be administered to anyone.
In order to make POMS applicable to text corpora, we expand the POMS lexicon. The need for this can be illustrated using the following example. Say, a user wants to indicate that he's angry. Instead of using the word angry(or any of the POMS lexicon words for angry), the user may use the word furious. Using WordNet 3.0, we expand the POMS-bi lexicon from 72 adjectives to 638 adjectives. We call our expanded lexicon extended POMS-bi.
So now, if any user uses any of the words from the extended POMS-bi lexicon, the word is mapped back to its original POMS term, and from there, using the POMS scoring key, to its POMS dimension. For eg: the word furious will be mapped back to its original POMS term angry and from there to its POMS dimension: angry.

2.2 Text Procesing

2.2.1 Preprocessing

Before foraging for sentiment, each tweet is processed as follows:

(a) Separation of terms on whitespace boundaries

(b) Removal of all non-alphabetic terms(except time and date stamp for each tweet)

(c) Removal of all hyperlinks in a tweet(In [2], the authors remove tweets which have any hyperlink; we however process such tweets, but remove the hyperlink, since such a tweet may also be expressing some sentiment)

(d) Conversion to lower case of all remaining characters

(e) Removal of 658 common stop words

(f) Porter stemming of remaining words

2.2.2 Scoring

Each tweet is then scored using the extended POMS-bi lexicon. Each of the terms in the lexicon are also porter-stemmed. Any term in a tweet that matches a term in the expanded lexicon is mapped back to its respective term in the original POMS-bi lexicon and then to its respective dimension via the POMS scoring key. In this manner, a mood vector for each tweet can be computed. The resulting mood vector is then normalised to produce a unit mood vector.

$$\hat{m} = \frac{m}{||m||}$$

The mood vector for a particular day can be computed by averaging mood vectors of all the tweets submitted on that particular day.

$$m_d = \frac{\sum_{\forall t \in T_d} \hat{m}}{||T_d||}$$

The time series of aggregated daily mood vectors over a k day period maybe defined as such:

$$\theta_{m_d}[i,k] = [m_i, m_{i+1}, \ldots m_{i+k}]$$

The time series of daily mood vectors thus computed, exhibit considerable variation over time. We observe from figure 1, that the variance increases as the number of tweets decreases. This makes it difficult to analyse changes in the series over time. We need a measure of the series independent of the number of tweets submitted on a particular day.

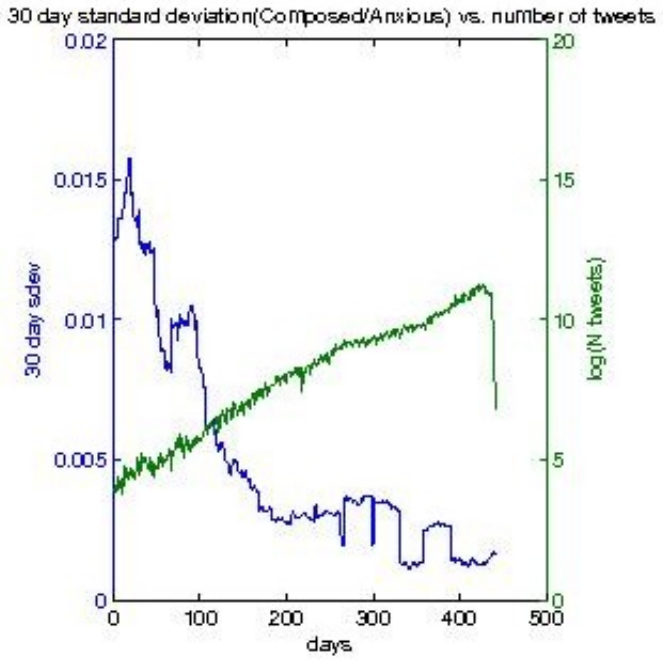


Figure 1

The mood values are thus, converted to z-values, where the z-value is defined as:

$$\tilde{m_i} = \frac{\hat{m_i} - \bar{x}(\theta[i, \pm k])}{\sigma(\theta[i, \pm k])}$$

where $\bar{x}(\theta[i, \pm k])$ and $\sigma(\theta[i, \pm k])$ represent the mean and standard deviation over the series in a k-day window.
A comparison of raw-scores and z-scores is given in figure 2. As a result, the normalized time series fluctuates around a mean of zero and its fluctuations are expressed on a common scale, namely the standard deviation regardless of the number of tweets submitted on a particular date[1].

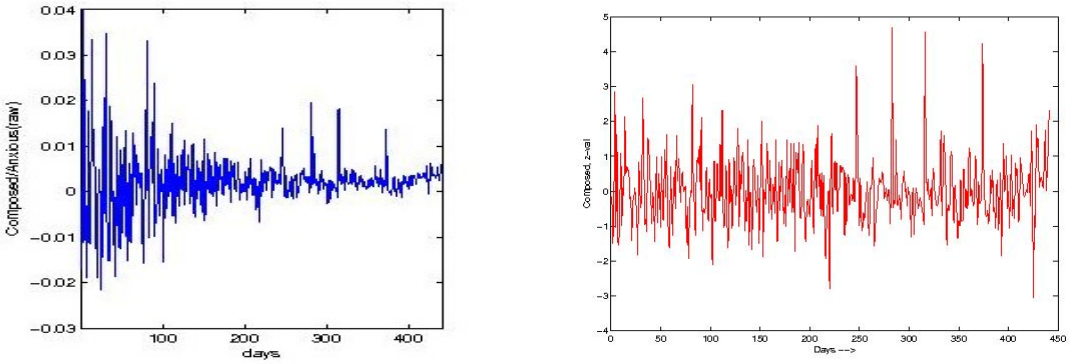


Figure 2

2.3 Results

One notices immediately the large fluctuations in the z-score graphs of the POMS mood dimensions. Since we're interested in public mood, its worth asking if some of the peaks and falls in the diagram relate with some of the major world happenings. By eyeballing the diagram closely, it is easy to make out a couple of observations and relate them with some important events in the world. We mention two events here:

- **January 15** - US Airways flight makes an emergency landing on the Hudson river, has a narrow escape. The next day, 'composed' levels jump by +2 standard deviations.
- **December 25** - The period of Christmas; almost all emotions witnesss high levels. Most notably, levels of elated and energetic(the other pole of tired) jump.

A number of events can be related to our findings in a similar manner. Detailed results may be viewed at the webpage: http://home.iitk.ac.in/<tilde>jaysha/cs365/projects/timeline.pdf.

2.4 DJIA correlation

We proceed to test whether the mood series thus obtained can be used to predict changes in the stock market. For this, DJIA closing values are retrieved for the same period as of the corpus from http://www.djaaverages.com. We apply Granger causality analysis to the two series. The Granger causality analysis rests on the assumption that if a variable X causes Y, then changes in X will systematically occur before changes in Y[2]. We might thus expect lagged values of the mood series to exhibit significant statistical correlation with values of DJIA, if indeed the mood series contains predictive information about the stock exchange. Such a correlation however does not prove causation.
The DJIA daily series is defined thus:

$$D_t = DJIA_t - DJIA_{t-1}$$

We test the following model:

$$L : D_t = \alpha + \sum_{i=1}^{n} \beta_i D_{t-i} + \sum_{i=1}^{n} \gamma_i X_{t-i} + \epsilon_t$$

, where X is a POMS mood dimension.
The results of Granger causality analysis are shown in table 1 below. Based on our results, we see that the mood dimension, 'confused' exhibits low $p-$values, and is thus significantly correlated with the DJIA time series. Over lags ranging from 1-7 days, it consistently exhibits $p-$values $<$ 0.05. Here, it must be mentioned that the authors in their paper [2], used the dimensions happy, sure, kind, calm, alert and vital, whereas we have used the original POMS-bi dimensions. The authors found the heaviest correlation with the 'calm' dimension, whereas we have found for the dimension 'confused'. A comparison as such, becomes difficult. Also, we can observe that the dimension 'confident' also seems reasonably correlated with the DJIA time series.
Another interesting observation be that values for one day lags exhibit the maximum correlation. This seems to support the dynamic nature of stock markets.

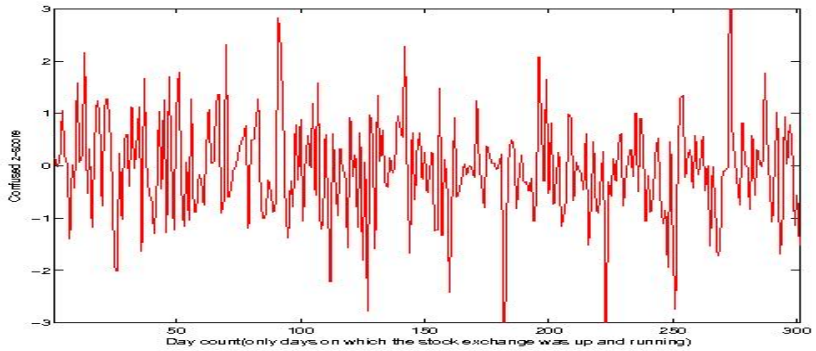
Table 1
Statistical significance(p-values) of bivariate Granger-causality correlation between moods and DJIA in period January 1, 2009 to March 14, 2010.

Lag(in days)	Composed	Agreeable	Elated	Confident	Tired	Confused
1	0.3332	0.2786	**0.0992***	**0.0153****	0.1595	**0.0019****
2	0.4744	0.3550	0.1247	**0.0538***	0.3138	**0.0089****
3	0.6675	0.2325	0.2742	0.1434	0.4141	**0.0279****
4	0.2036	0.3224	0.2688	**0.0412****	0.3424	**0.0201****
5	0.2273	0.1973	0.3326	**0.0750***	0.2096	**0.0276****
6	0.4005	0.1692	0.3164	0.1250	**0.0946***	**0.0572***
7	0.2830	0.0944	0.3336	0.1026	0.1044	**0.0241****

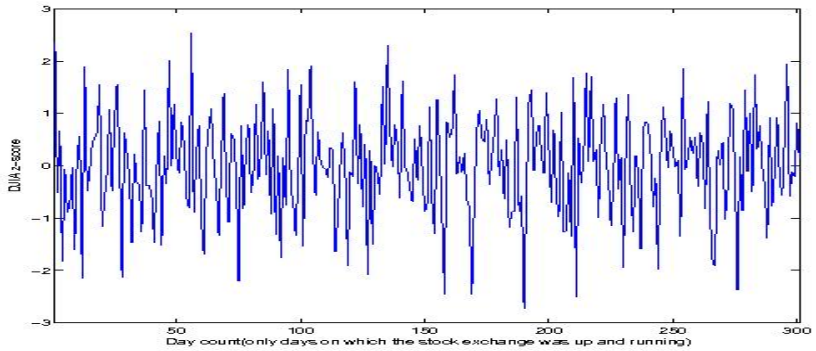
* $p < 0.1$
** $p < 0.05$

To get a better idea of the correlation, we have an accompanying graph, where the DJIA values have also been z-valued.

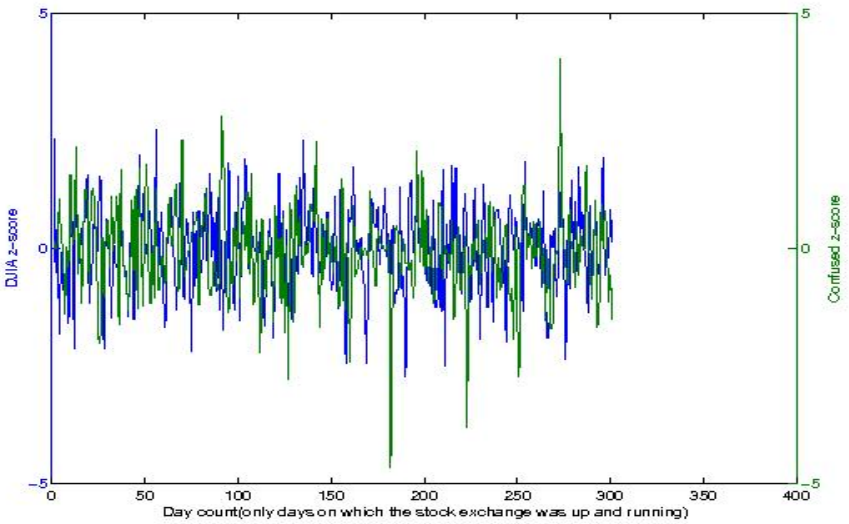
Confused z-values



DJIA z-values



A comparison of z-values of DJIA and POMS confused time-series



3 Conclusion

We are able to employ a a well-established psychometric tool to get a measure of public mood from twitter. The mood such obtained relates well with real-world events. Our work lends weight to the authors' argument that "sentiment analysis of minute text corpora(such as tweets) is efficienty obtained via a syntactic term-based approach that requires no training or machine learning"[1]. In addition we are also able to demonstrate the strong correlation between one of the POMS mood dimensions and the Dow Jones Average Index.

4 Improvement and Further Work

- Work on another method of expanding POMS, using the Web 1T n-gram database
- Employing methods to actually predict stock exchange movements, in addition to testing correlation

5 References

[1] Bollen, J., Pepe, A., & Mao, H. (2009). Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena. arXiv.org, arXiv:0911.1583v0911 [cs.CY] 0919 Nov 2009.
[2] Bollen, J.; Mao, H.; and Zeng, X.-J. 2010. Twitter mood predicts the stock market. Journal of Computational Science 2(1):18.
[3] McNair, D,; Loor, M; and Droppleman, L. 1971. Profile of Mood States
[4] Pepe, A., and Bollen, J. 2008. Between conjecture and memento: shaping a collective emotional perception of the future. In Proceedings of the AAAI Spring Symposium on Emotion, Personality, and Social Behavious.
[5] Z. Cheng, J. Caverlee, and K. Lee. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In Proceeding of the 19th ACM Conference on Information and Knowledge Management (CIKM), Toronto, Oct 2010.