Characteristics and predictability of Twitter sentiment series

A. Logunov [a], **V. Panchenko**[b]

[a]*School of Economics, University of New South Wales, Sydney, New South Wales, Australia, 2052*
[b]*School of Economics, University of New South Wales, Sydney, New South Wales, Australia, 2052*
Email: anatoly.logunov@unsw.edu.au

Abstract: In this paper we generate Twitter sentiment indices by analysing a stream of Twitter messages and categorising messages in terms of emoticons, pictorial representations of facial expressions in messages. Based on emoticons we generate daily indices. Then we explore the time-series properties of these indices by focusing on seasonal and cyclical patterns, persistence and conditional heteroscedasticity. In particular, we find significant day-of-the-week effect present in all indices, high persistence and significant degree of conditional heteroscedasticity. Then, using individual emoticon-based indices we generate an aggregate Twitter sentiment index and demonstrate that the index is good in capturing major world event such as major festive days and natural disasters. Using tests for linear and nonlinear Granger-causality we investigate whether the Twitter sentiment index contains extra information which could be used in a real-time prediction, but fail to detect any predictability at the moment. Our approach is inspired by two recent papers by Bollen et al. [2011] and O'Connor et al. [2010] who explore the relationship between Twitter sentiment and stock markets and economic indicators and find certain predictability. We significantly simplify the computational feasibility of the existing methodologies by experimenting with alternative sentiment classification processes, which may be the reason for reduced predictability of the index.

Keywords: Twitter, sentiment index, time series analysis

1 INTRODUCTION

In the present day, entire populations of networked individuals can not only connect in real-time, but have also at their fingertips the power to create and access more information than at any other time in the history of humanity. Enabled by fast mobile hardware and easy-to-use social networking technologies such as Facebook and Twitter, entire populations can now share information, organise public activities and meetings, and act together to respond to or create changes in the world around them. The 2011 popular uprisings and eventual revolution in Egypt, for example, were partly fueled by coordinated online activism and informed through social media channels such as YouTube.

Consequently, inquiry into the communication patterns of networked human populations has become an increasingly important research agenda. From investment banks to military think-tanks, organizations are increasingly exploring innovative methodologies to closely listen into, rather than hear from afar, the daily cacophony of user-generated content in order to better understand consumer behavior, predict and prepare for future trends in populations, and make better informed socio-economic decisions.

Leveraging large amounts of user-generated Twitter posts (Tweets) and pertinent research from an exciting and rapidly evolving field of information technology, we contribute by generating computationally trivial Twitter sentiment indices designed to convey the magnitude and polarity of both relative and aggregate daily emotional sentiment expressed by Tweeting individuals. The emotional sentiment of a Twitter message is extracted by means of analysing *emoticons* - popular representations of emotions and facial expressions by way of punctuation and letters. We then investigate the properties of the constructed indices.

2 DATA

A collection of Tweets was obtained by continually querying and archiving the Twitter Streaming API service, an official script which provides a sample of the newly submitted publicly-available Tweets at any time, along with information about the tweeting user and the geo-location data, when available. The streaming API service makes data available to us in the JSON data interchange format, a data object with key/value pairs representing information about the Tweet and the authoring user, with a single large text file of saved JSON objects existing for each sampled day. The actual text of the Tweet itself is represented by the value of the *text* key, for example `Yay! I just got a new job, so happy! :)`. The authoring user information includes attributes such as a user's self-reported description and location, profile picture, homepage URL and time-zone. As is often popular on Twitter, the authoring user may include a hyperlink in their message, referencing a photograph, or some further information, for the readers to access. A caveat for researchers, however, is that spam tweets generated by bots or individuals seeking traffic often also include a hyperlink. It is for this reason that similar analyses have opted to filter out tweets that contain the text "http:" or "www." (Bollen et al. [2011], Eisenstein et al. [2010]). For the purposes of our analysis, however, we have not performed any prior filtering of the day's Tweets during the pre-processing stage.

Our Twitter dataset spans a time period of 510 days from late December 2009 to late May 2011, ranging from approximately 1,000,000 tweets per day in early 2010 to over 10,000,000 archived tweets per day by May 2011. An estimated 50,000,000 global users have authored almost 4,000,000,000 total archived tweets. With the Twitter API delivering just a set proportion of all daily tweets, this result further highlights the phenomenal growth that Twitter has experienced.

2.1 Pre-processing

JSON is a popular and effective format for exchanging information between web services due to its easy compatibility with JavaScript. Unfortunately, at several terabytes of raw Tweet data, it is a difficult format to work with for efficient information extraction and query. During pre-processing, a series of scripts were executed to convert the Tweet JSON objects into the more manageable SQLite flat-file database format for efficient data mining. The aggregate daily JSON file (of approximately 15-30Gb) was converted into two daily SQLite files (of approximately 1-2Gb), one for the Tweets (with timestamp, tweet id, text and user id) and another with the details of all the users (with user id, and the other fields) which were made available that day.

3 Twitter sentiment indices: construction and analysis

Online communication, usually brief and text based, can often carry with it ambiguous or unknown interpretations of the author's mood or emotion. To solve this problem, and increase the depth and information density of short online communications, individuals often add an arrangement of letters and punctuation (*emoticons*) that have come to represent a certain emotional sentiment - usually due to the arrangement's likeness to a human face during an emotional expression, such as the happy emoticon *:-)* which bears resemblence to a rotated smiling face. Research by Yuasa et al. [2011] has shown that emoticons do indeed serve as emotional indicators similarly to other nonverbal means, activating physiological responses in the right inferior frontal gyrus region of the brain which, they report, is typically activated in emotion discrimination tasks.

Past forays into sentiment mining of Twitter streams, such as pioneering efforts by Bollen et al. [2011] and O'Connor et al. [2010], have employed the potent, albeit computationally intensive, text sentiment analysis software OpinionFinder. OpinionFinder is capable of tagging words with their contextual polarity - positive or negative [Wilson et al., 2005] and is thus suitable for processing Tweets. O'Connor et al. [2010], for example, calculate day-to-day sentiment scores out of the ratio of total positively tagged Tweets and total negatively tagged Tweets for that day.

In our classification methodology, referred to as *Naive Emoticon Classification*, we propose a computationally trivial approach for classifying a given Tweet's sentiment. If a Tweet contains an emoticon corresponding to a particular emotional sentiment, then it is simply assumed that the rational author actually felt and/or sought to convey the stated sentiment. The methodology buys its speed at the cost of foregoing analysis of any of the actual text or its contextual implications. One of the important analytical caveats to consider, however, is that without detailed contextual analysis of the content, emoticon-based methodologies such as ours could be exposed to the risks discussed by Vogel and Janssen [2009]. Specifically they remind us of the confusion and double dissociation that can sometimes arise when negative emoticons are used as a sympathetic response to an adverse situation and when positive emoticons are instead used to temper a negative text.

A plethora of different emoticons exist, with variation based on geography, cultural demographics, systematic linguistic differences (such as left-to-right and right-to-left emoticons) and differences based on sub-culture within the online communities themselves - such as the unique styles of emoticons commonly used in East Asia or Japan. Notwithstanding the fact that each of the hundreds of different emoticons can be used to convey a particular variation of sentiment, for the initial exploratory purposes of creating an aggregate sentiment index we focus only on a parsimonious expanded dichotomy of basic emotional sentiments summarised as

Emotional Sentiment	*Emoticon Proxy*
Happy	:) :-)
Sad	:(:-(
Joy (Very Happy)	:D :-D
Cry (Very Sad)	:'(

These are partly inspired by the emoticon reference guides of two of the largest instant messaging providers Windows Live Messenger and Yahoo! Messenger, who suggest *:)* or *:-)* emoticons for the *smile/happy* sentiment and *:(* or *:-(* for *sad* (Microsoft Corporation [2011] and Yahoo! [2011]). To estimate the frequency of alternative emoticon constructions, we considered a typical, randomly selected day with no major events (1 Nov 2010) and found that on this day the combined incidence of Tweets with emoticons *:)* and *:-)* makes up over 90% of the total number of Tweets with any incidence of the various known constructions of *happy* emoticons (*>:]*, *:>*, *:c)*, *:^}*, *:o)*, *:}*, *:]*, *=]*, *8)*, *:-)* and *:)*). Likewise, on the same day, the combined incidence of Tweets with emoticons *:(* and *:-(* makes up over 97% of the total number of Tweets with any incidence of the various known constructions of *sad* emoticons (*:-<*, *:<*, *:-c*, *:-[*, *:{*, *:[*, *:c*, *:-(*, *:(*). Given their relatively low frequencies of occurrence, we can omit the alternative emoticons without any major implications for our analysis.

For a given day t and emotion k, we construct two types of indices: (1) proportion of tweets containing emotion k out of the total number of all messages in this day (measured in %) and (2) share of Tweets

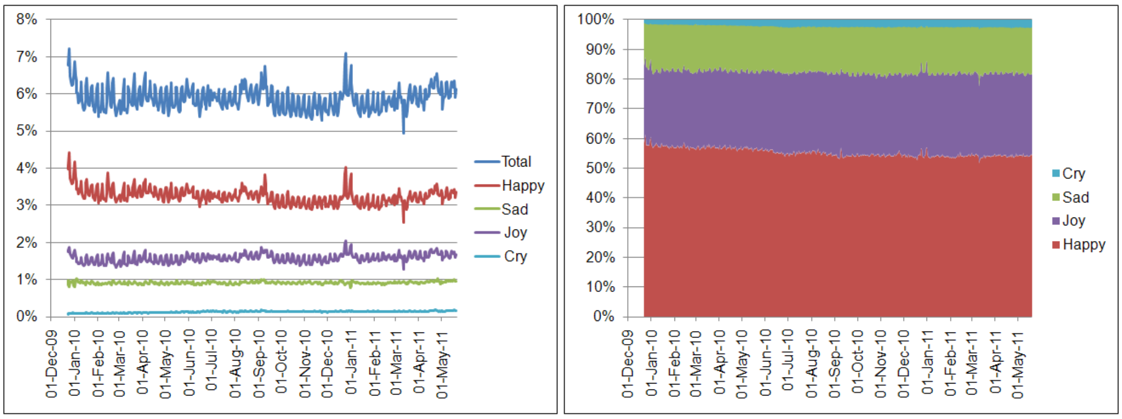

Figure 1: Twitter emotion series. Left panel shows time series of proportion (in %) of messages containing any considered emoticons (labeled Total) and specific considered emotion to the total number of all Tweets in a given day. Right panel shows the time evolution of the shares of the considered emoticons.

with emotion k out of all tweets containing any of the considered emoticons on this day (also measured in %). Figure 1 shows a time-series plot of the constructed indices. On average 5.88% of all Tweets use emoticons, out of which more than half belongs to Happy emoticons (55.25%), followed by Joy (26.89%), Sad (15.63%) and the relatively scarcely used Cry (2.23%). The proportion of Tweets using all considered emoticons is relatively stable over time, while the rarer Joy and Cry exhibit a slight rise over time replacing the traditional Happy and Sad emoticons. Our results are comparable to findings of Vogel and Janssen [2009] who analyse positive and negative emoticon use in the English posts of various online science and politics newsgroups and find the relative share of messages with positive emoticons equal to 61.23%.

Autocorrelation analysis reveals a significant seasonality pattern attributed to the day-of-the week variation. We run OLS regression on a constant and 6 dummy variables corresponding to the days of the week, holding Wednesday as a base, the results of which are given in Table 1. Relatively to Wednesday, we

Table 1: Day-of-the week effect in proportions (%) of emoticon use to the total number of messages

	Total	**Happy**	**Sad**	**Joy**	**Cry**
Monday	0.079 *	0.028	0.018 ***	0.028 *	0.005
	(1.96)	(0.93)	(3.73)	(1.94)	(1.53)
Tuesday	0.029	0.011	0.005	0.011	0.002
	(0.72)	(0.37)	(1.06)	(0.77)	(0.57)
Thursday	0.011	0.010	-0.011 **	0.013	-0.001
	(0.28)	(0.33)	(-2.24)	(0.89)	(-0.16)
Friday	0.223 ***	0.158 ***	-0.020 ***	0.081 ***	0.003
	(5.51)	(5.25)	(-4.08)	(5.61)	(0.89)
Saturday	0.456 ***	0.238 ***	0.020 ***	0.184 ***	0.014 ***
	(11.22)	(7.87)	(4.05)	(12.69)	(3.97)
Sunday	0.472 ***	0.245 ***	0.031 ***	0.182 ***	0.015 ***
	(11.63)	(8.07)	(6.35)	(12.52)	(4.47)
Intercept	5.697 ***	3.150 ***	0.911 ***	1.510 ***	0.125 ***
	(199.73)	(148.03)	(268.75)	(148.12)	(52.31)
Adjusted R^2	37.38%	23.43%	23.62%	41.58%	6.61%

***, **, * significant at a 1%, 5% and 10% level, respectively. T-statistics is reported in parenthesis

Table 2: AR(1)-ARCH(1) estimates on day-of-the-week filtered proportions (%) of emoticon use relative to the total number of messages

	Happy	**Sad**	**Joy**	**Cry**
AR				
c	-0.0226	-0.0004	-0.013*	-0.0028*
	(-1.16)	(-0.18)	(-1.79)	(-1.91)
ρ	0.8081***	0.5837***	0.74***	0.8725***
	(36.18)	(18.66)	(41.64)	(43.74)
ARCH				
γ	0.0064***	0.0004***	0.0018***	.00003***
	(35.42)	(14.54)	(32.34)	(14.49)
α	0.697***	0.2575***	0.445***	0.3357***
	(36.18)	(4.63)	(6.57)	(4.83)
Log likel.	451.24	1205.29	801	1875.65

***, **, * significant at a 1%, 5% and 10% level, respectively. T-statistics is reported in parenthesis

observe significant increase in emoticon use on the weekend, Friday and to a lesser degree on Monday. The weekend increase is significant for all emoticon categories, whilst the Friday increase is attributed to the rise of Happy and Joy emoticon use, but decrease in Sad emoticons. Monday, on the other hand, is dominated by the use of the Sad emoticon. The relatively high adjusted R^2 indicates a reasonable fit for all indices except for Sad emoticons.

To obtain a deeper intuition, we also look into the day-of-the week pattern of the share of a specific emoticon use relative to the total number of messages containing emoticons. Interestingly, we find that the share of Happy emoticon messages out of all messages with emoticons is stable over the week with a mild increase on Friday, while the relative share of Joy emotions show large increases on the weekend and a lower, but significant increase on Friday. The share of Sad emoticons exhibits a significant decrease on the weekend and Friday, and a slightly lower decrease on Thursday relatively to Wednesday and the beginning of the week, while the share of Sad emotions is a bit elevated (although not significantly) on Sunday.

After removing the day-of-the week effect we inspected a filtered series using autocorrelation and partial autocorrelation functions and found that there remains a significant degree of persistence. An ADF test rejected the null of unit root for all series. The series exhibited behaviour consistent with an AR(1) model. Moreover, some conditional heteroscedasticity was discovered after we controlled for the autocorrelation. Overall, we found that the behaviour of the emoticon time-series after controlling for the day of the week effect can be well represented by an AR(1)-ARCH(1) model:

$$\begin{aligned} I_t &= c + \rho I_{t-1} + \varepsilon_t,\, \varepsilon_t \sim N(0, \sigma_t) \\ \sigma_t &= \gamma + \alpha \varepsilon_{t-1}. \end{aligned} \tag{1}$$

The results of the AR(1)-ARCH(1) model, run on the day-of-the-week filtered indices of the proportions of specific emoticon use relative to the total number of Tweets, are given in Table 2. All emoticon indices are persistent with the highest first-order correlation coefficient attributed to Cry index and the lowest first-order correlation of Sad index. We also observe relatively high conditional heteroscedasticity for all emoticon indices with the highest level observed for the Happy index. High conditional heteroscedasticity indicates that periods of large swings in emoticon indices tend to cluster. Outbursts of happiness are typically observed around holidays and high swings may be attributed to the hangover or back to reality effect. Moreover some holidays also tend to cluster.

The evidence of clustering and persistence may be attributed to a *snowball* effect in the properties of the sentiment network. That is, the mood of an individual may be affected by the content of a Tweet they

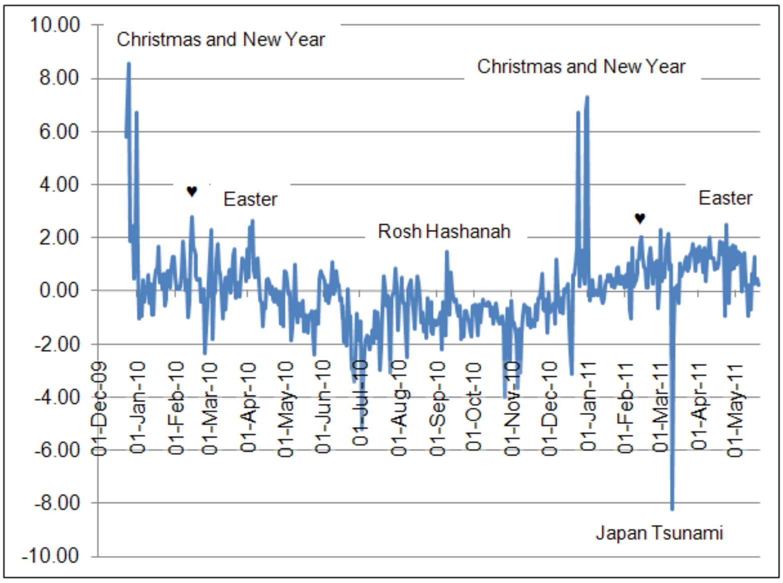

Figure 2: Time-series of Aggregate Twitter Sentiment index.

read and then, with some delay, may effect the Tweets this person then sends. Investigating the extent and magnitude of sentiment propagation in the Twitter network, and thus discerning between an author introducing a new, original sentiment reaction into the network or just acting based on a sentiment they have recently perceived, is not an easy task. However, the size and structure of our dataset which we are using does not readily lend itself to such large networked graph analysis. An effective dataset would collect all the Tweets of a large range of users, their followers, and the users they follow, to have enough information to track the behaviour of subsequent user activity in consideration of all the prior activity of the users they follow. Our dataset is a random sample of a large selection of messages of random users and does not include *all* their messages or all the messages of users specifically connected to them. Naveed et al. [2011], in their investigation of Twitter user's retweet behaviour, find that Tweets containing ”annoying or displeasant” content, ”exciting and intense” content, or a negative emoticon such as ”:-(” are more likely to be retweeted by other users, while the inclusion of a positive emoticon such as ”:-)” lowers the probability of a retweet.

4 Aggregate Sentiment Index

While in this paper we investigate time-series properties of emoticon-based indices the ultimate goal of this research program would be the construction of a Twitter-based aggregate index that is able to reflect overall emotional sentiment and, perhaps, can be useful in predicting other socio-economic and financial indicators. The above analysis indicates that constructing such an aggregate index is a challenging task. An aggregate index adding Happy or Joy and subtracting Sad and Cry indices will be dominated by the variations in Happy and Joy indices which may not be a desirable feature. Moreover, idiosyncracies in the day-of-week effect may also affect the properties of an aggregate index. To mitigate these two issues we construct an aggregate sentiment index by using residuals of the day-of-the-week regressions of the proportions of the emoticons messages to the total number of Tweets. All residuals are first standardised to a unit variance. For any given day t we sum the Happy and Joy residuals and subtract Sad and Cry residuals. The overall mood corresponds to positive and negative values in the index, which has mean zero by construction. A time-series plot of the index is shown in Figure 2 and includes reference to some notable events. At the moment the index is good in capturing the effect of holidays and major disasters. Preliminary linear Granger causality analysis did not reveal significant relationship between the constructed sentiment index and other socio-economic time-series (including returns on Dow Jones, consumer sentiment, etc). This is in contrast to Bollen et al. [2011] who do find some predictability with Dow

Jones index returns, but also use a more sophisticated index construction which includes OpinionFinder content analysis of Tweets. Non-parametric Granger causality showed some results in the commodity space, but the results were not conclusive and are left for further ongoing research.

In this paper we constructed various Twitter sentiment series using emoticons as proxies for Happy, Sad, Joy and Cry emotions. We found that all indices exhibit significant day-of-the-week effects. In particular, relatively more emoticons are used during Friday and the weekend relative to the other days of the week. At the same time these days see a smaller relative use of Sad emoticons and an increased use of Joy emoticons as one can reasonably expect. More detailed time-series analysis of the Twitter sentiment indices, after controlling for the day-of-the-week effect, revealed high first-order correlations and significant conditional heteroscedasticity. The high persistence can be attributed to the network spill-over effects. We also attempted to construct an index aggregating the individual emoticon indices. We have shown that the aggregate index is good in tracking important world events, including holidays and natural disasters. However, preliminary Granger causality tests applied to Dow Jones Index returns and consumer sentiment indices did not reveal any predictability. This may be due to an oversimplified construction of the index.

Future work will include extraction of messages more relevant to economic activity and attempts to more accurately deduce sentiment by using OpinionFinder. We also plan to use geo-location data for construction of location specific indices which may be better suited for predictability.

ACKNOWLEDGMENTS

The authors acknowledge Brendan O'Connor for his generous assistance with the Twitter data and the UNSW High Performance Computing center for providing the compute resources necessary for our data processing and analysis. Further research on this topic is forthcoming in A. Logunov's Honours thesis.

REFERENCES

Bollen, J., H. Mao, and X. Zeng (2011). Twtiter mood predicts the stock market. *Journal of Computational Science 2*, 1–8.

Eisenstein, J., B. O'Connor, N. A. Smith, and E. P. Xing (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, Stroudsburg, PA, USA, pp. 1277–1287. Association for Computational Linguistics.

Microsoft Corporation (2011). MSN Messenger Emoticons. `http://messenger.msn.com/Resource/Emoticons.aspx`. [Online; accessed 01 October 2011].

Naveed, N., T. Gottron, J. Kunegis, and A. C. Alhadi (2011). Bad news travel fast: A content-based analysis of interestingness on Twitter. In *WebSci '11: Proceedings of the 3rd International Conference on Web Science*.

O'Connor, B., R. Balasubramanyan, B. R. Routledge, and N. A. Smith (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pp. 122–129.

Vogel, C. and J. F. Janssen (2009). Multimodal signals: Cognitive and algorithmic issues. Chapter Emoticonsciousness, pp. 271–287. Berlin, Heidelberg: Springer-Verlag.

Wilson, T., J. Wiebe, and P. Hoffmann (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, Stroudsburg, PA, USA, pp. 347–354. Association for Computational Linguistics.

Yahoo! (2011). Yahoo! Messenger Emoticons. `http://messenger.yahoo.com/features/emoticons`. [Online; accessed 01 October 2011].

Yuasa, M., K. Saito, and N. Mukawa (2011). Brain activity when reading sentences and emoticons: an fMRI study of verbal and nonverbal communication. *Electronics and Communications in Japan 94*, 17–24.