# A multiscale theory for the dynamical evolution of sentiment in novels

Jianbo Gao
Institute of Complexity Science
and Big Data Technology
Guangxi University, Guangxi, China
Email: Email: jbgao.pmb@gmail.com

Matthew L. Jockers
Department of English
University of Nebraska-Lincoln
Lincoln, NE 68588

John Laudun
English Department
University of Louisiana at Lafayette
Lafayette LA 70504

Timothy Tangherlini
Department of Asian Languages and Cultures
UCLA, LA, CA, 90095

***Abstract*—Recent work in literary sentiment analysis has suggested that shifts in emotional valence may serve as a reliable proxy for plot movement in novels. The raw sentiment time series of a novel can now be extracted using a variety of different methods, and after extraction, filtering is commonly used to smooth the irregular sentiment time series. Using an adaptive filter, which is among the most effective in determining trends of a signal, reducing noise, and performing fractal and multifractal analysis, we show that the energy of the smoothed sentiment signals decays with the smoothing parameter as a power-law, characterized by a Hurst parameter $H$ of $1/2 < H < 1$, which signifies long-range correlations. We further show that a smoothed sentiment arc corresponds to the sentiment of fast playing mode or sentiment retained in one's memory, and that for a novel to be both captivating and rich, $H$ has to be larger than 1/2 but cannot be too close to 1.**

## I. Introduction

Affective computing and sentiment analysis are important tasks of AI, with applications ranging from automated analysis of reviews and social media for purposes of marketing and customer service, to the monitoring of political issues, among many others [1]. This stems from the assumption that emotions play an important role in communications among human beings, as well as in rational learning. While significant efforts have been made to detect sentiment [1], [2], [3], the analysis carried out thus far has largely been confined to detecting a polarity, or a mood, according to a limited set of emotions. Computational sentiment analysis thus has contributed little to a deep understanding of the theories of emotions, which are thought to involve many components, such as motivation, feeling, behavior, physiological changes, and evolution. Is the time ripe for computational sentiment analysis to go one step further to help gain some insights into any of the components for the theory of emotions? We show here that the sentiment dynamics of literature offer an excellent venue for studying the dynamical evolution of sentiment.

Recent work in literary text analysis has suggested that, shifts in sentiment can serve as a useful proxy for plot development [4], [5]. Remarkably, sentiment time series can now be consistently extracted from a novel by a few different methods [6], [7]. As the raw sentiment time series is very irregular, the raw data are typically filtered to yield a fairly smooth curve representing the macro-scale trend in the sentiment time series. Two examples are shown in Fig. 1 and 2, where the irregular blue curves are the raw sentiments, and the green and red curves are the filtered sentiments with the filter time scale parameter indicated in part (a) of the figures. Since the red curves are close to zero, the values are rescaled to the range of $[-1, 1]$ before they are plotted. This is shown as the red curves in part (b) of the figures. Since even the red curves are often not sufficiently smooth to reveal the macro-shape of the

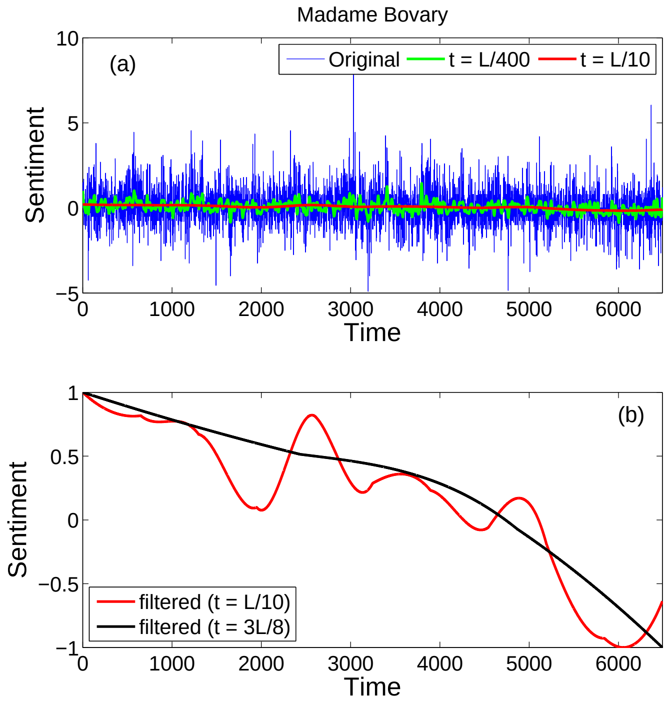

Fig. 1. Sentiment time series of Madame Bovary, where blue, red, and green are for raw, smoothed data with $w = 29$, and smoothed data with $w = 501$, respectively.

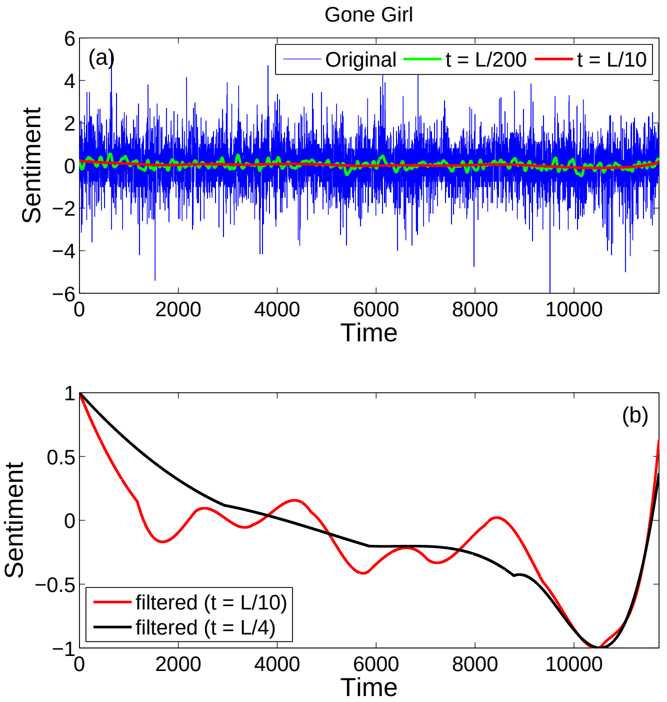

Fig. 2. Sentiment time series of Gone Girl. where blue, red, and green are for raw, smoothed data with $w = 29$, and smoothed data with $w = 501$, respectively.

series, they are either further filtered, or the raw sentiment is filtered with a time scale parameter larger than that for obtaining the red curves. Two examples are again shown as the black curves in part (b) of Fig. 1 and 2. The filter employed here is a nonlinear adaptive filter [8], [9], [10], which is among the most effective in determining trends, reducing noise, and performing fractal and multifractal analysis. The details of the filter will be presented in the Methodology section.

Human validation experiments indicate that machine derived sentiment arcs are closely correlated with sentiment arcs derived from human coded data [11], [6]. All of this prompts one to ask: what can one infer from a smooth sentiment time series about the plot development of a novel? When one seriously thinks along this line, one will realize that any smoothed sentiment time series is important; yet, no single smooth sentiment time series corresponding to a specific time scale provides sufficient, or complete, information about the plot development. This, in turn, leads one to ask a fundamental question: How does the sentiment "decay with time"? We show here that an elegant generic multiscale theory about sentiment can be developed based on random fractal theory.

One of the main models in random fractal theory is the $1/f^{\alpha}$ processes, where $\alpha = 2$ corresponds to the standard Brownian motion. Activities of many complex systems are characterized by such processes. A sub-class of such processes, denoted as $1/f^{2H+1}$, is called processes with long-range correlations (or long memories) characterized by a Hurst parameter $H$. Depending on whether $0 < H < 1/2$, $H = 1/2$, or $1/2 < H < 1$ [14], they are said to have antipersistent correlations, memoryless or only short-range correlations, or persistent long-range correlations. Prominent examples of such processes include vision [15], DNA sequences [16], [17], [18], [19], [20], human cognition [21] and coordination [22], posture [23], cardiac dynamics [24], [25], [26], [27], as well as the distribution of prime numbers [28], to name but a few. Here, we will show that sentiment time series extracted from novels always possess long-range correlations characterized by a Hurst parameter $H$ greater than 1/2.

## II. Methods

A covariance stationary stochastic process $X = \{X_t : t = 0, 1, 2, \ldots\}$, with mean $\mu$, variance $\sigma^2$, and autocorrelation function $r(k), k \geq 0$, is said to have long-range temporal correlation if the autocorrelation function $r(k)$ is of the form [12]

$$r(k) \sim k^{2H-2}, \quad as \quad k \to \infty, \tag{1}$$

where $0 < H < 1$ is the Hurst parameter. When $1/2 < H < 1$, $\sum_k r(k) = \infty$, leading to long-range temporal correlation. The process $X$ has a power-spectral density (PSD) of $1/f^{2H-1}$. Its integration, called a random walk process (and cumulative sentiment when applied to sentiment), has a PSD of $1/f^{2H+1}$. Being a $1/f$ process, it cannot be aptly modeled by a Markov process or an ARIMA model [13], since the PSD for those processes are distinctly different from $1/f$. To adequately model a $1/f$ process, a fractional order process has to be used. A well-known process of this class is the fractional Brownian motion model [14].

Since smoothing is a key issue in sentiment analysis, let us consider the effect of smoothing irregular sentiment time series $X = \{X_t : t = 0, 1, 2, \ldots\}$ by constructing a new time series

$$X^{(n)} = \{X_t^{(n)} : t = 1, 2, 3, \ldots\}, \ n = 1, 2, 3, \ldots,$$

obtained by simple nonoverlapping averaging,

$$X_t^{(n)} = (X_{tn-n+1} + \cdots + X_{tn})/n, \quad t \geq 1 \ . \tag{2}$$

For ideal fractal processes, there is an interesting scaling law for the variance of $X_t^{(n)}$ on the aggregation level $n$ [29], [30]

$$var(X^{(n)}) = \sigma^2 n^{2H-2} \tag{3}$$

where $\sigma^2$ is the variance of the original data. Eq. (3) offers an excellent means of understanding $H$. For example, if $H = 0.50$, $n = 100$, then $var(X^{(n)}) = \sigma^2/100$. When $H = 0.75$, in order to have $var(X^{(n)}) = \sigma^2/100$, then we need $n = 10^4$, which is much larger than $n = 100$ for the case of $H = 0.50$. On the other hand, when $H = 0.25$, if we still want $var(X^{(n)}) = \sigma^2/100$, then $n \approx 21.5$, which is much smaller than $n = 100$, the case of $H = 0.50$. An interesting lesson from such a simple discussion is that if a time series is short while its $H$ is close to 1, then smoothing is not a viable option for reducing the variations there.

While theoretically, moving average smoothing is a valid filter, in practice, it is among the least effective. Therefore, we need a better filter. Here we employ a nonlinear adaptive filter, which has been shown to be among the most effective in determining trends, reducing noise, and estimating the Hurst parameter [8], [9], [10]. The method works as follows. It first partitions a time series into segments (or windows) of length $w = 2n+1$ points, where neighboring segments overlap by $n+1$ points. While this has ensured symmetry, it also introduces a time scale of $\frac{w+1}{2}\tau = (n+1)\tau$, where $\tau$ is the sampling time. For each segment, we fit a best polynomial of order $M$. Note that $M = 0$ and 1 correspond to piece-wise constant and linear fitting, respectively. Denote the fitted polynomial for the $i$-th and $(i+1)$-th segments by $y^{(i)}(l_1)$, $y^{(i+1)}(l_2)$, $l_1, l_2 = 1, \cdots, 2n+1$, respectively. Note the length of the last segment may be smaller than $2n+1$. We define the fitting for the overlapped region as

$$y^{(c)}(l) = w_1 y^{(i)}(l+n) + w_2 y^{(i+1)}(l), \qquad l = 1, 2, \cdots, n+1 \tag{4}$$

where $w_1 = \left(1 - \frac{l-1}{n}\right), w_2 = \frac{l-1}{n}$ can be written as $(1 - d_j/n), j = 1, 2$, where $d_j$ denotes the distances between the point and the centers of $y^{(i)}$ and $y^{(i+1)}$, respectively. This means the weights decrease linearly with the distance between the point and the center of the segment. Such a weighting ensures symmetry and effectively eliminates any jumps or discontinuities around the boundaries of neighboring segments. In fact, the scheme ensures that the fitting is continuous everywhere, is smooth at the non-boundary points, and has the right- and left-derivatives at the boundary. The method can effectively determine any kind of trend signal. Note that with the adaptive filter described here, Eq. (3) is still valid if one equates the time scale $n$ to $(w+1)/2$ and $X_t^{(n)}$ to the trend signal, noticing that the time scale introduced by the adaptive filter with $w = 2n+1$ is $n+1$.

## III. Results and discussions

We have computed the variance of the trend signals for a wide range of time scales $t$ for the machine derived sentiment time series in 13 novels, and we have examined whether the scaling relation described by Eq. (3) holds or not. The answer is positive. Four examples are shown in Fig. 3, with $H$ indicated in each sub-plot. For all the 13 sentiment time series, $H$ is in between 1/2 and 1. Therefore, all 13 sentiment time series possess long-range correlations.

The long-range correlations in sentiment time series described by Eq. (3) give a comprehensive description on how sentiment may change with a smoothing parameter. In particular, we note that the trend signal corresponding to a time scale of $t$ amounts to the sentiment of fast playing mode with certain speed, or the memory of emotions once aroused.

In this sense, the raw, irregular sentiment corresponds to the "instantaneous" sentiment experienced by a reader on a sentence by sentence basis. Alternatively, when the sentiment data is smoothed with a larger time scale, the remaining sentiment variation reflects how the development of a novel is remembered by an "average" brain after the novel has been read. Sentiment in memory is necessarily weak, even though initially an emotion could run exceedingly high. With this understanding, it is clear how a trend signal $s(t)$ may serve as a constraint for faithful adaptation/abbreviation of a long novel. Such an interpretation serves as the starting point for a multiscale theory of sentiment.

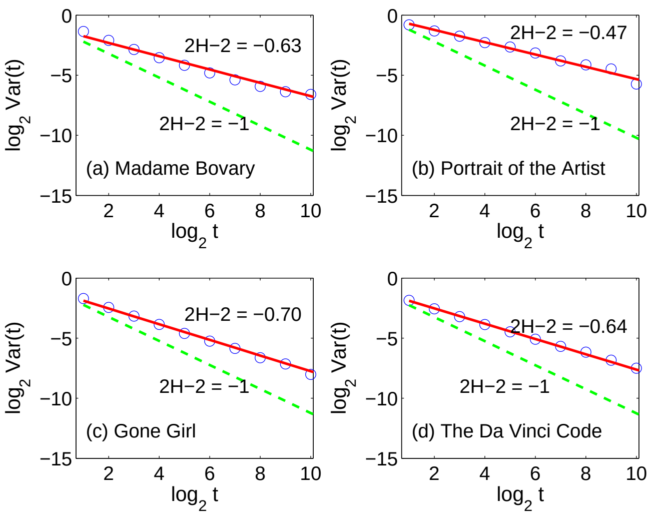

Fig. 3. Variance time analysis of four raw sentiment time series. The blue circles denote the results from the sentiment time series, the red line is the best linear least squares fitting, and the dashed green lines correspond to the case of $H = 1/2$.

We also note that although the adaptive filter employed here does not suffer from edge problems, sentiment in edges may not be resolved by $s(t)$ with large $t$, since the time scales do not match. This is a fundamental property of sampling: to resolve variations of certain frequency, the sampling frequency has to be at least twice the frequency of the variation.

To better appreciate why $1/2 < H < 1$, we emphasize that $H > 1/2$ captures the very fact that an emotion aroused by certain plot development will not instantly die out; rather, it will persist for a while, due to the memory effect. Therefore, this is a necessary condition for a novel to be captivating. On the other hand, in order to have a rich and varied plot development, $H$ cannot be close to 1.

While sentiment time series of different novels may have a comparable Hurst parameter, it is important to keep in mind that the spikes, troughs, and zeros of the smooth trend signals of sentiment are unique to each novel. That is the fundamental reason that sentiment may be considered a proxy or skeleton of unique plot development. Of course, to comprehensively characterize a novel, one needs to address many other elements, such as stylistics, thematics, character interactions and their dynamical evolution, and the interplay between what might be characterized as "action" scenes and digressions of a more philosophical or meditative nature (which are very likely connected with periods of "zero" sentiment, etc.). This is a gigantic task, and hence a vast field, rich in potential.

What is important is that the dynamical structure of sentiment is a self-similar process with long-range correlations. Moreover, the fractal long-range correlations in the sentiment evolution of plots in literature may have important implications to sentiment analysis of reviews, web contents, and public opinion monitoring. In particular, when considering the space of a large number of people expressing positive or negative sentiments, the temporal long-range correlations will bring about a well-defined polarity.

## Acknowledgment

This research was performed while the authors were visiting the Institute for Pure and Applied Mathematics (IPAM) at U-CLA, which is supported by the National Science Foundation.

## References

[1] E. Cambria, Affective computing and sentiment analysis. *IEEE Intelligent Systems* **31** (2): 102-107 (2016).

[2] E. Cambria, B. Schuller, Y.Q. Xia, C. Havasi, New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, March/April 15-21 (2013).

[3] W. Frankenstein, K. Joseph, and K. M. Carley, Contextual Sentiment Analysis, 9th International Conference on Social, Cultural, and Behavioral Modeling (SBP-BRiMS), Washington, DC, USA, June 28 July 1, 2016.

[4] Archer, J & Jockers, M. *The Bestseller Code*. New York: St. Martins Press (2016).

[5] Jockers, M. L. Syuzhet: Extracts Sentiment and Sentiment-Derived Plot Arcs from Text. https://cran.r-project.org/web/packages/syuzhet/index.htm (2016).

[6] Jockers, M. More Syuzhet Validation. http://www.matthewjockers.net/2016/08/11/more-syuzhet-validation/ (2016)

[7] Reagan, Andrew J., Mitchell, Lewis, Kiley, Dilan, Danforth, Christopher M., Dodds, & Peter Sheridan. The emotional arcs of stories are dominated by six basic shapes. arXiv: 1606.07772 (2016)

[8] Gao, J.B., Hu, J. & Tung, W.W. Facilitating joint chaos and fractal analysis of biosignals through nonlinear adaptive filtering. *PLoS ONE* **6**, e24331 (2011).

[9] Tung, W.W., Gao, J.B., Hu, J. & Yang, L. Detecting chaotic signals in heavy noise environments. *Phys. Rev. E* **83**, 046210 (2011).

[10] J.B. Gao, H. Sult an, J. Hu, and W.W. Tung, Denoising nonlinear time series by adaptive filtering and wavelet shrinkage: a comparison. *IEEE Signal Processing Letters* **17**, 237-240 (2010).

[11] Jockers, M. L. That Sentimental Feeling. http://www.matthewjockers.net/2015/12/20/that-sentimental-feeling/ (2015).

[12] Cox, D.R. in *Statistics: An Appraisal*. (eds. David, H.A. & Davis, H.T.) 55-74 (The Iowa State University Press, Ames, Iowa, 1984).

[13] Box, G.E.P. and Jenkins, G.M. *Time series analysis: forecasting and control*. 2nd ed. San Francisco: Holden-Day (1976).

[14] B.B. Mandelbrot, *The Fractal Geometry of Nature*. San Francisco: Freeman (1982).

[15] J. Gao, V. Billock, I. Merk, and et al, Inertia and memory in ambiguous visual perception, *Cognitive Processing* **7**, 105 -112 (2006).

[16] W. Li and K. Kaneko, *Europhys. Lett.* **17**, 655 (1992).

[17] R. F. Voss, *Phys. Rev. Lett.* **68** 3805 (1992).

[18] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, *Nature* **356**, 168 (1992).

[19] J. Gao, Y. Qi, Y. Cao, and W. Tung, Protein coding sequence identification by simultaneously characterizing the periodic and random features of dna sequences, *Journal of biomedicine and biotechnology* **2**, 139-146 (2005).

[20] J. Hu, J. Gao, Y. Cao, E. Bottinger, and W. Zhang, Exploiting noise in array cgh data to improve detection of dna copy number change, *Nucleic Acids Research* **35**, e35 (2007).

[21] D. L. Gilden, T. Thornton, and M. W. Mallon, *Science* **267** 1837 (1995).

[22] Y. Chen, M. Ding, and J. A. Scott Kelso, *Phys. Rev. Lett.* **79** 4501 (1997).

[23] J. J. Collins and C. J. De Luca, *Phys. Rev. Lett.* **73**, 764 (1994).

[24] P. C. Ivanov, M. G. Rosenblum, C. K. Peng, J. Mietus, S. Havlin, H. E. Stanley, and A. L. Goldberger, Scaling behaviour of heartbeat intervals obtained by wavelet-based time-series analysis, *Nature* **383**, 323 (1996).

[25] L. A. N. Amaral, A. L. Goldberger, P. C. Ivanov, and H. E. Stanley, Scale-independent measures and pathologic cardiac dynamics, *Phys. Rev. Lett.* **81**, 2388 (1998).

[26] P. C. Ivanov, M. G. Rosenblum, L. A. N. Amaral, Z. R. Struzik, S. Havlin, A. L. Goldberger, and H. E. Stanley, Multifractality in human heartbeat dynamics, *Nature* **399**, 461 (1999).

[27] P. Bernaola-Galvan, P. C. Ivanov, L. A. N. Amaral, and H. E. Stanley, Scale invariance in the nonstationarity of human heart rate, *Phys. Rev. Lett.* **87** 168105 (2001).

[28] M. Wolf, 1/f noise in the distribution of prime numbers, *Physica A* **241**, 493 (1997).

[29] Gao, J.B., Cao, Y.H., Tung, W.W. & Hu, J. *Multiscale Analysis of Complex Time Series — Integration of Chaos and Random Fractal Theory, and Beyond* (Wiley-Interscience, 2007).

[30] Gao, J.B. et al. Assessment of long range correlation in time series: How to avoid pitfalls. *Phys. Rev. E* **73**, 016117 (2006).