Sociocultural trend signatures in minimal persistence and past novelty

Kristoffer L. Nielbo[1,2], Peter Bjerregaard Vahlstrup[1,2], Jianbo Gao[3,4], Anja Bechmann[2]

[1]Center for Humanities Computing Aarhus, Aarhus University, Denmark
[2]DATALAB, Aarhus University, Denmark
[3]Center for Geodata and Analysis, Faculty of Geographical Science, Beijing Normal University, China
[4]Institute of Automation, Chinese Academy of Sciences, China

Corresponding author: Kristoffer L. Nielbo , kln@cas.au.dk

INTRODUCTION

Sociocultural trends from social media platforms such as Twitter or Instagram have become an important part of knowledge discovery. The 'trend' construct is however ambiguous and its estimation from unstructured sociocultural data complicated by several methodological issues. This paper presents an approach to trend estimation that combines ('intersects') domain knowledge of social media with advances in information theory and dynamical systems. In particular, we show how *trend reservoirs* (i.e., signals that display trend potential) can be identified by their relationship between novel and resonant behavior, and their minimal persistence.

In the typical case of trend estimation for social media, a query term (e.g., 'AI') is used to extract a signal based on the term's frequency, associated queries, and rating systems. While researchers agree that a trend has direction (e.g., an increase in AI-related posts) and tendency (e.g., "AI is the new black"), accurate estimation is a matter of debate [11]. In its simplest form, a trend's tendency is detected as a 'novelty spike' in the query's temporal distribution and the direction is estimated as the slope coefficient of the query's frequency fitted on time, e.g., [19, 17]. This *standard approach* suffers from several problematic issues: 1) by focusing on spiky behavior, it equates a sociocultural trend detection with that of natural catastrophes and epidemics; 2) it makes strong assumptions on the trend's shape; 3) it treats atomic words as semantically meaningful; and in pre-selecting query terms it 4) can fail to establish a proper baseline; and 5) reverse time order by nominating queries that show a spiky behavior in the past as future trends.

These five issues can be remedied by techniques from information theory and dynamical systems. Recent studies have shown that windowed relative entropy can generate signals that capture information *novelty* as a reliable difference from the past and *resonance* as the degree to which future information conforms to the novelty [1, 21, 23]. Several studies have used latent semantic models to summarize the data set's co-occurrence structure as an alternative to atomistic query terms [5, 25]. Regarding the trend shape, a smoothing function that fits piecewise polynomials to the data makes no assumption about the shape [11, 26]. Recently, dynamical systems approaches have indicated that adaptive functions hold great promise for smoothing sociocultural data [6, 7, 22, 15, 28]. This paper combines these insights to propose a new approach to trend estimation that studies 'trend reservoirs' which are characterized by a strong novelty-resonance association and short-range dependencies ('minimal persistence') in comparison to a random baseline.

RESULTS

We illustrate estimation of trend reservoirs on Reddit data in a single factor design that compares human annotated 'trending' subreddits with randomly selected subreddits (see Appendix A). To

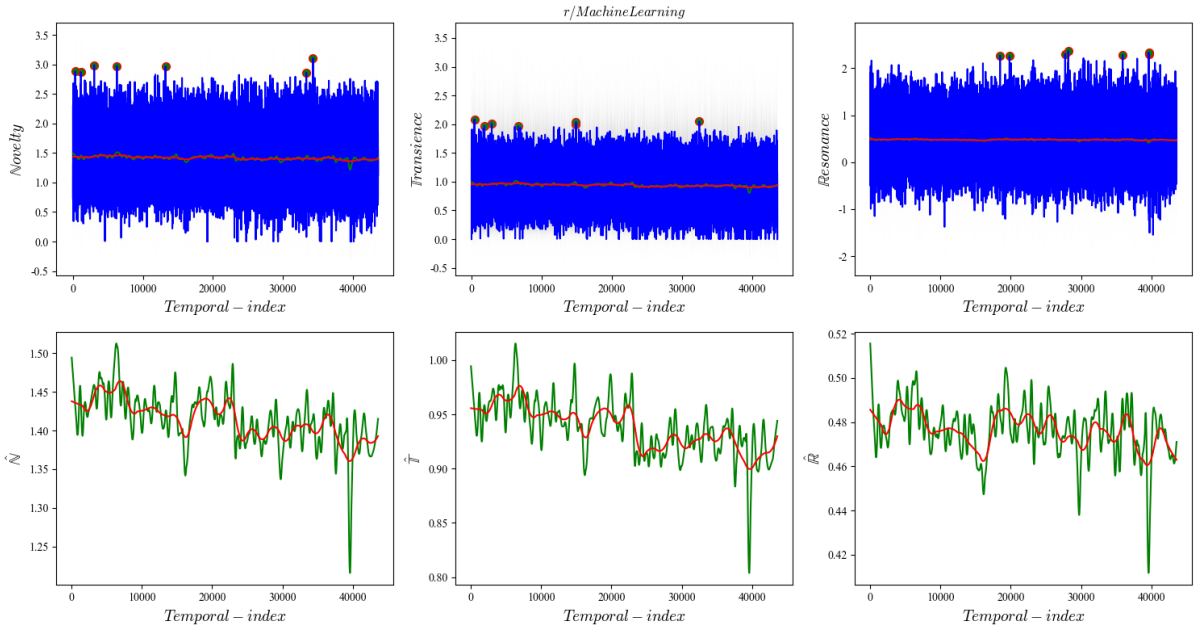

Figure 1: Novelty, transience and resonance for *r/MachineLearning* with adaptive filtering (green: $w = 56$, red: $w = 256$). *r/MachineLearning* shows a weak negative novelty tendency, while resonance stays almost constant due to a decrease in transience.

generate a signal, we train an LDA model on titles for each Subreddit and estimate the novelty, transience and resonance of over time (Figure 1, Appendix A) [1]. Novelty (left panels) captures how much, in a window of three days, the content diverge from previous titles. Similarly, transience captures the degree to which the content differs from future content (middle panels). Finally, resonance is the difference between novelty and transience, such that posts with high novelty and low transience introduce novel content that changes the future. The subreddits' trend potential is estimated as the linear slope coefficient ($\mathbb{N} * \mathbb{R}$) of its post's resonance on novelty (see Figure 2). In comparison with the baseline slope, $M = 0.74$, $SD = 0.03$, trending Subreddits show significant slope increase, $M = 0.79$, $SD = 0.01$, $t_{498} = 27.89, p < .0001$ indicating that $\mathbb{N} * \mathbb{R} > 0.77$ is a signature of trend reservoirs (Figure 3, left panel).

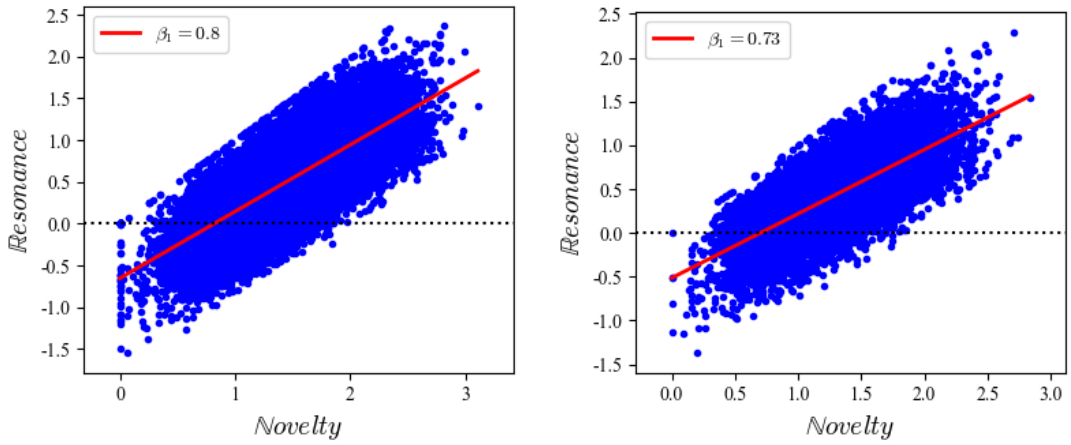

Figure 2: $\mathbb{N} * \mathbb{R}$ slopes for a trending (left) and random (right) subreddit.

Fractal analysis can accurately discriminate between the global dynamics of sociocultural sys-

tems [6, 10]. Some signals show long-range dependencies (i.e., correlations at multiple time scale), while other signals only have short-range dependencies (i.e., correlation between neighboring data points). For trend reservoirs, Hurst exponent H (i.e., an estimate of long-range dependencies), functions as a discrimination signature (see Appendix A). On average trending subreddits show a significantly higher H, $M = 0.5$, $SD = 0.02$ than the baseline, $M = 0.34$, $SD = 0.04$: $t_{498} = 59.05, p < .00001$ (Figure 3, right panel). $H \approx 0.5$ indicates that trend reservoirs only display short-range dependencies, likely due to a larger influx of diverse information, while $H < 0.5$ indicates that the baseline shows anti-persistent and rigid behavior [8]. H and $\mathbb{N} * \mathbb{R}$ are uncorrelated within condition (no-trend: $r = -0.008, p = .9$; trend: $r = -0.1, p = .11$).

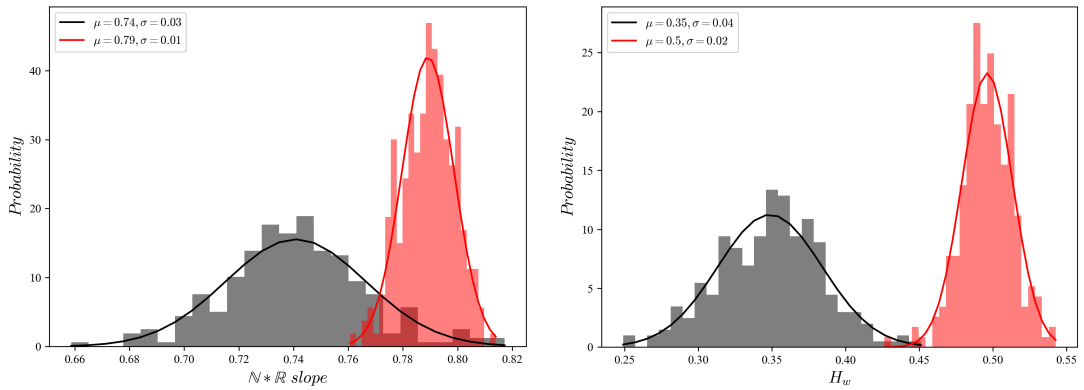

Figure 3: Distributions of $\mathbb{N} * \mathbb{R}$ slopes (left) and H (right) for random (gray) and trending (red). H has a stronger discrimination power than $\mathbb{N} * \mathbb{R}$ for the two conditions.

CONCLUDING REMARKS

This paper presents an approach to trend estimation that identifies trend reservoirs according to their relationship between novelty and resonance, and their degree of persistence. It shows that trend reservoirs have steeper $\mathbb{N} * \mathbb{R}$ slope and higher H in comparison to a random baseline. Importantly, these 'signatures' capture different properties of trend reservoirs, information stickiness and multi-scale correlations respectively, that both have discrimination power.

APPENDIX A: METHODS

The appendix describes data and equations involved in estimation of trend reservoirs. It is important to point out that the approach generalizes to other data sources (e.g., Twitter) and types (e.g., images). For stable estimates, a signal has to consist of a minimum of 265 data points (e.g., posts in a subreddit).

Data and samples

The study uses all post titles from two samples of subreddits from Reddit.com. Subreddits are niche fora that discuss topics related to a forum subject (e.g., *r/MachineLearning*) and titles represent a uniform and comparable data element across all subreddits (i.e., titles relies only on natural language and are hosted at Reddit.com). Power calculations were made for two samples of $n = 25$, but the planned sample sizes were increased by a factor 10 for representativity, resulting in a design with $n = 250$ for each condition. At the time of writing estimates have been made for 25 subreddits for each condition because of the computational requirements. The process is however ongoing.

Design and Statistical analysis
The study uses single a factor design for independent samples that compares human annotated 'trending' subreddits with randomly selected subreddits. Trending subreddits were sampled using sets of trending concepts created by human experts, e.g. *AI* ={*ai, facial recognition, machine leaning ...*} [12]. The trending sample consists of the subreddits with the greatest word overlap in their description (Community Details and Rules) for the each set (e.g., $r/artificial$ and $r/MachineLearning$ for AI) with the constraint of minimum 265 posts. The baseline was randomly selected without replacement, no overlap with the trending sample, and subject to the same minimum number of posts. Statistical tests were conducted with an α-level of .005 [2].The full samples were simulated using parameter estimates from the collected data set under the assumption of Gaussian distributions. Before hypothesis testing, the Shapiro-Wilk W test was used to confirm that the data did not deviate significantly from normality [24].

Novelty and Resonance
For estimates of Novelty and Resonance, a Latent Dirichlet allocation model was trained for each subreddit in order to create dense low-rank vector representations [3]. A grid search was carried out for each model in order to determine the parameter K (number of topics) from 10 to 250 in steps of 10 and the loglikelihood of each model was used as evaluation metric. Novelty ($\mathbb{N}$), transience ($\mathbb{T}$) and resonance ($\mathbb{R}$) were estimated for a window (w) of three days and based on the following equations from [1]:

$$\mathbb{N}_w(j) = \frac{1}{w}\sum_{d=1}^{w} D_{KL}(s^{(j)} \mid s^{(j-d)}) \tag{1}$$

$$\mathbb{T}_w(j) = \frac{1}{w}\sum_{d=1}^{w} D_{KL}(s^{(j)} \mid s^{(j+d)}) \tag{2}$$

$$\mathbb{R}_w(j) = \mathbb{N}_w(j) - \mathbb{T}_w(j) \tag{3}$$

Where s is a K-dimensional document distribution in the LDA model and D_{KL} is the Kullback-Leibler divergence:

$$D_{KL}(s^{(j)} \mid s^{(k)}) = \sum_{i=1}^{K} s_i^{(j)} \times \log_2 \frac{s_i^{(j)}}{s_i^{(k)}} \tag{4}$$

Because LDA can give less than optimal results for short documents, the performance of each model was compared to a model trained on the same data using Non-negative Matrix Factorization and cosine distance [20]. Signal properties were robust across models and LDA chosen for continuity with previous studies.

Nonlinear Adaptive Filtering

Nonliear adaptive filtering is used because of the inherent noisiness of trend signals [11]. First, the signal is partitioned into segments (or windows) of length $w = 2n + 1$ points, where neighboring segments overlap by $n + 1$. The time scale is $n + 1$ points, which ensures symmetry. Then, for each segment, a polynomial of order D is fitted. Note that $D = 0$ means a piece-wise constant, and $D = 1$ a linear fit. The fitted polynomial for ith and $(i + 1)th$ is denoted as $y^{(i)}(l_1), y^{(i+1)}(l_2)$, where $l_1, l_2 = 1, 2, ..., 2n + 1$. Note the length of the last segment may be shorter than w. We use the following weights for the overlap of two segments.

$$y^{(c)}(l_1) = w_1 y^{(i)}(l + n) + w_2 y^{(i)}(l), l = 1, 2, ..., n + 1 \tag{5}$$

where $w_1 = (1 - \frac{l-1}{n}), w_2 = 1 - w_1$ can be written as $(1 - \frac{d_j}{n}), j = 1, 2$, where d_j denotes the distance between the point of overlapping segments and the center of $y^{(i)}, y^{(i+1)}$. The weights decrease linearly with the distance between point and center of the segment. This ensures that the filter is continuous everywhere, which ensures that non-boundary points are smooth.

Adaptive Fractal Analysis

Assuming that stochastic process $X = X_t : t = 0, 1, 2, ...$, with stable covariance, mean μ and σ^2, the process' autocorrelation function for $r(k), k \geq 0$ is:

$$r(k) = \frac{E\left[X(t)X(t + k)\right]}{E\left[X(t)^2\right]} \sim k^{2H-2}, as \quad k \to \infty \tag{6}$$

where H is called the Hurst parameter[18]. For $0.5 < H < 1$ the process is characterized by long-range temporal correlations such that increments are followed by increases and decreases by further decreases. For $H = 0.5$ the time series only has short-range correlations; and when $H < 0.5$ the time series is anti-persistent such that increments are followed by decreases and decreases by increments.

Detrended fluctuation analysis (DFA) is the most widely used method for estimating the Hurst parameter, but DFA may involve discontinuities at the boundaries of adjacent segments. Such discontinuities can be detrimental when the data contain trends [14], non-stationarity [16], or nonlinear oscillatory components [4, 13]. Adaptive fractal analysis (AFA) is a more robust alternative to DFA [9, 27]. AFA consists of the following steps: first, the original process is transformed to a random walk process through first-order integration $u(n) = \sum_{k=1}^{n}(x(k) - \overline{x}), n = 1, 2, 3, ..., N$, where $\overline{x}$ is the mean of $x(k)$. Second, we extract the global trend $(v(i), i = 1, 2, 3, ..., N)$ through the nonlinear adaptive filtering. The residuals $(u(i) - v(i))$ reflect the fluctuations around a global trend. We obtain the Hurst parameter by estimating the slope of the linear fit between the residuals' standard deviation $F^{(2)}(w)$ and w window size as follows:

$$F^{(2)}(w) = \left[\frac{1}{N}\sum_{i=1}^{N}(u(i) - v(i))^2\right]^{\frac{1}{2}} \sim w^H \tag{7}$$