

Trend Reservoir Detection: Minimal Persistence and Resonant Behavior of Trends in Social Media

Kristoffer L. Nielbo^{a,b}, Peter B. Vahlstrup^{a,b}, Anja Bechmann^b and Jianbo Gao^{c,d}

^aCenter for Humanities Computing Aarhus, Jens Chr. Skous Vej 4, Building 1483, 3rd floor, DK-8000 Aarhus C, Denmark

^bDATALAB, School of Communication and Culture, Aarhus University, Helsingforsgade 14, DK-8200 Aarhus N, Denmark

^cCenter for Geodata and Analysis, Faculty of Geographical Science, Beijing Normal University, China

^dInstitute of Automation, Chinese Academy of Sciences, China

Abstract

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 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. This approach contrasts with trend estimation that relies on linear or polynomial techniques to study point-like novelty behavior in social media, and it completes approaches that rely on smooth functions of time.

Keywords

Trend detection, Social media, Information dynamics, Fractal analysis

1. 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. This approach contrasts trend estimation that use linear or polynomial techniques to study point-like novelty behavior in social media, and it completes approaches that rely on smoothing functions [11].

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

CHR 2020: Workshop on Computational Humanities Research, November 18–20, 2020, Amsterdam, The Netherlands


✉ kln@cas.au.dk (K.L. Nielbo); imvpbv@cc.au.dk (P.B. Vahlstrup); anjabechmann@cc.au.dk (A. Bechmann); jbgao.pmb@gmail.com (J. Gao)

🌐 <https://knielbo.github.io/> (K.L. Nielbo)

📄 0000-0002-5116-5070 (K.L. Nielbo); 0000-0002-5588-5155 (A. Bechmann)

© 2020 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

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 [7, 6, 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.

2. Methods

This section 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, FB or 4Chan) and types (e.g., images, video). For stable estimates, a signal has to consist of a minimum of 265 data points (e.g., posts in a subreddit).

2.1. 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. Runtime of the algorithm for a subreddit of approximately 1,763,527 posts, e.g., *r/technology*, on a machine with 12 i7-8750H CPUs at 2.28 GHz, 32 GB RAM, and an NVIDIA GeForce RTX 2080 (only used for preprocessing with a LSTM) is 9344 seconds.

2.2. 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. $AI = \{ai, facial\ recognition, machine\ learning \dots\}$ [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.

In this study *trending subreddits* therefore refer to thematically curated sets of posts (i.e., subreddits) that in their description use more trending keywords within a given domain, e.g., AI . The trend value of a keyword is naturally context-dependent (e.g., *rule-based* and *fixed* knowledge systems have less trend value in AI today than in the previous century) and is not the object of this study. Instead we model the difference between subreddits that either use or do not use keywords that are classified as trending by human annotators [12]. For the simple comparison, we use a baseline that was randomly selected without replacement, have no overlap with the trending sample, and are subject to the same minimum number of posts. This results in two samples of subreddits, a *trending* and random baseline (referred to as *non-trending*), on which we model and compare the information theoretical and dynamic properties. 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].

2.3. 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]:

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

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

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

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

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

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.

2.4. Nonlinear Adaptive Filtering

Nonlinear 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 i th and $(i + 1)$ th is denoted as $y^{(i)}(l_1), y^{(i+1)}(l_2)$, where $l_1, l_2 = 1, 2, \dots, 2n + 1$. Note the length of the last segment may be shorter than w . We use the following weights for the overlap of two segments.

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

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.

2.5. Adaptive Fractal Analysis

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

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

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 $u(n) = \sum_{k=1}^n (x(k) - \bar{x}), n = 1, 2, 3, \dots, N$, where \bar{x} is the mean of $x(k)$. Second, we extract the global trend ($v(i), i = 1, 2, 3, \dots, N$) 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:

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

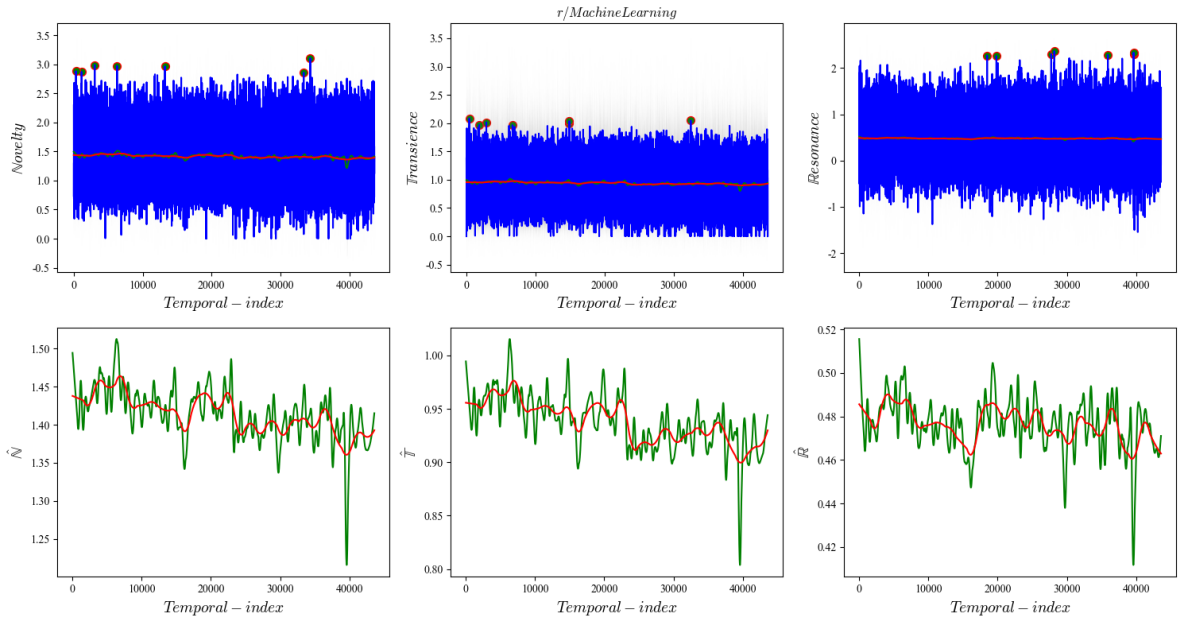


Figure 1: Novelty, transience and resonance for *r/MachineLearning* with adaptive filtering (green: $w = 56$, red: $w = 256$). *r/MachineLearning* shows a weak negative novelty tendency, while resonance stays almost constant due to a decrease in transience.

3. Results

We illustrate estimation of trend reservoirs on Reddit data a single factor design that compares human annotated ‘trending’ subreddits with randomly selected subreddits. To 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) [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).

Fractal analysis can accurately discriminate between the global dynamics of sociocultural systems [7, 8]. 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. On average trending subreddits show a significantly higher H , $M = 0.5$, $SD = 0.02$ for resonance 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 [10]. H and $\mathbb{N} * \mathbb{R}$ are uncorrelated within condition (no-trend: $r = -0.008$, $p = .9$; trend:

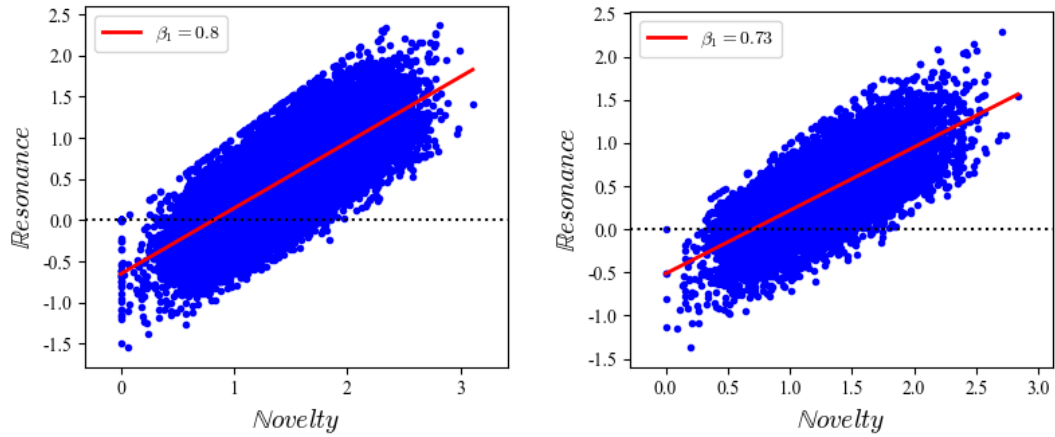


Figure 2: $\mathbb{N} * \mathbb{R}$ slopes for a trending (left) and random (right) subreddit.

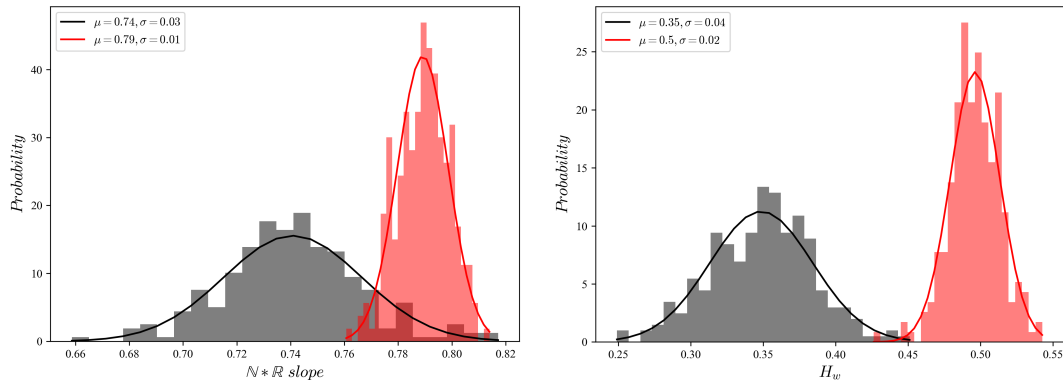


Figure 3: Distributions of $\mathbb{N} * \mathbb{R}$ slopes (left) and H (right) for random (gray) and trending (red). H has a stronger discrimination power than $\mathbb{N} * \mathbb{R}$ for the two conditions.

$r = -0.1, p = .11$).

4. Conclusion

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. Importantly, this paper identifies a statistically reliable difference between these two groups irrespective of the validity of the sampling procedure. The findings actually support that Gyodi et al. [12] did indeed identify relevant structure with their keywords. Some of the most direct application domains of these findings are decision support and recommender systems in order to identify and curate subsets of streaming data that provide information on any given set of topic. For discussion boards, this amounts to recommending the relevant

subreddits that have produced and are most likely to continue to produce trending posts on a given subject. Similarly, the results could be used for classification and early detection of critical states in patients, when the resonant and novel properties of their journals only display short-range correlations over time. Importantly, both application examples are only tentative suggestions and need further testing.

References

- [1] A. T. J. Barron et al. “Individuals, Institutions, and Innovation in the Debates of the French Revolution”. In: *arXiv:1710.06867* (2017), p. 8.
- [2] D. J. Benjamin et al. “Redefine statistical significance”. In: *Nature Human Behaviour* (2017). (Visited on 09/11/2017).
- [3] D. M. Blei, A. Y. Ng, and M. I. Jordan. “Latent dirichlet allocation”. In: *the Journal of machine Learning research* 3 (2003), pp. 993–1022. (Visited on 01/08/2015).
- [4] Z. Chen et al. “Effect of nonlinear filters on detrended fluctuation analysis”. In: *Phys. Rev. E* 71.1 (Jan. 2005), p. 011104. DOI: 10.1103/PhysRevE.71.011104.
- [5] A. Chinnov et al. “An Overview of Topic Discovery in Twitter Communication through Social Media Analytics”. en. In: *Twenty-first Americas Conference on Information Systems*. 2015, p. 10.
- [6] J. Gao et al. “A multiscale theory for the dynamical evolution of sentiment in novels”. en. In: *Behavioral, Economic and Socio-cultural Computing (BESC)*. 2016.
- [7] J. Gao et al. “Culturomics meets random fractal theory: insights into long-range correlations of social and natural phenomena over the past two centuries”. en. In: *Journal of The Royal Society Interface* 9.73 (Aug. 2012), pp. 1956–1964. (Visited on 05/19/2016).
- [8] J. Gao, P. Fang, and F. Liu. “Empirical scaling law connecting persistence and severity of global terrorism”. en. In: *Physica A: Statistical Mechanics and its Applications* 482 (Sept. 2017), pp. 74–86. ISSN: 03784371. DOI: 10.1016/j.physa.2017.04.032. (Visited on 09/15/2017).
- [9] J. Gao, J. Hu, and W.-w. Tung. “Facilitating Joint Chaos and Fractal Analysis of Biosignals through Nonlinear Adaptive Filtering”. en. In: *PLoS ONE* 6.9 (Sept. 2011). Ed. by M. Perc, e24331. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0024331. (Visited on 10/18/2017).
- [10] J. Gao et al. *Multiscale Analysis of Complex Time Series: Integration of Chaos and Random Fractal Theory, and Beyond*. English. 1 edition. Hoboken, N.J: Wiley-Interscience, Sept. 2007. ISBN: 978-0-471-65470-4.
- [11] K. L. Gray. “Comparison of Trend Detection Methods”. en. In: *Graduate Student Theses, Dissertations, & Professional Papers* 228 (2007), p. 98. URL: <https://scholarworks.umt.edu/etd/228>.
- [12] K. Gyodi, L. Nawaro, and M. Palinski. *Keyword frequency in popular tech media*. 2019. URL: [Zenodo.%20http://doi.org/10.5281/zenodo.2554116](http://doi.org/10.5281/zenodo.2554116).
- [13] J. Hu, J. Gao, and X. Wang. “Multifractal analysis of sunspot time series: the effects of the 11-year cycle and Fourier truncation”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2009.02 (Feb. 2009), P02066. ISSN: 1742-5468. (Visited on 02/04/2018).

- [14] K. Hu et al. “Effect of trends on detrended fluctuation analysis”. en. In: *Physical Review E* 64.1 (June 2001). ISSN: 1063-651X, 1095-3787. (Visited on 02/04/2018).
- [15] Q. Hu. et al. “Dynamic evolution of sentiments in Never Let Me Go”. In: *HAL preprint hal-02143896* (2019).
- [16] J. W. Kantelhardt et al. “Multifractal detrended fluctuation analysis of nonstationary time series”. In: *Physica A: Statistical Mechanics and its Applications* 316.1-4 (2002), pp. 87–114.
- [17] A. Madani, O. Boussaid, and D. E. Zegour. “What’s Happening: A Survey of Tweets Event Detection”. en. In: *INNOV 2014 : The Third International Conference on Communications, Computation, Networks and Technologies*. 2014, p. 7.
- [18] B. Mandelbrot. *The Fractal Geometry of Nature*. English. Updated ed. edition. San Francisco: Times Books, 1982. ISBN: 978-0-7167-1186-5.
- [19] M. Mathioudakis and N. Koudas. “TwitterMonitor: trend detection over the twitter stream”. en. In: *Proceedings of the 2010 international conference on Management of data - SIGMOD '10*. Indianapolis, Indiana, USA: ACM Press, 2010, p. 1155. ISBN: 978-1-4503-0032-2. (Visited on 10/13/2019).
- [20] D. Moyer et al. “Determining the Influence of Reddit Posts on Wikipedia”. en. In: *Proceedings from the 2015 ICWSM Workshop*. 2015, p. 8.
- [21] J. Murdock, C. Allen, and S. DeDeo. “Exploration and Exploitation of Victorian Science in Darwin’s Reading Notebooks”. In: *arXiv preprint arXiv:1509.07175* (2015). (Visited on 01/11/2016).
- [22] K. L. Nielbo et al. “A curious case of entropic decay: Persistent complexity in textual cultural heritage”. In: *Digital Scholarship in the Humanities* (Oct. 2018). ISSN: 2055-7671. DOI: 10.1093/llc/fqy054. (Visited on 07/02/2019).
- [23] K. L. Nielbo et al. “Automated Compositional Change Detection in Saxo Grammaticus’ Gesta Danorum”. en. In: *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*. 2019, p. 13.
- [24] S. S. Shapiro and M. B. Wilk. “An analysis of variance test for normality (complete samples)”. In: *Biometrika* 52.3/4 (1965), pp. 591–611. (Visited on 05/22/2013).
- [25] S. Stieglitz et al. “Social media analytics – Challenges in topic discovery, data collection, and data preparation”. en. In: *International Journal of Information Management* 39 (Apr. 2018), pp. 156–168. ISSN: 02684012. DOI: 10.1016/j.ijinfomgt.2017.12.002. (Visited on 07/25/2019).
- [26] D. Y. Tenen. “Toward a Computational Archaeology of Fictional Space”. en. In: *New Literary History* 49.1 (2018), pp. 119–147. ISSN: 1080-661X. DOI: 10.1353/nlh.2018.0005. (Visited on 10/13/2019).
- [27] W.-w. Tung et al. “Detecting chaos in heavy-noise environments”. en. In: *Physical Review E* 83.4 (Apr. 2011). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.83.046210. (Visited on 10/14/2019).
- [28] M. Wevers, J. Gao, and K. L. Nielbo. “Tracking the Consumption Junction: Temporal Dependencies between Articles and Advertisements in Dutch Newspapers”. en. In: *arXiv:1903.11461 [cs]* (Mar. 2019). arXiv: 1903.11461. URL: <http://arxiv.org/abs/1903.11461> (visited on 10/13/2019).