The shape of attention: detecting and classifying anomalous lexical activity in eighty years of the daily press

Elias Echikr Benoît de Courson Simon Coste

August 22, 2026

Abstract

blabla

Contents

1 Introduction

motivation: [7], [1], [14]

Gallicagram: [2] + exemples d'utilisation ?

Sornette [17] and Crane [6]

Lehmann [13]

studies on word/token counts [10], [15]

change detection [8, 11]

kewords: reflexivity, time change, news?

main goal of Bouchad et al: show that jumps are not caused by news or volume, but by "endogeneous" processes. what if it's already the case... for news themselves? is there a structure across the journals or do some activity spikes appear out of nowhere?

2 Data collection

Our corpus gathers the articles published by $J = 36$ French media outlets: national and regional dailies, weeklies and online news sites, including the main national dailies — *Le Monde*, *Le Figaro*, *Les Échos*. It currently contains more than 16 million articles. The archive depth varies across outlets: the oldest one, that of *Le Monde*, goes back to 1944, and all 36 outlets are covered from 2008 onwards. The corpus is updated daily.

Each article is segmented into words (*tokens*), and the counts are aggregated at the publication-day level over a vocabulary common to all media. For each media j, each word w in the vocabulary and each day t, we thus record the number of occurrences $X_t^{w,j}$ and the total number of tokens N_t^j published by this outlet on this day.

These counts define the two corpora used in the sequel. The *per-media corpus* is the collection of the individual series $(X_t^{w,j}, N_t^j)$: it tracks the activity of a given journal. The *unified corpus* aggregates the 36 outlets into a single journal, by summing occurrences and exposures day by day:

$$X_t^w = \sum_{j=1}^{J} X_t^{w,j}, \qquad N_t = \sum_{j=1}^{J} N_t^j.$$

3 Detecting anomalous activity

3.1 Vocabulary selection

We will study the word or n-gram frequency in the unified corpus, across time. The most frequent words often do not bear any interesting information (such as « the », « and », « of », etc.). We will therefore only consider words in a carefully selected vocabulary: we choose words and n-grams

1. that are not connection words, simple verbs, demonstrative pronouns, etc.
2. ?

3.2 Statistical model and fitting

Fix a word w in the vocabulary. For each publication day $t \in \{1, \dots, T\}$ we observe the count $X_t \equiv X_t^w \in \mathbb{N}$ of occurrences of w, and the *exposure* of that day, which is the total number of tokens N_t published that day. The (normalized) word frequency is

$$f_t = 10^5 \, \frac{X_t}{N_t}.$$

Our preliminary goal is to detect days where f_t is unusually high; but we have to account for the fact that N_t varies from day to day. A day with low N_t is more likely to have a high f_t simply because it has fewer tokens to distribute, and thus f_t would have a higher variance. To automate the detection of these unusually high frequencies while taking into account N_t, we will first fit a statistical model to the count data $(X_t^w)_{t=1}^T$, and then we use the model to compute p-values for each day.

We model the daily count as a mixture: a Bernoulli variable decides whether the word is used at all that day (there are many days where the word is not used at all). Then, conditionally on being used (ie $X_t^w > 0$), we model $X_t^w - 1$ as a negative binomial[1] variable whose mean is proportional to N_t. Formally, with parameters $\theta = (p, \mu, r) \in (0,1) \times (0,\infty)^2$,

$$_\theta(X_t = k) = \begin{cases} p, & k = 0, \\ (1-p)\, f(k-1;\ \mu N_t,\ r)\,, & k \geq 1, \end{cases} \tag{1}$$

with $f(\cdot \mid m, r)$ being the density of the negative binomial with mean m and variance $m + m^2/r$, that is

$$f(k \mid m, r) = \frac{\Gamma(k+r)}{\Gamma(k+1)\Gamma(r)} \left(\frac{r}{r+m}\right)^r \left(\frac{m}{r+m}\right)^k, \tag{2}$$

In the sequel, we simply note $X_t \sim (p, \mu N_t, r)$.

In the econometrics and applied statistics litterature, models like (1) with a probability p of being zero, and a probability $1-p$ of being random with a certain parametrized distribution, are known under the name of *hurdle models* (see [5, 16]) or *zero-inflated* models, as in [12].

Conditionnally on being used (ie $X_t > 0$), the mean of X_t is proportional to N_t: this is standard in the litterature for regression of count data [4, 9]. The moments of the model are

$$_\theta[X_t] = (1-p)\,(1+\mu N_t), \tag{3}$$

$$_\theta(X_t) = (1-p)\left(\mu N_t + \frac{(\mu N_t)^2}{r}\right) + p(1-p)\,(1+\mu N_t)^2, \tag{4}$$

the second term of (4) being the extra variance due to the Bernoulli component.

We will fit models like (1) to each word w in the vocabulary, using maximum likelihood estimation. The likelihood of (1) is easy to evaluate: writing $n_0 = \#\{t : X_t = 0\}$ for the number of inactive days and $\mathcal{A} = \{t : X_t \geq 1\}$ for the set of active days, the log-likelihood is

$$\ell(p, \mu, r) = \underbrace{n_0 \log p + (T - n_0)\log(1-p)}_{\text{Bernoulli}} + \underbrace{\sum_{t \in \mathcal{A}} \log f(X_t - 1 \mid \mu N_t,\ r)}_{\text{counting}}, \tag{5}$$

in which p appears only in the first term and (μ, r) only in the second. Hence the maximum likelihood estimator gives $\hat{p} = n_0/T$ on the first hand, and then $(\hat{\mu}, \hat{r})$ maximises a negative binomial log-likelihood on the active days $\mathcal{A}$. This can be done using standard methods implemented in the Scipy library.

3.3 p-values and spike detection

Once the parameters $\hat{\theta} = (\hat{p}, \hat{\mu}, \hat{r})$ are fitted for a certain word, we can compute the likelihood of any count X_t under the fitted model, given the exposure N_t: for days with activity (ie $X_t > 0$), the likelihood is

$$p_t = (1 - \hat{p}) f(X_t - 1 \mid \hat{\mu} N_t, \hat{r}), \tag{6}$$

and we call

$$_t = -\log_{10}(p_t)$$

[1] Elementary context on the negative binomial distribution is given in Appendix A.

the *surprise* of the day t. Given a threshold H, a day is declared anomalous[2] when $_t \geq H$, that is $p_t < 10^{-H}$. In the sequel, we use $H = 6$ to be very conservative. If the statistical fit is correct, that would correspond to a one-in-a-million event.

Remark 3.1. Two-stage fitting and robustification

Remark 3.2. Given fitted parameters $\hat{\theta}$, the quantiles of $\mathrm{BNB}(\hat{p}, \hat{\mu}N, \hat{r})$ are decreasing with the exposure N: with a low exposure N, the count X has a higher variance and thus a higher probability to be large. Consequently, two different days $t \neq t'$ with the same word count $X_t = X_{t'}$ can have a very different surprise, if the exposure of the two days are different, $N_t \neq N_{t'}$. Indeed, if $N_t < N_{t'}$, then $s_t < s_{t'}$.

4 Statistics of activity spikes

4.1 Anomalous days

total number of spikes | number of words with a spike | distribution of number of spikes per word (histogram) | histogram of the surprises especially above the threshold

4.2 Co-jumps

For any day t, we note K_t the number of words having a spike at this day.
time series of of K_t | fit with heavy tail and quantiles and max |

5 Classification of single jumps

5.1 Methodology

The goal of this section is to find some regular time patterns around activity spikes. For this, we perform a PCA on time windows around spikes. More precisely, let w be a word for which there is an anomalous activity spike at a certain day t. Let D be a window size, to be chosen as a hyperparameter — typically, $D = 15$ days. We extract a window of D time units before and after the spike at time t, obtaining a vector of size $2D+1$, $Z_t^w = (X_{t-D}, \ldots, X_{t-1}, X_t, X_{t+1}, \ldots, X_{t+D})$. If n is the total number of anomalous activity spikes detected in the corpus, we thus obtain a matrix $Z = (Z_t^w)_{\text{spike for } (w,t)}$ of size $n \times (2D+1)$.

Non-maximal suppression . talk abt NMS [3]

Standardization of the windows . Before performing a PCA, we need to standardize the lines of Z to have mean 0 and variance 1. We could compute the standardized z-score for each window, but it is more sound to compute the mean and standard deviation in the window, *excluding* the spike at time t. Thus, let Z_s be a spike. We note

$$m_s = \frac{1}{2D} \sum_{|i-t| \leq D, i \neq t} X_i,$$

$$\sigma_s = \sqrt{\frac{1}{2D} \sum_{|i-t| \leq D, i \neq t} (X_i - m_s)^2}.$$

and we replace Z_s by its standardized version, $Z'_s = (Z_s - m_s)/\sigma_s$. We then perform a PCA (or a variant) on the standardized matrix Z'.

[2] We only test for excess activity and we did not consider the opposite, an anomalous absence of activity.

5.2 Results

plots: valeurs singulières avec fit marchenko-pastur, nombre d'outliers
part de la variance expliquée (K50, K90) y compris pour des hyperparamètres
formes des vecteurs propres et interprétations

Remark 5.1. We explored different variants of the classical PCA, with no noticeable improvements. In particular, Kernel PCA did not explain a significantly higher part of the data variance, showing that the data does primarily have a linear structure. Further results on these explorations are given in Appendix B.

6 The structure of jumps across journals

We now arrive to the main part of the paper, where we show patterns of activity jumps across the different J *journals*. To do so, we have to split the windows of the unified corpus into per-media windows. Our methodoly is as follows: for each detected spike $s = (w, t)$, we extract J windows of activity around the spike and stack them. If

$$Z_{s,j} = (X_{t-D}^{w,j}, \dots, X_{t-1}^{w,j}, X_t^{w,j}, X_{t+1}^{w,j}, \dots, X_{t+D}^{w,j})$$

is the window of activity around the spike for the word w and the journal j, then we set

$$Z_s = (Z_{s,j})_{j=1}^J \in \mathbb{R}^{J \times (2D+1)}$$

the window of activity around the spike for the word w and the journals.

QUESTIONS: quelle normalisation adopter ?

7 Conclusion

Data and code availability

Acknowledgements

References

[1] Cecilia Aubrun, Rudy Morel, Michael Benzaquen, and Jean-Philippe Bouchaud. Riding wavelets: A method to discover new classes of financial price jumps. *Proceedings of the National Academy of Sciences of the United States of America*, 122(6):e2409156121, 2025.

[2] Benjamin Azoulay and Benoît de Courson. Gallicagram: un outil de lexicométrie pour la recherche. *Corpus*, (24), 2023.

[3] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-NMS: improving object detection with one line of code. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5561–5569, 2017.

[4] A. Colin Cameron and Pravin K. Trivedi. *Regression Analysis of Count Data*. Cambridge University Press, 2 edition, 2013.

[5] John G. Cragg. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39(5):829–844, 1971.

[6] Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105 (41):15649–15653, 2008.

[7] David M Cutler, James M Poterba, and Lawrence H Summers. What moves stock prices?, 1988.

[8] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501, 2016.

[9] Joseph M. Hilbe. *Negative Binomial Regression*. Cambridge University Press, 2 edition, 2011.

[10] Slava M. Katz. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2(1):15–59, 1996.

[11] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*, pages 625–635, 2015.

[12] Diane Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

[13] Janette Lehmann, Bruno Gonçalves, José J. Ramasco, and Ciro Cattuto. Dynamical classes of collective attention in Twitter. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*, pages 251–260, 2012.

[14] Riccardo Marcaccioli, Jean-Philippe Bouchaud, and Michael Benzaquen. Exogenous and endogenous price jumps belong to different dynamical classes. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(2):023403, 2022.

[15] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[16] John Mullahy. Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365, 1986.

[17] Didier Sornette, Fabrice Deschâtres, Thomas Gilbert, and Yann Ageon. Endogenous versus exogenous shocks in complex networks: an empirical test using book sale rankings. *Physical Review Letters*, 93(22):228701, 2004.

A Negative binomial distribution

The usual parametrization of the negative binomial distribution is with a success parameter p and a dispersion parameter $r > 0$, with probability mass function given by

$$\frac{\Gamma(k+r)}{\Gamma(k+1)\Gamma(r)} p^r (1-p)^k \qquad (k = 0, 1, 2, \dots) \tag{7}$$

When r is an integer, this distribution is often interpreted as the number of failures before the r-th success in a sequence of Bernoulli trials with success probability p, and known under the name of Pascal distribution. However, in general r needs not be an integer. The mean and variance are given by

$$(X) = \frac{r(1-p)}{p}, \tag{8}$$

$$(X) = \frac{r(1-p)}{p^2}. \tag{9}$$

Noting $m = r(1-p)/p$, a simple computation gives $(X) = m + m^2/r$. The parameter r thus accounts for overdispersion: when it is small, the negative binomial has a higher variance that the corresponding Poisson distribution, which has mean and variance equal to m. It is thus more intuitive to parametrize the negative binomial distribution in terms of m and r. To go back to p, we can use the relation $p = r/(r+m)$ and $1 - p = m/(r+m)$. With this new parametrization, the form of the probability mass function is then

$$f(k \mid m, r) = \frac{\Gamma(k+r)}{\Gamma(k+1)\Gamma(r)} \left(\frac{r}{r+m}\right)^k \left(\frac{m}{r+m}\right)^r, \tag{10}$$

as in (2).

B Further results on PCA

Ici : kernel PCA etc