Media Flows Analysis for Sustainable Development of Urban Environment

Vladislav Evstigneev, Dmitry Voronin, and Pavel Kuznetsov

Sevastopol State University, Universitetskaya st., 33, 299053 Sevastopol, Russia

Abstract. An objective analysis of multidimensional media flows in the implementation of the concept of sustainable development of the urban environment is based on methods of statistical data processing, the sample of which is subject to certain constraints to improve the reliability statistical conclusions. In particular, the presence of serial correlation in the series of a random variable reduces the reliability of statistical estimates. This paper proposes a method for assessing the intensity of information flows generated by a separate media source against the background of two different dominant informational agendas: before and after the outbreak of coronavirus infection. The technique makes it possible to assess the probability of the media generating messages of a given discourse taking into account the autocorrelation of the series. The developed and implemented algorithm for assessing the likelihood of information messages on a given topic can be used in the future to study the stability of information "intrusions" into the target media space, as well as to simulate the consequences of the implementation of measures for the digital transformation of the urban environment within the concept of its sustainable development.

1 Introduction

In the recent decade special attention is paid to scientific problems associated with the analysis of social networks. Extraction of information from social networks is actively used in the promotion of services, sales, healthcare and even in politics. Analyzing media governments aim to identify the interest of citizens in government activity or to detect possible fraudulent and terrorist groups. During the 2016 US presidential election, the number of related Twitter posts exceeded more than a billion the day before the election. The authoritative Internet publication TechCrunch confirms that social networks were much more effective than traditional polls and exit polls in predicting the final election result [1, 3, 8].

The study of social media on a rigorous scientific basis allows one to solve a number of important sociological problems, one of which is the assessment of the reintegration potential of post-conflict societies. By using modern information technologies of agent-based modeling and Big Data, it becomes possible to study the processes of forming public opinion in social networks.

In general, the data generated in social networks creates the basis for their modeling (the so-called Social Network Analysis) and predicting their reactions to certain external challenges and news stories [9]. However, the most important stage of social computing [1, 2, 10] is still the primary analysis of social media data. The process of representing,

analyzing, and extracting valid insight from social media data has been called Social Mining [8].

One of the most important tasks of Social Mining is to analyze all information flows of media in order to classify them, to identify the natural patterns or to assess the impact of external informational intrusion.

An objective analysis of multivariate sociological data is based on methods of statistical data processing, the sample of which is subject to certain constraints in order to increase the reliability of statistical conclusions. In particular, the presence of serial correlation in the series of a random variable reduces the reliability of statistical estimates.

This paper proposes a method for assessing the intensity of information flows generated by a separate media source against the background of two different dominant information agendas: before and after the outbreak of coronavirus infection. The technique makes it possible to assess the probability of the media generating messages of a given discourse, taking into account the autocorrelation of the series.

2 Method

The study was based on dataset of social media messages having been downloaded using online service Medialogia for monitoring of social media. In order to validate the method to be developed a test sample of striking topic messages was extracted by Medialogia API. Data extraction by keywords "Crimea, sanctions, annexation" was engaged for two separate periods: March-April 2019 and March-April 2020. The key media source for the procedure was the Ukrainian Internet edition Korrespondent.net. Two identical calendar periods differ in their information background: in 2019, the dominant information agenda was the presidential elections in Ukraine; spring 2020 is the peak of the spread of the COVID19 coronavirus pandemic.

The data was presented in the form of a daily timeseries of a binary random variable, where 0 means the absence of messages on a given topic for the given day, 1 – appearing of at least one message. A simple comparison of the descriptive characteristics of a random binary variable may face several problems. On the one hand, the sample sizes are insufficient for a reliable statistical estimation and a normal approximation of the distribution of the binomial value. On the other hand, the emerging information "burst" can last for several days and manifest itself in the form of autocorrelation of the timeseries. In this regard, it is necessary to develop the most generalized method for assessing the frequency of occurrence of messages, which has been done in this study.

The paper poses the problem of determining the probability of occurrence of information messages by one source. Having sufficiently long series of daily registration of messages of a given topic, it becomes possible to estimate the probability p and calculate σ – standard error of its calculation. Let for the information source $x^{(L)}$ stands for sequence of binary elements of length L . The true value of the element corresponds to the situation of the appearance of messages in the current time interval. The probability of the event "presence of a message" is calculated by the formula (1).

$$p = \frac{\sum_{i=1}^L x_i}{L} = \frac{r}{L} \quad (1)$$

For reliable calculations of the frequency of "presence of a message" r/L , σ should be estimated (sampling error), which is formed due to the problem of representativeness of samples as well as due to the presence of autocorrelation. As a consequence of calculation

error, it is required to develop a method for assessing the frequency of events using binomial distribution (Eq.2) [4, 6, 7].

$$\text{Bin}(r|p, L) = \binom{L}{r} p^r (1-p)^{L-r} \quad (2)$$

In this study we used β -distribution $\phi \sim \text{Beta}(\varepsilon, \zeta)$ for the autoregressive coefficient. For the stochastic case, the value of the total variance of r/L can be calculated using Eq.3 or simplified to the form of Eq.4. The first summand in the right part of Eq.4 includes variance of a sample mean conditioned by serial correlation. This variance may be calculated by the Eq.5, where V describes the bias of the sample variance of the mean (see Eq.6). Transforming Eq.5 by taking into account the β -distributions of the error of p estimate and the parameter ϕ one can obtain Eq.7. In accordance with the fact that these random variables obey β -distribution, the final form of the formula has the form of Eq.8. The Eq.8 allows to estimate the standard error of the proportion of "presence of a message" events $\sigma = \sqrt{D(r/L)}$.

$$\mathbf{D}\left(\frac{r}{L}\right) = \mathbf{E}\left[\mathbf{D}\left(\frac{r}{L} \middle| p, \phi\right)\right] + \mathbf{E}\left[\mathbf{D}\left(\mathbf{E}\left[\frac{r}{L} \middle| p, \phi\right] \middle| p\right)\right] + \mathbf{D}\left[\mathbf{E}\left(\frac{r}{L} \middle| p\right)\right] \quad (3)$$

$$\mathbf{D}\left(\frac{r}{L}\right) = \mathbf{E}\left[\mathbf{D}\left(\frac{r}{L} \middle| p, \phi\right)\right] + \mathbf{E}[\mathbf{D}(p|p)] + \mathbf{D}(p) = \mathbf{E}\left[\mathbf{D}\left(\frac{r}{L} \middle| p, \phi\right)\right] + \mathbf{D}(p) \quad (4)$$

$$\mathbf{D}\left(\frac{r}{L} \middle| p, \phi\right) = V \cdot \frac{s_x^2}{L} = \frac{1}{L} \cdot p(1-p) \cdot V \quad (5)$$

$$V = 1 + 2 \sum_{j=1}^{L-1} \frac{j}{L} \phi^{L-j} \quad (6)$$

$$\begin{aligned} \mathbf{D}\left(\frac{r}{L}\right) &= \mathbf{E}\left[\frac{1}{L} \cdot p(1-p) \cdot V\right] + \mathbf{D}(p) = \frac{1}{L} \cdot \mathbf{E}[p - p^2] \cdot \mathbf{E}(V) + \mathbf{D}(p) = \\ &= \frac{1}{L} \cdot [\mathbf{E}(p) - \mathbf{D}(p) - \mathbf{E}^2(p)] \cdot \left(1 + \frac{2}{L} \sum_{j=1}^{L-1} j \frac{\Gamma(\varepsilon + \zeta) \Gamma(\varepsilon + L - j)}{\Gamma(\varepsilon) \Gamma(\varepsilon + \zeta + L - j)}\right) + \mathbf{D}(p) \end{aligned} \quad (7)$$

$$\mathbf{D}\left(\frac{r}{L}\right) = \frac{ab}{L(a+b)^2(a+b+1)} \cdot \left[(a+b) \left(1 + \frac{2}{L} \sum_{j=1}^{L-1} j \frac{\Gamma(\varepsilon + \zeta) \Gamma(\varepsilon + L - j)}{\Gamma(\varepsilon) \Gamma(\varepsilon + \zeta + L - j)}\right) + L\right] \quad (8)$$

In order to make estimation of σ reliable, $x^{(L)}$ should be splitted on a set of N subsequences $x^{(n)}_l$ ($l=1..N$) of smaller length $n < L$. Taking into account the p.d.f. of r and p one can express unconditional distribution of r as:

$$f(r) = \int_0^1 f(r|p) \cdot f(p) dp = \int_0^1 \text{Bin}(r|p, n) \cdot \text{Beta}(r|a, b) dp$$

which is a definition of beta-binomial p.d.f:

$$f(r) = \text{BB}(r|n, a, b) = \frac{\Gamma(a+b) \Gamma(a+r) \Gamma(n-r+b) \Gamma(n+1)}{\Gamma(a) \Gamma(b) \Gamma(a+b+n) \Gamma(r+1) \Gamma(n-r+1)} \quad (9)$$

Eq.9 enables to construct a likelihood function:

$$LK(a,b)=\prod_{l=1}^N BB(r_l|n,a,b). \tag{10}$$

Maximization of Eq.10 allows one to estimate parameters $(\hat{a},\hat{b})=\arg\max_{a,b} LK(a,b)$. A point estimate of the parameters $(\hat{\varepsilon},\hat{\zeta})$ of β -distribution of the autoregressive coefficient φ can be given by the same maximum likelihood method. In order to do this, it is enough to build a functional:

$$LK(\varepsilon,\zeta)=\prod_{l=1}^N Beta(\phi_l|\varepsilon,\zeta),$$

where ϕ_l – a lag-1 serial correlation coefficient estimated from the sequence $x^{(n)}_l$. It is preferable to use the logarithm of the likelihood function $\ln LK$. Based on the above formulas, a computational algorithm was developed within the framework of the statistical programming environment R [5].

3 Validation of the proposed approach

The assessment of the likelihood of a message of a given topic on a selected media resource on the Internet was determined using data from Medialogia API online service for two periods - March-April of 2019 and 2020. For each year, an estimate was obtained for the probability p and the error σ using the above equations. Table 1 shows the numerical estimates of the parameters of the beta-binomial and beta distribution for p and autoregressive coefficient φ .

Table 1. Parameters of probability distribution functions of Eq.4

Years	Parameters of distribution function			
	a	b	E	ζ
2019	38,9	19,4	1,6	6,4
2020	7,0	9,1	3,6	9,4

Figure 1 shows the beta-binomial distribution curves for the number of news messages generated by the media source. It can be seen that in 2020 the number of messages decreased by about one and a half times compared to the same period (March-April) in 2019. In 2019 the confidence region is narrowed; in 2020 this region turns out to be quite wide, probably due to the limited statistical sample of data. It is also worth noting that the distribution of messages in 2020 has a symmetrical shape, close to Gaussian. The shape of the 2019 curve has some asymmetry (skewness was -0.2).

Based on the obtained estimates of the parameters the standard error σ of the probability of occurrence of a message for a given topic was calculated using Eq.4. For 2019 $\sigma = 0.18$, for 2020 this value was 0.22. To assess the significance of differences between samples of different years one may conduct a standard t-test under the assumption that the sample mean obeys the normal distribution according to the Central Limit Theorem. Calculations have shown that the average number of messages in two adjacent years differs significantly (confidence level of 95%).

Fig. 1. Distribution density curves of probability of message occurrence on media source

This result is quite clear accounting for the key news that formed the main agenda of the given media source. On the one hand, in spring 2019, an active pre-election campaign of candidates for the President occurred in Ukraine. The likelihood of the appearance in the media space of messages relevant to the set of keywords increased significantly during that period. On the other hand, the beginning of 2020 was marked by the global coronavirus pandemic and all news agendas were in one way or another related to this topic. Obviously, the politically oriented content of many media sources decreased slightly during this period.

4 Conclusions

The processes of transition from a technology-oriented paradigm of the formation of a smart urban environment to a person-oriented concept, taking into account the individual needs of individual citizens and various social groups, requires the creation of developed information technologies for analyzing media streams when verifying measures for the digital transformation of the urban environment.

The developed and implemented algorithm for assessing the probability of occurrence of information messages of a given topic will be useful in future to study the stability of information "intrusions" into the target media space. It focuses on providing decision-makers with the necessary information support, allowing them to form reasoned, balanced decisions on the development of urban areas.

Acknowledgements

The reported study was funded by RFBR and Sevastopol Government (project №20-47-920006) and by Sevastopol State University (project №28/06-31 "Development of methods of agent modeling and Big Data for the analysis of social media in post-conflict societies").

References

1. A.A. Azarov, E.V. Brodovskaya, A.Yu. Bubnov, A.A. Glazkov, R.Ya. Gibadulin, O.V. Dmitrieva, A.M. Dolgorukov, R.I. Zhdanov, V.F. Musina, V.D. Nechaev, A.S. Ognev, A.V. Sirotkin, A.V. Suvorova, V.V. Titov, A.L. Tulupyev, T.V. Tulupyeva, A.A. Filchenkov, R.M. Yusupov, *Proceedings of SPIIRAS*, **3** (**26**), 136 (2013)
2. N.L. Johnson, S. Kots, N. Balakrishnan, *One-dimensional continuous distributions: Part 2*, 600 (2012)
3. TechCrunch – Startup and Technology News (2018), <https://techcrunch.com/>
4. C.G. Bowsher, P.S. Swain, *Proceedings of the National Academy of Sciences*, **109** (**20**), 1320 (2012)
5. Core Team R, <http://www.r-project.org/>
6. MacKay, Cambridge University Press, 628 (2003)
7. D. Wilks, *Journal of Climate*, **10**, 65 (1997)
8. R. Zafrani, Cambridge University Press, 380 (2014)
9. Q. D. Truong, Q. B. Truong, T. Dkaki, *International Conference on Nature of Computation and Communication*, 276 (2016)
10. J. Yang, J. McAuley, J. Leskovec, *13th International Conference on Data Mining*, 1151 (2013)