Empirical Economics (2024) 66:467–481
https://doi.org/10.1007/s00181-023-02463-1

Selecting sensitive web info via conditional probabilities to model economics and financial variables

Andrea Monaco[1] · Adamaria Perrotta[1] · Joseph Mulligan[2]

Received: 9 August 2022 / Accepted: 29 June 2023 / Published online: 13 July 2023

Abstract
In this paper, we propose a methodology to identify relationships between web data and social/economic variables, such as inflation. Our method enables the selection of relevant time series from a large data sample by employing a criterion based on a few hypotheses regarding their dynamics. Specifically, we examine the correlation between web activities and the dynamics of two macroeconomic variables: the unemployment rate and US automotive sales. We demonstrate how changes in the search volume of specific keywords, as measured by corresponding Google Trends data, are reflected in the underlying dynamics of these variables. The findings presented in this paper, along with the versatility of our approach, suggest the potential extension of this study to other economic variables.

Keywords Web dynamics · Web info · Google search · Unemployment rate · Inflation rate · Macroeconomic variables · Consumption variables · Conditional probability

1 Introduction

The recent availability of time series data on the volume of web searches has opened up new possibilities for utilizing web information in forecasting social behavior and economic variables. This represents a significant advancement, backed by statistical significance, in our ability to explore the potential of web data. Compared to traditional sources such as market research or behavioral studies, leveraging web data offers distinct advantages. It provides access to real-time information and allows for the analysis of large population samples. Additionally, the unique characteristics of web information enable the monitoring of dynamic phenomena that are often challenging

✉ Andrea Monaco
andrea.monaco@ucd.ie

[1] School of Mathematics and Statistics, University College Dublin, Belfield, Dublin 4, Ireland

[2] Department of Mathematics, Imperial College London, Exhibition Rd, South Kensington, London SW7 2BX, UK

to measure using conventional methods. This application is commonly referred to as nowcasting in the literature (Giannone et al. (2008); Castle et al. (2009)).

In this context, the increased availability of new indices related to web activity, such as Google Trends (I_j), i.e., the index of the volume of web searches published by Google [1], has led to a surge in publications investigating the relationship between web data and the dynamics of social events and economic/financial variables (Choi and Varian (2012), Choi and Varian (2009); Ginsberg et al. (2009); Preis et al. (2013); Huang et al. (2020); Askitas and Zimmermann (2009); Shimshoni et al. (2009); Desai et al. (2012); Carneiro and Mylonakis (2009); Pelat et al. (2009)). This new source of information represents a significant opportunity to support economic analysis, and web searches have been successfully used to model both macroeconomic variables (Koop and Onorante (2019); Kholodilin et al. (2009)) and private consumption (Vosen and Schmidt (2011); Goel et al. (2010)).

Macroeconomic variables, including the unemployment rate and GDP, along with private consumption indexes, have been extensively researched in the field of economics. These variables play a vital role in predicting one of the most significant economic indicators, namely inflation (Mankiw (2001)). Inflation is closely monitored by central banks as it informs their decision-making process in effectively managing the economy. Despite the considerable attention given to this topic, the selection of reliable models to accurately describe inflation remains an ongoing challenge. Traditional economic studies on inflation primarily focus on analyzing macroeconomic variables such as the unemployment rate and GDP (Blanchard (2020)). Modern theories, utilizing nowcasting techniques, concentrate on measuring private consumption indexes (Li et al. (2015); Seabold and Coppola (2015); Reinsdorf and Schreyer (2020)). These approaches aim to provide more timely and accurate assessments of inflation dynamics.

Among these studies, only a small fraction of the existing literature on social and economic events utilizes web data, (specifically Google Trends), to identify crucial web searches for forecasting or nowcasting techniques. The selection of relevant searches, in fact, poses a challenge due to the vast amount of data and various assumptions involved. Most studies adopt either aggregate data, which includes searches related to the same category of interest, or specific word searches. In the latter case, assumptions about the dependence of the search on the social phenomenon under study are made in advance, without prior data analysis.

To contribute to research on social and economic events based on web information and address the economic puzzle, we propose a method that combines classical and modern approaches. This method aims to identify web information that is sensitive to both a macroeconomic variable, such as the US unemployment rate, and a consumption variable, such as US car sales. Additionally, we apply our method to variables that previous studies have shown can benefit from the analysis of web data for forecasting purposes.

We have developed a method for identifying web searches that exhibit sensitivity to a specific phenomenon represented by a time series U within a large dataset of web searches. Our method involves the selection of relevant information for modeling the

[1] https://trends.google.com/trends/.

reference phenomenon through a systematic scan of the data and the application of an objective criterion for information selection, based on a minimal set of assumptions. To accomplish this, we measure the conditional probability of changes in U given changes in search volumes of a particular term (ΔI_j) across various time windows. This allows us to identify web searches that demonstrate sensitivity to the dynamics of U.

Our method involves developing a systematic approach to select relevant information for modeling the reference phenomenon U within a large dataset of web searches. This is achieved through a scan of the data and the application of an objective criterion for information selection based on a minimal set of assumptions. To accomplish this, we measure the conditional probability of changes in the reference phenomenon given changes in the search volumes of specific terms ΔI_j across different time windows. This enables us to identify web searches that demonstrate sensitivity to the dynamics of the reference variable U.

By analyzing the semantic and logical relevance of these terms in relation to the reference phenomenon, we can select a subset of Google searches to develop more sophisticated models that explore the connection between I_j and U. Our method offers advantages as it leverages a crucial parameter, namely the conditional probability, to differentiate searches with the highest predictive power for a simple forecasting model. This sets it apart from other selection methods such as the Granger test or the Wald–Wolfowitz test.

To validate our hypothesis regarding conditional volume changes across various time windows, we conducted an analysis using a diverse sample of 512 web searches. We compared the observed dynamics of volume searches with those generated by a pure random process in order to identify words that exhibit both a strong semantic/logical connection with the reference phenomenon and a statistically significant relationship with its dynamics. To confirm the statistical significance of the selected words, we examined the complete probability distribution of our sample, with the conditional probability of the chosen words located at the extreme percentiles of the distribution.

Finally, we explored the sensitivity of each search term across different time horizons, providing valuable insights into the underlying mechanisms that drive the mutual influence between the dynamics of web information and the specific social/economic event.

2 Methodology

In this section, we explore how the dynamics of the search volume index I_j for specific words on the Google search engine can provide insights into the dynamics of social and economic variables represented by a reference time series U. We begin by assuming that both I_j and U follow the same discrete time structure, denoted as $t_1, .., t_i, .., t_N$. Our hypothesis tests whether a change in the reference signal U at time t_i, represented as $\Delta U(t_i) = U(t_i) - U(t_{i-1})$, is followed or preceded by systematic changes in the search volume of specific terms over a time horizon τ, denoted as $\Delta I_j(t_{i-s}, n_\tau) =$

$I_j(t_i) - \hat{I}_j(t_{i-s}, n_\tau)$. In this equation, the index j represents the j-th selected word, and $\hat{I}_j(t_{i-s}, n_\tau) = \frac{1}{n_\tau}\sum_{k=i-n_\tau-s}^{i-s} I_j(t_k)$ calculates the average search volume over the time horizon τ, while n_τ corresponds to the number of time steps within τ. By increasing n_τ, we enlarge the reference time window lag, including the lag itself, which determines the starting point in the past.

The search volume data for a specific word are obtained from the corresponding Google Index I_j, which indicates the number of searches conducted on that word within a monthly period. The data are normalized to a value of 1.0 for the period with the highest levels and are available on a monthly basis starting from January 1, 2004[1].

To investigate the mutual influence between I_j and U, we assume only two possible behaviors: a change in their levels in the same direction or a change in opposite directions. We introduce a parameter, θ, to mathematically describe these cases, setting $\theta = 1$ for the first occurrence and $\theta = -1$ for the second. Once the value of θ is determined, the corresponding hypothesis regarding the mutual relationship between I_j and U is satisfied if the following inequality holds:

$$\Delta U(t_i) \cdot \Delta I_j(t_{i-s}, n_\tau) \cdot \theta \geq 0. \tag{1}$$

Thus, to test the validity of one of the two hypotheses ($\theta = \pm 1$) over the time interval $[0, t_N]$, we calculate the probability ($\hat{P}$) that the condition expressed in (1) is met. In other words, we compute:

$$\hat{P}(\Delta U \cdot \Delta I_j(s, n_\tau)\theta \geq 0, t_N) = \frac{1}{N}\sum_{i=0}^{N} \delta(\Delta U(t_i) \cdot \Delta I_j(t_{i-s}, n_\tau) \cdot \theta \geq 0), \tag{2}$$

Here, $\hat{P}$ represents the probability that changes in ΔU are systematically followed ($\tau > 0$) or preceded ($\tau < 0$) by changes in $\Delta I_j(s, n_\tau)$ over the time interval $[0, t_N]$.

It is important to note that although $\hat{P}$ depends on t_N, if the behavior of $\hat{P}$ is stationary, then for a fixed j, it can be shown that $\hat{P}$ converges to a well-defined value: $\hat{P} \to \tilde{P}$ as $N \to \infty$, making $\hat{P}$ independent of t_N.

Furthermore, if ΔI_j exhibits purely random behavior, then $\hat{P}$ converges to 0.5. This indicates that there is no preferred configuration between the two possible relative changes of ΔI_j and ΔU, and the two variables do not have any significant mutual influence. Therefore, by monitoring the behavior of $\hat{P}$ and its convergence toward the random level of 0.5, we can assess the reliability of one of the two hypotheses ($\theta = \pm 1$).

To test the randomness of the marginal contribution to the conditional probability $\hat{P}$, we compared the contribution of a specific Google search term ΔI_j to that of a purely random process. The selection criterion does not make any specific assumptions about the distribution of the conditional probability.

Finally, the knowledge of $\tilde{P}$ (i.e., $\hat{P}$ measured over long time series assuming stationary behavior) for changes of I_j in the past (i.e., $\Delta I_j(s, n_\tau < 0)$) can be utilized to make predictions about future changes of U. In fact, $\tilde{P}$ can be expressed as a conditional probability conditioned on one of two possible behaviors of $\Delta I_j(s, n_\tau <$

0), namely $\Delta I_j(s, n_\tau)\theta \geq 0$ or $\Delta I_j(s, n_\tau)\theta < 0$:

$$\tilde{P}(\Delta U \cdot \Delta I_j(s, n_\tau)\theta \geq 0) = \begin{cases} \tilde{P}(\Delta U \geq 0 \mid \Delta I_j(s, n_\tau)\theta \geq 0), \ \mathit{if} \Delta I_j(s, n_\tau) \geq 0, \\ \tilde{P}(\Delta U < 0 \mid \Delta I_j(s, n_\tau)\theta < 0), \ \mathit{if} \Delta I_j(s, n_\tau) < 0. \end{cases}$$

Based on the previous relationship, it becomes possible to forecast the direction of future changes in U, ΔU, by leveraging the knowledge of $\tilde{P}$ and the historical information of $\Delta I_j(s, n_\tau < 0)$. This approach falls under the category of non-parametric predictive-causality testing as it avoids making assumptions about the underlying mechanisms of influence between I_j and U. The main distinction between this method and traditional predictive-causality tests, like the Granger test (Granger (1969)), lies in the fact that the latter relies on a linear regression model to depict the stochastic processes. To use it, we need to test some properties on the data we are working with, like that the time series are stationary and that they are linearly related. Our research shows that we can ignore these assumptions and still get good results, making our method more versatile than other approaches. In fact, to apply our method, we do not need to make any assumptions about the stationary nature of the time series involved or their linearly related (meaning that one variable affects linearly the other). The effectiveness of our method, compared to the Granger test, is confirmed when both tests are applied to the same dataset. The highest-ranked selected words differ between the two methods, and the words identified by the Granger test lack the semantic relevance of the ones we have identified.

2.1 Data preprocessing

To utilize the described method, data preprocessing and cleaning are necessary. One important consideration is handling missing or constant values within the time series for I_j and U.

Regarding missing data in I_j, our approach depends on their distribution within the time series. If the majority of missing data are located at the beginning or end of the series, we exclude that specific period or the entire time series from our analysis. If the missing data are scattered throughout the series, we exclude the entire time series only if the ratio of missing data to the total time series exceeds 20%. This ensures the reliability of the calculated conditional probabilities. After excluding the missing data, any remaining gaps are filled using linear interpolation.

Regarding constant values in the time series (including the reference signal time series U), we adopt a neutral probability hypothesis. This means that when no changes are detected in the I_g/U signal, we replace the constant values with randomly generated positive or negative changes. This approach does not impact the non-random long-range probability component and remains conservative in terms of conditional probability levels. On average, positive changes are offset by negative changes, while still considering this data in our analysis.

3 Results

In this section, we examine $\hat{P}$ for extensive word samples (512 words) in relation to two indices: the US automotive sales index and the unemployment rate. We observe that certain web searches, which exhibit semantic or logical relevance to the studied activities, demonstrate a higher $\hat{P}$ level than what would be expected from a purely random pattern of I_j. These searches exhibit a leading behavior compared to the dynamics of the investigated phenomenon.

3.1 US automotive sale market

The methodology described in the previous section was applied to the US automotive sales market index using the time series data provided by the US Census Bureau.[2] The choice of this time series was driven by to primary reasons: firstly, it was due to its significance in forecasting the inflation rate, as mentioned in the introduction; secondly, to evaluate the methodology's reliability. The relevance of Google searches in modeling US automotive sales is supported by a study conducted by Choi and Varian (2012), which demonstrates how such information can enhance forecasting accuracy. However, the mentioned study did not employ any quantitative criteria for selecting key web searches. In contrast, our method enables the retrieval of a specific set of reference Google searches from a large sample, facilitating the prediction of the dynamics of the target variable (in this case, US automotive sales) and resulting in improved predictive performance.

The conditional probability of changes in the search index I_j given changes in the automotive sales index U was computed for a set of 512 terms. These terms were selected from the most searched terms in the US on January 12, 2020. Specifically, we chose the top 25 searches for each of the 25 categories of interest defined on https://www.google.it/trends within the geographical area of the US. We then eliminated synonymous web searches from this sample.

Using the time series data for I_j and U, we computed the probability (2) for different values of n_τ covering the entire available history of I_j up until January 1, 2020. The values of τ were chosen as multiples of the time step of the reference time series U, which, in the case of automotive sales, was monthly. Finally, the levels of $\hat{P}$ for each word, corresponding to different values of τ, were sorted to identify terms that not only exhibited clear relevance to the analyzed phenomenon but also had $\hat{P}$ values significantly deviating from the random level of 0.5.

Table 1presents the top 20 terms ranked according to their $\hat{P}$ values. The terms that are highly relevant to the automotive sales phenomenon are highlighted in bold. It is important to note that different words may have the same $\hat{P}$ values due to the discrete and limited number of time steps in the U and I_j time series. Thus, a higher resolution of information for both U and I_j would be required to enhance the resolution of $\hat{P}$ dynamics.

Notably, searches such as *audi* and *mercedes* exhibit significant values of $\hat{P}$ over a period of 1–3 months in the past (i.e., $n_\tau = 1, 2, 3$). These web searches indicate

[2] https://www.census.gov/retail/marts/www/adv441x0.txt.

Table 1 $\hat{P}(\Delta U \cdot \Delta I_j(s, n_\tau)\theta \geq 0, t_N)$ with U: US Car sales, $t_N \simeq 16$ years

N	$n_\tau = 1$		$n_\tau = 2$		$n_\tau = 3$	
1	Weather radar	0.6030	Fire	0.6010	Uti	0.5939
2	Ice cream	0.6030	Homes for sale	0.5959	Navy	0.5939
3	Zip code	0.5879	Apartments for rent	0.5959	**Audi**	0.5888
4	Homes for sale	0.5879	**Mercedes**	0.5858	Haircut	0.5888
5	Fedex	0.5829	Cleveland	0.5858	Tires	0.5888
6	Cvs	0.5778	Autozone	0.5858	House	0.5888
7	**Mercedes**	0.5728	Apartment	0.5858	**Mercedes**	0.5837
8	Cleveland	0.5728	Rv	0.5808	Marriott	0.5837
9	Bus	0.5728	Real estate	0.5808	Hilton	0.5837
10	Bath and body	0.5728	Nhl	0.5808	Hair	0.5837
11	Restaurants	0.5678	Houses	0.5808	Fire	0.5837
12	Mlb	0.5678	Restaurants	0.5757	Ocean	0.5786
13	Tiempo	0.5628	Rent	0.5757	Bus	0.5786
14	Solitaire	0.5628	Panera	0.5757	Petco	0.5736
15	Rv	0.5628	House for sale	0.5757	Jeep	0.5736
16	Real estate	0.5628	Fanfiction	0.5757	Hyundai	0.5736
17	Marriott	0.5628	Chipotle	0.5757	Houses	0.5736
18	Gmc	0.5628	Cat	0.5757	Hotels	0.5736
19	Fire	0.5628	Car rental	0.5757	Homes	0.5736
20	Autozone	0.5628	Vet	0.5707	Harbor	0.5736

future changes in U, as there is a substantial probability of changes in U conditioned on changes in I when $\tau < 0$ is set.

Figure 1 illustrates the progression of $\hat{P}$ leading up to January 1, 2020, for all the words (represented by red lines). The blue line corresponds to the trajectory associated with the search term *audi*, while the green lines indicate the levels of paths that exhibit a purely random behavior for the index I_j. The solid line centered around 0.5 represents the average value, while the dashed line represents the confidence levels corresponding to $\pm\sigma$. These random levels were estimated by simulating 1000 paths, where at each time step t_i, the change in the level of I_j was obtained through random extraction.

The narrowing of the distribution of the red path as t_i increases indicates a decrease in statistical errors during the computation of $\hat{P}$. Figure 1 also demonstrates that as the statistical significance of the calculated probability rises, the distribution of $\hat{P}$ follows nearly symmetrical paths toward levels that fall within the range associated with the $\pm\sigma$ confidence levels of a hypothetical pure random process. This clearly indicates the non-random nature of the *audi* search term, as the corresponding path on January 1, 2020, is situated at extreme values, far from those corresponding to the confidence level $\pm\sigma$ predicted by a pure random process.

Similar non-random behavior is observed in the web search related to *mercedes* for different values of τ (specifically, $n_\tau = 2$ and $n_\tau = 3$), as illustrated in Fig. 2.

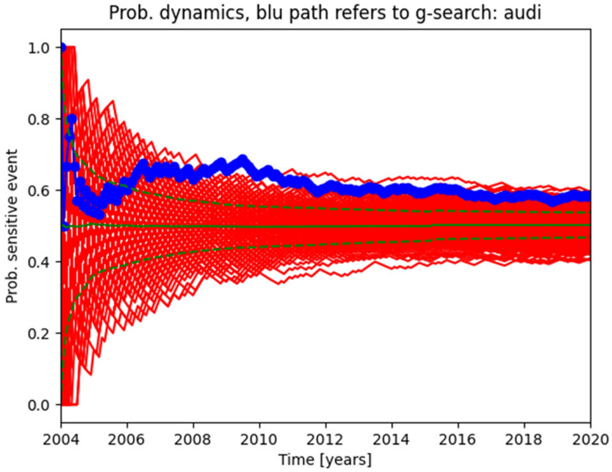

Fig. 1 Red lines indicate $\hat{P}(\Delta U \cdot \Delta I_j(s, n_\tau)\theta \geq 0, t_N)$ with $n_\tau = 3, \theta = 1, U =$US Automotive sales, the blue line is the path associated to the *audi* search term, the solid green line represents the mean value associated to the random behavior of I_j, while dashed green lines indicate for the same random process the confidence level $\pm\sigma$

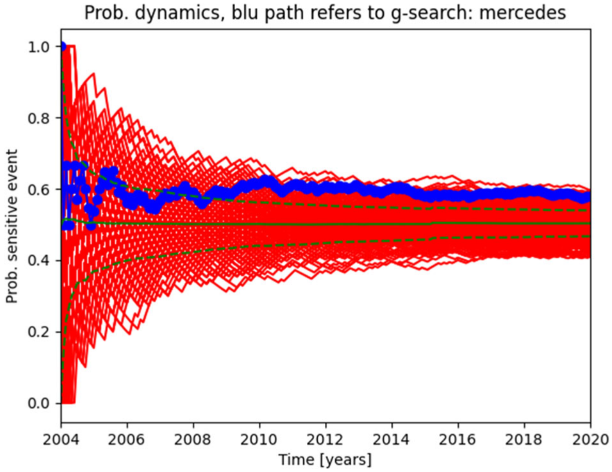

Fig. 2 Red lines indicate $\hat{P}(\Delta U \cdot \Delta I_j(s, n_\tau)\theta \geq 0, t_N)$ with $n_\tau = 2, \theta = 1, U =$US Automotive sales, the blue line is the path associated to the *mercedes* search term, the solid green line represents the mean value associated to the random behavior of I_j, while dashed green lines indicate for the same random process the confidence level $\pm\sigma$

In Fig. 3, we present the distribution of $\hat{P}$ calculated for a time period of approximately 16 years, using the time series data from January 1, 2004, to January 1, 2020, for all 512 terms. The distribution exhibits a symmetrical shape centered around the 0.5 level. However, the circular symbols corresponding to web searches that are clearly relevant (such as *audi* and *mercedes* for $n_\tau = 3$, and *mercedes* for $n_\tau = 2$) are distinctly

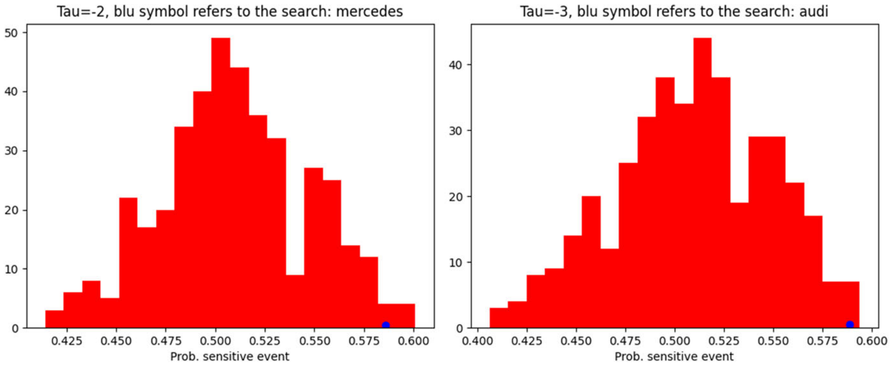

Fig. 3 Distribution of $\hat{P}(\Delta U \cdot \Delta I_j(s, n_\tau)\theta \geq 0, t_N)$ with $\theta = 1$, U =US Automotive Sales, $t_N \simeq 16$ years, the blue circular symbols represent web searches: *mercedes*, *audi*

positioned on the tails of the $\hat{P}$ distribution. This deviation from randomness suggests that these searches play an active role in the dynamics of US automotive sales.

In conclusion, our study has demonstrated a consistent relationship between changes in the volume of specific web searches and shifts in US automotive sales trends. We have identified certain selected words that actively contribute to the mutual influence between web information and US automotive sales. This finding aligns with a previous study conducted by Choi et al. (2012) and suggests the potential for developing improved forecasting models using logically and semantically related Google searches as predictors.

It is important to note that our approach for selecting relevant Google searches was based on a systematic analysis of a large sample, without any predetermined hypotheses. However, to assess the generalizability of our methodology, we deliberately excluded specific terms directly related to car sales or their meanings from the sample. Including such searches in future investigations could offer valuable insights into the underlying mechanisms of influence between public web opinion and US automotive sales, allowing for further refinement and enhancement of our methodology.

3.2 Unemployment rate

In this section, we investigate the interplay between the unemployment rate and search volumes of specific keywords. The relationship between the unemployment rate and web information dynamics has been extensively explored in the literature, with various models applied to forecast changes in unemployment rates using Google search index data (e.g., Choi and Varian (2009); Askitas and Zimmermann (2009); D'Amuri and Marcucci (2010); McLaren (2011); D'Amuri and Marcucci (2012); Fondeur and Karame (2013)). However, a systematic examination that scans a large sample of searches to select relevant data for modeling unemployment rates, without relying on preconceived hypotheses, is currently lacking.

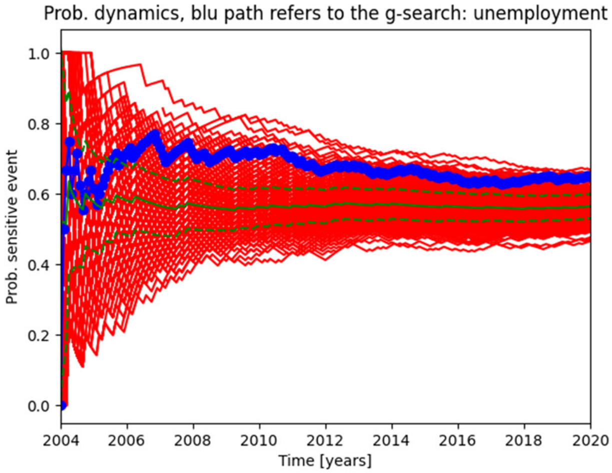

Fig. 4 Red lines indicate $\hat{P}(\Delta U \cdot \Delta I_j(s, n_\tau)\theta \geq 0, t_N)$ with $n_\tau = 1, \theta = 1, U =$ US unemployment rate, the blue symbols indicate pattern relative to term *unemployment*, the green line '-' represents the mean value associated to the random behavior of I_j, while dashed green lines ' - -' indicate for the same random process the confidence level $\pm\sigma$

To investigate this relationship, we utilized time series data provided by the US Bureau of Labor Statistics,[3] specifically focusing on non-seasonally adjusted data (Franses and De Bruin (2000)). Our analysis involved a sample of 512 terms selected based on their relative Google search index. Similar to the approach used for the US automotive sales market index, we identified the top 25 searched terms within each of the 25 categories defined by Google on their website *https://www.google.it/trends*. We then removed synonymous searches to ensure data consistency.

For each keyword, we computed the probability (2) by varying the values of n_τ. The selection of n_τ values was based on multiples of the time step of the unemployment rate series U, which has a monthly periodicity. Consequently, the search volume series I_j was averaged on a monthly basis.

Table 2 presents the values of $\hat{P}$ for each searched word, corresponding to different n_τ values. The keywords highlighted in bold are those that exhibit both a logical/semantic connection with the unemployment phenomenon and a substantial deviation of $\hat{P}$ from the value of 0.5. Notably, the reported $\hat{P}$ values in the table were computed using data spanning from January 1, 2004, to January 1, 2020.

Examining Table 2, we can clearly identify two noteworthy keywords: *unemployment* and *indeed jobs*.

The non-random nature of various words is illustrated in Figs. 4, 5, and 6, where the blue circular symbols corresponding to sensitive searches are positioned on the tails of the $\hat{P}$ distribution. This distribution was computed across all searches, considering different values of τ, and utilizing Google search volume data spanning almost ten years (approximately $t_N \simeq 16$ years).

[3] https://data.bls.gov/.

Table 2 $\hat{P}(\Delta U \cdot \Delta I_j(s, n_\tau)\theta \geq 0, N)$ with U: US unemployment rate, $t_N \simeq 16$ years

N.	$n_\tau = 1$		$n_\tau = 2$		$n_\tau = 3$	
1	**unemployment**	0.6985	India	0.6919	India	0.6904
2	Windows 7	0.6985	Facebook	0.6919	Facebook	0.6904
3	Nfl scores	0.6985	Windows 7	0.6869	Torrent	0.6853
4	Adobe	0.6985	Chat	0.6869	Face	0.6802
5	Map	0.6935	Nfl scores	0.6818	Colleges	0.6802
6	Mocospace	0.6884	Twitter	0.6768	Chat	0.6802
7	India	0.6884	Pof	0.6768	YouTube	0.6751
8	Plenty	0.6784	Jokes	0.6717	Mocospace	0.6751
9	Nintendo	0.6784	Chase bank	0.6717	Translator	0.6701
10	Myspace	0.6784	**indeed jobs**	0.6667	Mario	0.6701
11	Donald trump	0.6784	Plenty	0.6667	Exercise	0.6701
12	2012	0.6784	Usps tracking	0.6667	Scrabble	0.6650
13	Worldstarhiphop	0.6734	Government	0.6667	Java	0.6650
14	United	0.6734	Drudge	0.6667	Inspirational quotes	0.6650
15	Ixl	0.6734	**unemployment**	0.6616	Hotmail	0.6650
16	Halloween	0.6734	Xbox 360	0.6616	Cool math	0.6650
17	Chipotle	0.6734	Snake	0.6616	Traductor	0.6599
18	Vrbo	0.6683	Petco	0.6616	Spanish to English	0.6599
19	Twitter	0.6683	Periodic table	0.6616	Netflix	0.6599
20	Cheap flights	0.6683	Fb	0.6616	Irs	0.6599

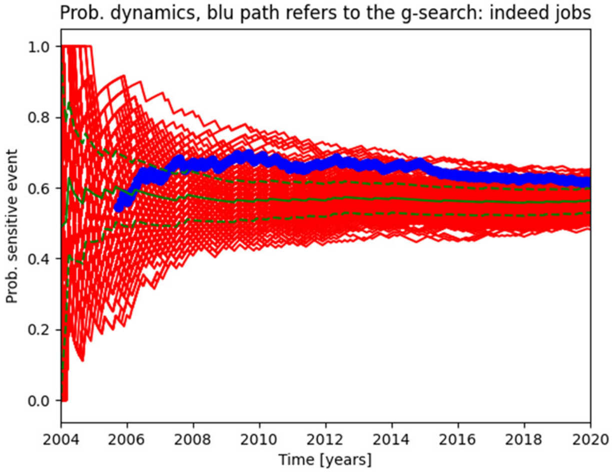

Fig. 5 Red lines indicate $\hat{P}(\Delta U \cdot \Delta I_j(s, n_\tau)\theta \geq 0, t_N)$ with $n_\tau = 2$, $\theta = 1$, $U =$ US unemployment rate, the blue symbols indicate pattern relative to term *indeed jobs*, the green line '-' represents the mean value associated to the random behavior of I_j, while dashed green lines ' - -' indicate for the same random process the confidence level $\pm\sigma$

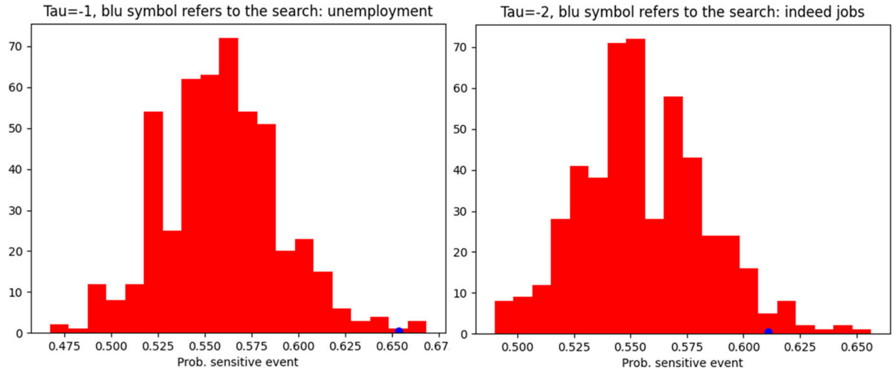

Fig. 6 Distribution of $\hat{P}(\Delta U \cdot \Delta I_j(s, n_\tau)\theta \geq 0, t_N)$ with $\theta = 1$, $U =$US Unemployment rate, $t_N \simeq 16$ years, the blue circular symbols represent web searches: *unemployment* and *indeed jobs*

Analyzing the $\hat{P}$ values of the chosen search terms, we observe that the absolute changes in U and I_j—denoted as ΔU and ΔI_j, respectively—may exhibit different signs for past time intervals ($\tau < 0$). Specifically, an increase in the past search volume for the term *unemployment* compared to the present level is associated with an increase in the unemployment rate. This behavior suggests that an upward trend in the search volume for *unemployment* serves as a predictor for an impending rise in the unemployment rate. Thus, similar to the case of US automotive sales, we find that variations in the past search volumes of specific keywords closely linked to unemployment precede changes in the level of the unemployment rate.

Our findings align with previous literature on forecasting the unemployment rate (Choi and Varian (2012); Choi and Varian (2009); Askitas and Zimmermann (2009);

D’Amuri and Marcucci (2010); McLaren (2011); D’Amuri and Marcucci (2012); Fondeur and Karame (2013)).

Similar to the US automotive sales analysis, we deliberately excluded specific terms related to the influence of web information on employment from our general Google search sample.

4 Discussion

This paper presents a methodology for identifying Google searches that reflect the dynamics of social and economic events. One of the main challenges in research utilizing web information for modeling such events is selecting relevant data that are associated with the target variable. To address this challenge, our systematic approach focuses on two complex aspects: analyzing complex data samples and applying an objective selection criterion based on a few hypotheses regarding the sample’s dynamic behavior.

To assess the effectiveness of our proposed method, we conducted an analysis using two reference time series: the unemployment rate (macroeconomic variable), and US automotive sales (consumption variable). These variables were chosen for their significance in modeling inflation rates and for comparison with existing literature. By considering the entire search volume history up to January 1, 2020, we computed the probability of search volumes for specific words moving in accordance with or against changes in the reference time series over various time periods. Our analysis encompassed 512 terms.

The selection method proposed in this study is closely linked to the nature of the analyzed data, specifically the changes in search volumes within a predefined time window with a time delay. If the data were different, such as absolute levels or percentage changes in volumes, the discrimination procedure would need to be adjusted. In such cases, the asymptotic level of conditional probability would not be 0.5, and the comparison of search dynamics with a random process would require reformulation, along with modifying the data cleaning process outlined in Sect. 2.1.

In conclusion, our proposed method combines two key elements: analyzing the dynamics of conditional probability across different time delays and employing a selection criterion based on simulating a pure random signal. To validate the statistical significance of the selected words, we examined the entire probability distribution of our sample. We discovered that the conditional probability associated with the selected words fell within the extreme percentiles of the overall distribution.

This analysis enabled us to identify relevant searches that not only demonstrated semantic/logic relevance to the studied phenomena but also exhibited a non-random change in direction relative to the reference time series. By employing both criteria, we successfully identified words capable of predicting the two investigated phenomena. Consequently, by testing a statistical hypothesis regarding the mutual influence of search volumes on the reference time series, we gained insights into the underlying mechanisms governing the correlation between web opinion and social/economic events.

Our findings align with the existing literature. We have discovered that the selected sensitive searches exhibit semantic and logical relevance to both US automotive sales and unemployment rates, similar to the terms previously identified in studies that enhance forecasting models (Choi and Varian (2009); Askitas and Zimmermann (2009); D'Amuri and Marcucci (2010); McLaren (2011); D'Amuri and Marcucci (2012); Fondeur and Karame (2013)).

As mentioned in the introduction, the selection of relevant searches poses challenges due to the vast amount of data and underlying assumptions. Many studies utilize aggregate web search data or rely on pre-established hypotheses regarding the dependence on a social phenomenon, without conducting prior data analysis.

To apply the proposed method, we recommend starting with a general criterion to create a reference sample of Google searches. We demonstrate a methodology for extracting only the search terms that are sensitive to a specific phenomenon from a large sample of web searches. Additionally, specific terms can be added to the sample based on their semantic and logical relevance to a particular reference signal. This addition could provide further insights into the relative influence of these terms on the same reference event.

The methodology described here can be valuable in selecting web data for modeling economic and financial variables, such as inflation, with unemployment and US car sales serving as foundational components.

The robustness of the methodology lies in its ability to perform well on large data sets and provide a scale to assess the relevance of each search term. However, it is important to note that the performance of the method depends on the length of the selected time period and the time resolution of the reference signal. Having access to data with a wide time window and a time sampling similar to that of Google searches would greatly enhance the method's performance.

Acknowledgements We acknowledge Dr Pierluigi Casale for the insightful discussions on the data analysis and Dr Graziano Moramarco for the macroeconomic ones.

Funding Open Access funding provided by the IReL Consortium

Declarations

Conflict of interest Author Andrea Monaco declares that he has no conflict of interest. Author Adamaria Perrotta declares that she has no conflict of interest. Author Joseph Mulligan declares that he has no conflict of interest.

Human and animals participants This article does not contain any studies with human participants or animals performed by any of the authors.

References

Askitas N, Zimmermann KF (2009) Google econometrics and unemployment forecasting. Appl Econ Q 55:107–120

Blanchard O (2020) Macroeconomics, 8th edn. Global edition, Pearson

Carneiro HA, Mylonakis E (2009) Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis 49(10):1557–1564

Castle J, Fawcett N, Hendry D (2009) Nowcasting is not just contemporaneous forecasting. Natl Inst Econ Rev 210(1):71–89

Choi, H. and Varian, H. (2009). Predicting initial claims for unemployment benefits

Choi H, Varian H (2012) Predicting the present with google trends. Econ Record 88:2–9

D'Amuri F and Marcucci J (2010) Google it! forecasting the us unemployment rate with a google job search index. Fondazione Eni Enrico Mattei Working Papers 421

D'Amuri, F. and Marcucci J (2012) The predictive power of Google searches in forecasting unemployment. Number 891. Banca d'Italia

Desai R, Hall A, Lopman B, Shimshoni Y, Rennick M, Efron N, Matias Y, Patel M, Parashar U (2012) Norovirus disease surveillance using google internet query share data. Clin Infect Dis 55(8):e75–e78

Fondeur Y, Karame F (2013) Can google data help predict french youth unemployment? Econ Modell 30:117–125

Franses P, De Bruin P (2000) Seasonal adjustment and the business cycle in unemployment. Studies Nonlinear Dyn Econ 4(2):1558–3708

Giannone D, Reichilin L, Small D (2008) Nowcasting: the real-time information content of macroeconomic data. J Monet Econ 55:665–676

Ginsberg J, Mohebbi M, Patel R, Brammer L, Smolinski M, Brilliant L (2009) Detecting influenza epidemics using search engine query data. Nature 457:1012–1014

Goel S, Hofman J, Lahaie S, Pennock D, Watts D (2010) Predicting consumer behavior with web search. In: Proceedings of the national academy of sciences of the United States of America PNAS 107(41):17486–17490

Granger C (1969) Investigating casual relations by economic models and cross-spectral methods. Econometrica 37:24–36

Huang M, Rojas R, Convery P (2020) Forecasting stock market movements using google trend searches. Empir Econ 59:107–120

Kholodilin K, Podstawski M, Siliverstovs B and Bürgi C (2009) Google searches as a means of improving the nowcasts of key macroeconomic variables. Discussion Papers

Koop G, Onorante L (2019) Macroeconomic nowcasting using google probabilities. Top Identif Ltd Depend Var, Partial Observability, Exp Flexible Model: Part A, Adv Econ 40A:17–40

Li X, Shang W, Wang S, Ma J (2015) A MIDAS modelling framework for Chinese inflation index forecast incorporating google search data. Electron Commer Res Appl 14(2):112–125

Mankiw NG (2001) The inexorable and mysterious tradeoff between inflation and unemployment. Econ J 111:45–61

McLaren N (2011) Using internet search data as economic indicators. Bank England Quarterly Bull 51(2):134–140

Pelat C, Turbelin C, Valleron A-J (2009) More diseases tracked by using google trends. Emerg Infect Dis 15(8):1327–1328

Preis T, Moat H, Stanley H (2013) Quantifying trading behaviour in financial markets using google trends. Sci Rep 3:1684

Reinsdorf M, Schreyer P (2020) Measuring consumer inflation in a digital economy. Measur Econ Growth Prod Found KLEMS Prod Model Ext 1:339–362

Seabold, S. and Coppola, A. (2015). Nowcasting prices using google trends: an application to central america. World Bank Policy Research Working Paper (7398):112–125

Shimshoni Y, Efron N, and Matias Y (2009) On the predictability of search trends. Google Research Blog

Vosen S, Schmidt T (2011) Forecasting private consumption: survey-based indicators vs. google trends. J Forecast 30(6):565–578