# Assessing causality among topics and sentiments: the case of the G20 discussion on Twitter

**Mauro Fonseca[1], Fernando Delbianco[2,3], Ana Maguitman[1,4], and Axel J. Soto[1,4]**

## Abstract

Although the identification of topics and sentiments from social media content has attracted substantial research, little work has been carried out on the extraction of causal relationships among those topics and sentiments. This article proposes a methodology aimed at building a causal graph where nodes represent topics and emotions extracted from social media users' posts. To illustrate the proposed methodology, we collected a large multi-year dataset of tweets related to different editions of the G20 summit, which was locally indexed for further analysis. Topic-relevant queries are crafted from phrases extracted by experts from G20 output documents on four main recurring topics, namely government, society, environment & health, and economics. Subsequently, sentiments are identified on the retrieved tweets using a lexicon based on Plutchik's wheel of emotions. Finally, a causality test that uses stochastic dominance is applied to build a causal graph among topics and emotions by exploiting the asymmetries of explaining a variable from other variables. The applied causality discovery process relies on observational data only and does not require any assumptions of linearity, parametric definitions or temporal precedence. In our analysis, we observe that although the time series of topics and emotions always show high correlation coefficients, stochastic causality provides a means to tell apart causal relationships from other forms of associations. The proposed methodology can be applied to better understand social behavior on social media, offering support to decision and policy making and their communication by government leaders.

## Keywords

Causality, Sentiment Analysis, Stochastic Dominance, Social Media, G20

## Introduction

Social concerns typically give rise to strong emotions, which are often shared on social media. As a consequence, the identification of emotions on social media is increasingly being adopted to help investigate and understand social, governmental, economical, and environmental problems from different perspectives. This can be accomplished by using social media data either to augment or as an alternative to opinion surveys [1].

Topics discussed on Twitter can spark emotions, and in turn, emotions can cause certain topics to be more actively discussed. Also, some emotions can give rise to other emotions, and some topics can generate discussion on other topics. These observations suggest the existence of a causal graph that can be learned from social media. In such a graph, nodes represent topics and emotions, while edges represent causal relations.

Although the study of the concept of causality is a central and long-standing issue in the field of econometrics [2], the relatively recent availability of large volumes of data opens up new opportunities to conjecture on possible causal relations. A causal graph of topics and emotions derived from social media offers an analytics data-driven mechanism. Causal graphs are useful for prediction, but most importantly, they allow to better understand the cause-effect connections among variables [3]. This cause-effect knowledge leads to more explainable models, which are of paramount value for understanding social, governmental, economical, and environmental problems.

A causal graph of topics and emotions allows us to assess whether an increase in the volume of posts and replies associated with a topic (e.g., the environment) has an effect on specific emotions (e.g., fear or anticipation). To illustrate the approach proposed in this article, we analyze a collection of tweets associated with G20 editions collected from three different time periods. Twitter posts in relation to G20 summits provide an excellent case study, as several issues such as international financial stability, climate change mitigation, and sustainable development are typically addressed by leaders and citizens. Such an analysis helps assess the flow of topics and their causal relations with emotions about different policies and governments. A practical objective of the proposed approach is to better understand social behavior on social media by analyzing, aggregating and structuring the topics and emotions associated with these data as a causal graph. Moreover, the proposed approach can support decision and policy making and their communication by government leaders.

[1] Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur, Argentina

[2] Departamento de Economía, Universidad Nacional del Sur, Argentina

[3] Instituto de Matemática de Bahía Blanca, INMABB (CONICET-UNS), Argentina

[4] Instituto de Ciencias e Ingeniería de la Computación, ICIC (CONICET-UNS), Argentina

**Corresponding author:**
Ana Maguitman, Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur, San Andrés 800, Bahía Blanca, Buenos Aires, 8000, Argentina. agm@cs.uns.edu.ar

The role of social media in issues related to economics and politics has been extensively investigated in the literature [4, 1, 5]. However, few studies have addressed the problem of sentimental causal rule discovery from text. The application of causality analysis to investigate emotion in texts and their relations to variables of interest has been mainly used in the context of financial prediction [6, 7, 8]. Other approaches have looked into the problem of extracting sentimental causal rules from text [9], identifying sentimental causal relations across time [10] and discovering causal relations between emotions and topical shifts [11].

Although the question of "pure causality" is a philosophical one, the study of "predictive causality" has been central in the field of econometrics for a long time [2, 12, 13, 14, 15]. This topic has attracted increasing attention in computer science and information science where causal discovery [16, 17, 18, 19] is distinguished from causal inference [20, 21, 22, 23, 24]. The goal of the former is to obtain causal knowledge directly from observational data while the latter aims to test whether two variables are related and assessing the impact of one on the other [25].

Econometricians have addressed causal discovery mainly by methods derived from the *Granger Causality test* [2], which is based on two principles: (1) a cause occurs before the effect, and (2) the cause produces unique changes in the effect, so past values of the cause help predict future values of the effect. However, requirement (1) is too restrictive for certain applications. For instance, news or social media may mention two events that represent cause and effect in a single announcement, in different announcements with no or insignificant time intervals in between, or even in reverse order. This would require supplementary sources of information to sort events by time and the availability of time series with arbitrarily small time granularity. In light of this limitation, Vinod [26] offers a statistical method that can help determine the direction of causal paths without requiring the cause to occur strictly before the effect. The analysis presented in this article takes advantage of Vinod's causality test to identify causal direction among topics and emotions.

Another issue that usually arises when studying social media data is that the association between two variables is typically investigated using regression and correlation analyses, which are useful in predictive modeling but inevitably leave the issue of causality open to question [27]. This work focuses on the causal association of variables representing topics and emotions, disregarding other types of relationships between two variables.

Our research question focuses on determining whether causality relations exist among topics that were addressed at different G20 summits and discussed on Twitter, as well as the sentiments that social media users expressed in their posts. If there are significant causality relations, then as a derived question, we aim at finding the connections of the resulting causal graph. Once a causal graph is derived, then it is possible to determine which topics or sentiments are more plausible to be the cause or effect of other variables.

The contributions of this work can be summarized as follows:

- A methodology for building causal graphs consisting of nodes representing topics and emotions extracted from a large volume of social media data.

- The application of a causality test that exploits the asymmetries of explaining a variable in terms of the other variables without requiring any assumptions of linearity, parametric definitions or temporal precedence. The causal discovery process can be assessed with observational data only, i.e., without requiring interventional or experimental settings.
- An analysis of the causal relationship existing among the four main recurring G20 topics (government, society, environment & health and economics), eight primary emotions based on Plutchik's theory (anger, fear, sadness, disgust, surprise, anticipation, trust, and joy), and no emotion.

We also make the research data available to allow reproducibility and for other researchers to use. These data include:

- A multi-year dataset with the tweet ids of nearly 50 million tweets related to different G20 editions and the posts from and to the main politicians from the participating countries of the G20 summit.
- Expert-extracted phrases that were regarded as relevant for the four main recurring G20 topics and a list of queries derived from these phrases.

Although the languages used to illustrate the proposed methodology are English and Spanish, and the analyzed topics were associated with the G20 summits, it is important to highlight that all the described methods can be applied to other target languages and topics as long as we have access to (1) expert-generated phrases for the selected domain topics in the target languages and (2) a language-specific emotion lexicon.

The article is structured as follows. In the next two sections, we present related work and the methodology applied to collect the data and to assess causality. In the subsequent section, we present the results and main findings, while in the last section we discuss the conclusions of our work.

## Related work

The interaction between topics and emotions in social media has been mostly analyzed by looking into how sentiments depend on specific topics. For such analyses, topics are represented either explicitly (i.e., using keywords or hashtags) or implicitly (i.e., as hidden or latent topics). The work presented by Meng et al. [28] is an early example of a framework that applies sentiment analysis on explicit topic representations resulting from grouping hashtags. Opinion summaries are generated by integrating the derived topics with sentiment classification towards entities extracted from the collected tweets. Another proposal that takes an explicit approach to represent topics is Social Sentiment Sensor [29]. The proposed system relies on hashtags to detect daily hot topics on Sina Weibo with the purpose of analyzing sentiment distribution towards the identified topics.

An example of implicit topic representation is given by the Hidden Topic Sentiment Model [30], where topic coherence and sentiment consistency are captured from text documents by extracting latent aspects and the corresponding sentiment polarities. Other approaches use a combination of natural language processing techniques and statistical

modeling to extract the sentiments and topics simultaneously from text data and to analyze the relationship between those variables. This gives rise to joint sentiment-topic models as described by Lin and He [31] and subsequent approaches that have been used to quantify the sentiment expressed towards topics identified in a dataset and to understand how sentiments towards different topics change over time [32, 33, 34, 35].

Some recent approaches attempt to capture the dynamics of sentiments on topics. This is the case of Liang et al [36], where the authors propose to use a Dynamic Bayesian Network to model the dynamics and interactions of the sentiment of topics on social media. In another recent example, Pathak et al. [37] dynamically extract topics at the sentence level using online latent semantic indexing with regularization constraints and then apply topic-level attention mechanism in an LSTM network to perform sentiment analysis. Another recent work that analyzes the dynamics of topic-level sentiment monitors the evolution of people's mental states across different topics or events related to coronavirus [38]. However, all these approaches focus on the interactions among topics and emotions without looking at their causal relation, as it is performed in our work.

Previous studies have applied a combination of causal analysis and sentiment analysis on social media data and news. Several authors have used these techniques for stock market or product pricing prediction. For instance, in Smailović et al. [6], Granger causality is applied to show that sentiment polarity based on public opinion on companies and their products collected from Twitter feeds can help predict stock price movements a few days in advance. Granger causality was also used to test whether and how the sentiment of online news articles impacts oil price [7]. The analysis concluded that sentiment series strictly Granger causes the price series, with a predictive lag order of three weeks. Another study has proposed a sentiment analysis engine that works at the phrase level and allows to extract collective expressions from large amounts of texts [8]. This last approach is shown to offer a helpful mechanism for analyzing the trends in a stock market index. Causal relations to understand other phenomena have been also proposed, such as for the analysis of COVID-19 statistics and its influence on attitudes towards tourism on social media [39].

More closely related to our work is the identification of causal relations between social media sentiments and political news, political events and even politicians' posts. It has been shown that words used by politicians have an effect on stock markets [40]. Likewise, public sentiment on a political topic, such as BREXIT, may influence Great Britain's currency exchange rate and its FTSE 100 index [41]. In a similar way, Scaramozzino et al. [42] use transfer entropy to quantify causality relations between daily stock price and daily social media sentiment for the top 50 companies in the Standard & Poor index during a two-year period. The analysis reveals that there is a causal flux of information that links those companies. The largest fraction of significant causal links is between prices and between sentiments, but there is also significant causal information that goes both ways from sentiment to prices and from prices to sentiment.

Other related works have focused on novel strategies to find causal relations across time. For instance, Preeti et al. [10] present the concepts of *temporal sentiment analysis* and *sentiment causal relations*. The combination of these two concepts is used to define a generalized prediction model that can be applied to predict the time period between

events and the sentiment of upcoming events. Similarly, Baumann et al. [11] focus on the interaction between emotion and topical shifts in a political context. The analysis is carried out on the Austrian Corpus of Parliamentary Records and the Austrian Media Corpus. Time series are built using each of the two corpora for three political parties and three variables, namely topical stability, valence and arousal. Granger causality is applied to discover causal relations between the time series.

In Dehkharghani et al. [9], a methodology for sentimental causal rule extraction from Twitter is proposed. A combination of machine learning- and lexicon-based approaches is applied for sentiment analysis while an association rule mining approach is used to extract causal rules. Their conclusions resemble ours in the sense that they stated that sentimental causal rules are an effective way to summarize important aspects and their causal relations from textual data, which can better support policymaking.

In causality analysis there are usually two branches: one in which the goal is to find causal relations, and the other where the goal is to make causal inference. These two branches can be combined resulting in a two-stage process in which causal discovery is used as a first stage for analyzing and creating models that capture relationships inherent in the data. Causal inference can be applied to the created models as a second stage to study the possible effects of altering a given system. As it has been pointed out in the literature (e.g., [43]) combining both approaches can result in statistical issues that invalidate the inference when we sum up the uncertainty of both stages. In this paper, we focus on causal discovery by proposing a methodology to obtain causal knowledge directly from textual data. Also, the motivation of this work is in line with recent information science literature that proposes moving beyond correlation [24].

Most existing proposals for sentimental causality apply classical causal discovery techniques, such as the Granger test, or causal association rule mining. The evolution of causality discovery methods opens up new opportunities to test them in different scenarios and settings, such as the ones explored in this article. In this work, we apply stochastic causality based on stochastic dominance [26] to build a causal graph of topics and emotions. As mentioned in the *Stochastic Causality* section of this article, this method can detect causal relations from data that are not necessarily arranged as time series.

Another salient aspect that distinguishes our approach from previous ones is the use of Plutchik's wheel of emotions [44], which considers eight primary emotions, i.e., anger, fear, sadness, disgust, surprise, anticipation, trust, and joy. This is in contrast to the most commonly adopted categorization for sentiment analysis, which relies on the use of positive, negative, and neutral polarities. To the best of our knowledge, no previous work has adopted Plutchik's wheel of emotions in causal analysis from social media.

## Methodology

The process of building a causal graph of topics and emotions from social media data requires several steps. The initial steps include data collection and indexing. Subsequent steps for the analysis of the collected data are query formulation, topic-based retrieval,

sentiment analysis, dataset construction with topics and emotions observed over time, and causal structure learning. Figure 1 presents a schematic overview of this process.
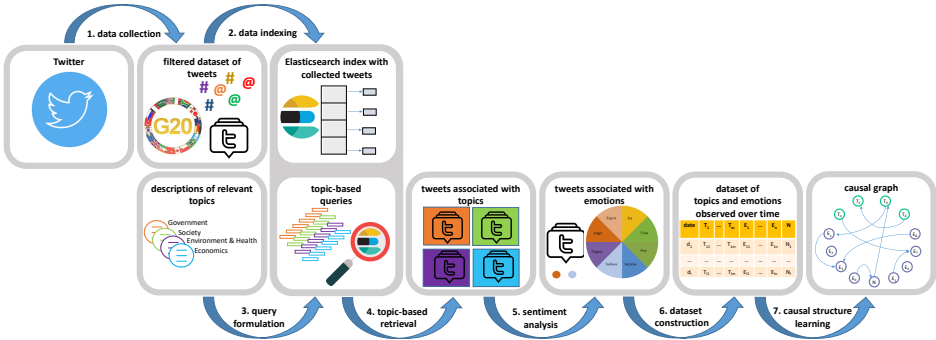


**Figure 1.** Steps applied in our proposed method to build a causal graph of topics and sentiments from social media data centered on G20 summits.

## Data collection and indexing

In the first step (described in Figure 1) social media data was collected from the Twitter API*. We aimed to create a multi-year dataset with tweets related to different G20 editions and the posts from and to the main politicians from the participating countries of the G20 summit. We crawled tweets at three main times: late 2018 (November and December), early 2020 and early 2021 (February and March).

Each year we filtered tweets that had any of the most frequently used hashtags related to the G20 summit. For instance, in 2018 we used: "#g20", "#g20summit" and "#g20argentina". Also, we included the official and personal user handles of all presidents or prime ministers (or the corresponding official channel for the presidency, e.g. "@WhiteHouse") from all countries participating in the G20 summits. We collected tweets posted by those user handles and tweets mentioning any of the monitored accounts. Every year the strategy for filtering stayed the same, except for the list of presidents and prime ministers, which was updated accordingly. Also, the list of hashtags was updated to reflect the hashtags of the corresponding year. The exact list of users and keywords used each year is provided as research data. Similarly, the list of tweet ids is made available for reproducibility of this work. A total of 49,030,053 tweets were collected. We then selected tweets whose language is either English or Spanish. The statistics about the volume of data are provided in Table 1. Finally, we indexed the entire dataset of Twitter posts in Elasticsearch (Figure 1 - Step 2) to support the retrieval of text.

---

*https://developer.twitter.com/en/docs/twitter-api

**Table 1.** Number of tweets in English and Spanish collected in three different periods of time: 2018 (November 6, 2018 to December 12, 2018), 2020 (February 21, 2020 to February 24, 2020 and March 2, 2020 to April 23, 2020) and 2021 (March 30, 2021 to April 10, 2021).

|         | 2018      | 2020       | 2021      |
| ------- | --------- | ---------- | --------- |
| English | 9,010,433 | 17,741,151 | 1,271,029 |
| Spanish | 1,497,151 | 338,666    | 235,828   |

## Data analysis

In order to identify salient topics discussed at different G20 editions, an expert with a background in international politics took different outcome documents[†] and selected four main recurring topics: government, society, environment & health and economics (as depicted in Step 3 in Figure 1). The expert also highlighted phrases in the document that were regarded as relevant for the topic. For instance, within economics the expert highlighted phrases such as "sustainable capital flows", and "global transportation routes and supply chains remain open, safe, and secure". The final list of phrases highlighted is made available as research data.

Once the expert finished with the highlighting of phrases, we built topical queries associated with those main phrases (Figure 1 - Step 4). We aimed at being precise with the retrieval, thus we aimed at matching exact phrases rather than individual words. For instance, "sustainable capital flows" must be matched as an exact phrase since individual words may not lead to topically relevant tweets. However, a phrase like "global transportation routes and supply chains remain open, safe, and secure" is likely too long and specific to retrieve any relevant tweets. Therefore, such cases are transformed into the following queries: "global transportation routes open", "global transportation routes safe", "global transportation routes secure", "supply chains open", "supply chains safe" and "supply chains secure". Despite not matching the original long phrase, an exact match with any of the substring queries is likely to retrieve relevant results for the topic. Finally, named entities of type "Organization" duplicate the number of queries, since the long and abbreviated forms are considered. For instance, the phrase "necessary reform of the World Trade Organization (WTO)" yields two queries: "necessary reform of the World Trade Organization" and "necessary reform of the WTO". Statistics of all the retrieved tweets broken down by topic and language can be found in Table 2. Figure 2 presents examples of tweets in English related to each of the analyzed topics.

Sentiment analysis refers to methods for determining the sentiment of some piece of text. In this work, sentiment analysis is carried out on tweets based on Plutchik's wheel of emotions [44]. According to Plutchik's theory, there are eight primary emotions: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation. Sentiment analysis techniques are typically based on machine learning [45, 46], lexicon methods [47, 48]

---

[†]`https://www.gov.za/speech-subjects/g20,` `https://www.mofa.go.jp/policy/` `economy/g20_summit/index.html` and `https://g20.argentina.gob.ar/en/` `ministerial-declarations-and-communiques`

**Table 2.** Number of tweets segregated by topic and language.

|  | Economics | Environment & Health | Society | Government |
|---|---|---|---|---|
| English | 202,044 | 165,671 | 201,810 | 161,251 |
| Spanish | 19,016 | 22,022 | 25,265 | 15,166 |

| Topic | Tweet | Words | anger | fear | sadness | disgust | surprise | anticipation | trust | joy |
|---|---|---|---|---|---|---|---|---|---|---|
| Economics | The US and Canada share the longest international border— and the most comprehensive trade and investment relationship, supporting millions of jobs in our two countries.  https://t.co/pE13TO69Yh https://t.co/MlTqzz4kUr | share<br>center<br>trade<br>supporting<br>**prevailing sentiments** |  |  |  |  |  | X |  X<br>X<br>X<br>X<br>**X** | X |
| Environment & Health | @realDonaldTrump The country is ready  for great debate about public health, environmental health, Universal healthcare , Universal income, free higher education, American Global leadership, philanthropy and love. People centered social policy, social justice & equity. Leaders see future not ownFACE | ready<br>public<br>income<br>love<br>policy<br>justice<br>**prevailing sentiments** |  |  |  | X |  | X<br>X<br>X | X<br>X<br>X<br>X<br>**X** | X<br>X |
| Society | @realDonaldTrump Made in America, hire American make America great. End corporate abuse of labor in the digital age and in manufacturing. Many areas to partner with India. Making tada and wipro and other Indian based consulting companies wealthy is not one of them by using H1Bs and offshoring. | hire<br>abuse<br>labor<br>**prevailing sentiments** | X | X | X | X | X | X<br>X<br>**X** | X<br>X<br>X<br>**X** | X<br>X<br>X<br>**X** |
| Government | @realDonaldTrump corruption is at the core of his businesses/money laundering and now NY bribery in the 90's. This family is about as dirty as they come!!! https://t.co/TuaB9grfXZ | corruption<br>money<br>bribery<br>dirty<br>**prevailing sentiments** | X |  |  | X<br>X<br>X<br>**X** | X | X | X | X |

**Figure 2.** Examples of tweets related to each topic (economics, environment & health, society, and government). Lexicon-based sentiment analysis based on Plutchik's wheel of emotions.

and linguistics [49]. This work applies a broadly used lexicon-based technique to carry out sentiment analysis on tweets by analyzing the tone of words comprising them. The lexicon used in our analysis is the English version and a manually curated Spanish version of the NRC Affect Intensity Lexicon [50], which contain a dictionary of words in each language, each labeled with the set of associated emotions. To score the sentiments of tweets we count the number of words associated with each sentiment based on the lexicon. The sentiment with the highest score is the one that is assigned to the entire tweet. If no emotion is associated with any of the words in the tweet, the tweet is assigned to the "no emotion" category. Otherwise, tweets are assigned to non-exclusive categories represented by Plutchik's primary emotions (Figure 1 - Step 5). Figure 2 illustrates the described sentiment analysis process.

The rationale for choosing a lexicon-based method instead of a machine learning-based method for sentiment classification relies on the fact that machine learning methods for sentiment analysis require a large volume of labeled data that is highly dependent on the specific domain. This means that to train a machine learning model, a huge labeling effort is required for each potential target topic or aspect under analysis. In contrast,

lexicon methods are generally simpler, more interpretable, and generalize better than machine learning methods to different domains. This makes it possible to keep up with the dynamic nature of emerging Twitter topics without retraining or requiring a larger annotation effort. Note, however, that the lexicon-based method adopted for sentiment analysis in the proposed methodology can be straightforwardly replaced by a machine learning-based method if sufficient labeled data is available to train a model to perform the required sentiment analysis step in the target domain.

After the sentiment analysis step is completed, we generate a dataset consisting of topics and emotions observed over time (Figure 1 - Step 6). This dataset is used for causal structure learning with the purpose of building a causal graph of topics and emotions (Figure 1 - Step 7). In this work, we adopt the method proposed by Vinod [26] for detecting stochastic causality, which is described next.

## Stochastic Causality

Vinod [26] developed kernel causality by extending Granger's ideas when data is not necessarily a time series. He mentions the following about Granger's notion:

> "However, this is needlessly restrictive and inapplicable for human agents (who read newspapers) acting strategically at time $t$ in anticipation of events at time $t + 1$".

To avoid the assumptions of linearity, parametric definitions or temporal precedence in the causality analysis, he constructs a test that exploits the asymmetries of explaining $X_i \sim f(X_j)$ and $X_j \sim f(X_i)$, where $f$ denotes a density function and $\sim$ implies a similarity or equivalence between the elements involved. This test has also the advantage of relying only on passively observed data. This is in contrast to the application of interventions and manipulations as part of a randomized experiment.

Vinod extends the concepts defined in [51], where in contrast to the theory of deterministic causality, causality can be defined in a probabilistic manner, which tolerates noise and violations of the causal path between cause and effect. Formally, we say that $X_i$ causes $X_j$ if

$$P\left(X_j \mid X_i\right) > P\left(X_j\right) \text{ a.e.,}$$

where "a.e." denotes almost everywhere. Then, instead of writing this inequality with probabilities, we could express it in terms of the conditional densities[‡]:

$$f\left(X_j \mid X_i\right) > f\left(X_j\right) \text{ a.e.}$$

When working with densities instead of probabilities, we have the advantage of using multiple regression to remove the effect of control variables, which are not readily available for probabilities of events. Taking into account that logically consistent probabilistic causality theory must retain robust asymmetry, i.e., not showing a causality

---

[‡]While the inequality is originally defined using conditional probabilities, for the sake of simplicity we use conditional densities. For more details see [52]

relation when cause and effect are swapped, Vinod's causality test can now be stated in terms of the differences between $f\left(X_j \mid X_i\right)$ and $f\left(X_i \mid X_j\right)$. Then, to exploit these asymmetries, Vinod [26] introduces a series of criteria to determine if a time series is stochastically dominant over another time series.

The definition of stochastic kernel causality assumes that three conditions should be met: (A1) conditional expectation functions are consistently estimated, (A2) data generating processes are standardized, and (A3) $X_k$ are control variables (no confounders). Then, we say that "$X_i$ causes $X_j$", i.e., $X_i \rightarrow X_j$, if and only the absolute errors predicting $X_j$ are smaller than the errors in predicting $X_i$, when using $\hat{f}$ as a prediction method. Instead of true (unknown) errors, we have the empirical residuals $e$ in the following form:

$$\left|e_{j \mid ik}\right| = \left|X_j - E\left[\hat{f}\left(X_j \mid X_i, X_k\right)\right]\right| < \left|X_i - E\left[\hat{f}\left(X_i \mid X_j, X_k\right)\right]\right| = \left|e_{i \mid jk}\right|.$$

Kernel regressions are used to obtain these residuals (hence the denomination "Stochastic Kernel Causality"). Vinod [26] points out two advantages of using this approach: (a) kernel regression fits are generally better than parametric linear or non-linear regressions (given the smoothness characteristics of the method); (b) kernel regressions do not place any unnecessary restrictions on the unknown conditional expectation functions (no need for assumptions on the distributional properties of parameters).

A kernel regression for a time series of length $T$ can be defined as:

$$X_{jt} = G_1\left(X_{it}\right) + \epsilon_{j \mid i}, \quad t = 1, \ldots, T,$$

and

$$G_1(X) = \frac{\sum_{i=1}^{T} X_j K\left(\frac{X_{it} - X}{h}\right)}{\sum_{i=1}^{T} \kappa\left(\frac{X_{it} - X}{h}\right)},$$

where $K(.)$ is a Gaussian kernel, $\epsilon_{j \mid i}$ is the empirical residual, and $h$ is the bandwidth chosen by leave-one-out cross-validation as in [53].[§]

To build an index that measures the strength of the causal relations based on the residuals resulting from the asymmetry exercise, Vinod proposes three criteria, if $X_i$ is the cause:

- Criterion 1: Model 1 defined as $X_k = G_1\left(X_i\right)$, is better than Model 2 ($X_i = G_2\left(X_k\right)$) in minimizing local kernel regression gradients.
- Criterion 2: The estimated absolute residuals of Model 1 should be smaller than those of Model 2.
- Criterion 3: The prediction accuracy of Model 1 is higher in terms of the coefficient of determination, i.e., $R^2$, than that of Model 2.

---

[§]For the sake of simplicity the definition is shown without the control variables $X_k$, which can be added to the equation.

The first two criteria are evaluated using statistical dominance [54]. With all the information obtained by the exercise of analyzing the three criteria above, the strength index is built, which takes values ranging from 0 to 100. In this work, we define a causal relation between two variables (topic and/or sentiment) if the strength index is exactly equal to 100, which implies that all the criteria signal the causal path in the same direction.

In an example used in [55], crime and deployment of police officers are displayed as a time series, and its correlation was 0.99. Despite this high correlation, the method correctly detects the direction of causality. The example shows that the causality detection method is useful in contexts of high correlation between variables.

The methodology can be summarized in a few steps for the practitioner. Given a set of variables representing time series:

(i) assume no confounder variables are present (all are control variables), this step requires theoretical a priori information;

(ii) standardize the time series;

(iii) define a desired threshold of strength in the statistical dominance index (this choice leads to a trade-off between precision and recall in the recovered causal relationships); and

(iv) build and analyze the resulting causal graph.

Next section presents the results obtained from applying the proposed methodology on real-world data. We used the R package *generalCorr*¶ provided in [26] for the causality analysis. To standardize the time series we used the function *standardize* from the R package *RobustHD*‖ [56], which transforms the values to have zero mean and a one unit standard deviation.

## Results and Discussion

This section presents the results obtained from applying the proposed methodology to the described dataset. By interpreting the obtained results, we offer insight into our research questions and emphasize the utility of the stochastic causality test adopted in our proposal.

To answer our research questions, we will now discuss how our different time series correlate and whether causal relationships exist. In Figures 3 and 4 we can see the difficulty of the problem at hand given that the time series of topics and sentiments show high correlation coefficients. This highlights the utility of the stochastic causality test, given that it has good properties in high correlation contexts, as it was mentioned in the methodology section.

For instance, the time series of tweets associated with the topics economics and government show a high correlation coefficient (0.95 as it is shown in Figure 5), and yet, there is no causality found between these time series (Figure 8). We see a similar

---

¶https://rdocumentation.org/packages/generalCorr/versions/1.2.1
‖https://rdocumentation.org/packages/robustHD/versions/0.7.2

relation between the time series corresponding to economics and environment (with a slightly higher correlation coefficient, i.e., 0.97) as shown in Figure 6, but in this case, a causality relation is found (i.e., economy-related tweets generate posts on environment concerns). As a last example with economics, we can see a relation with society topics, where the correlation is also highly significant, and the causal relation found with the test renders society as the cause (Figure 7).

**Figure 3.** Occurrences of topics (Y axis) over time (X axis).
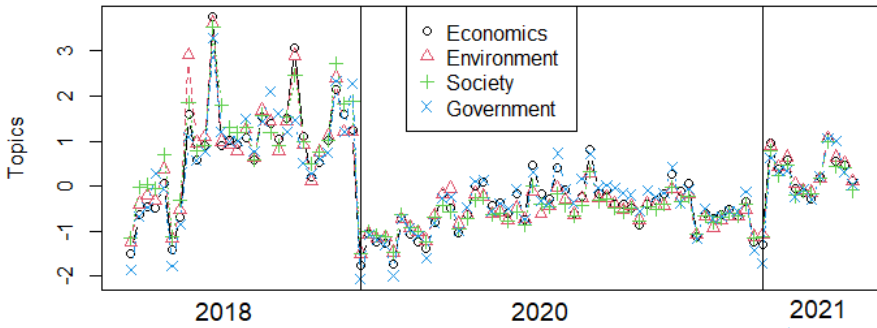


**Note:** The vertical lines delimit the three time periods, namely 2018, 2020 and 2021.

**Figure 4.** Occurrences of sentiments (Y axis) over time (X axis).
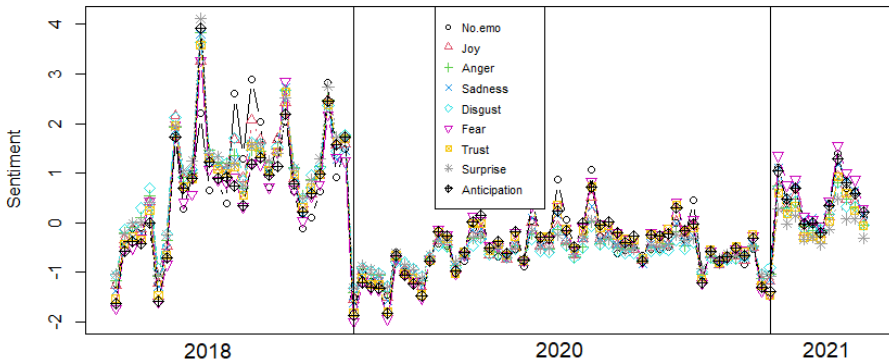


**Note:** The vertical lines delimit the three periods, namely 2018, 2020 and 2021.

The correlation matrix resulting from the generalized asymmetry test has correlation coefficients ranging from $0.9066$ to $0.9980$. In the cases used as examples, the coefficient for *economics* → *environment* is $0.9732$, while after the relation is flipped, the coefficient

**Figure 5.** Histograms and correlation between economics and government.
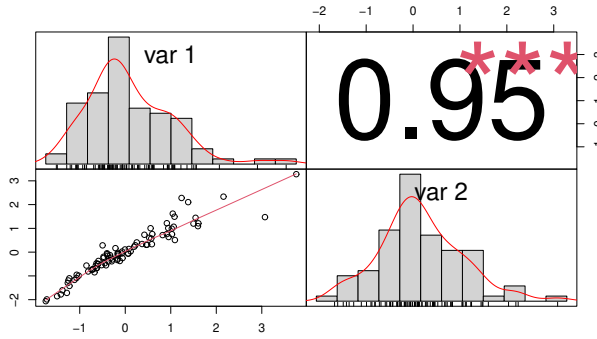


**Note:** $Var1$ is Economics and $Var2$ is Government. The top-right panel displays the correlation coefficient and its statistical significance (in this case, significant at 1% level).

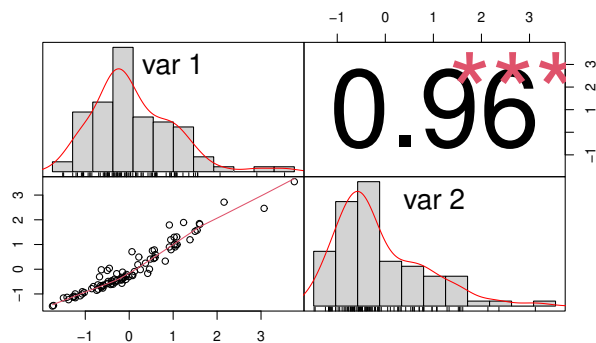**Figure 6.** Histograms and correlation between economics and environment
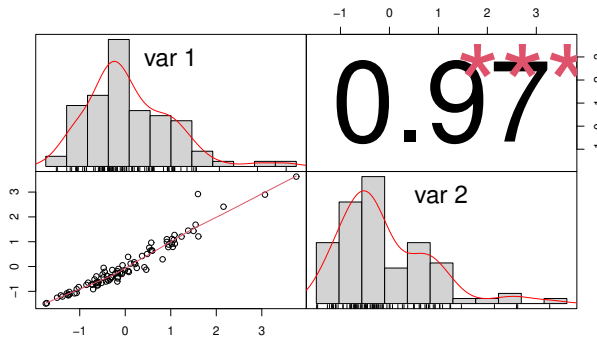


**Note:** $Var1$ is Economics and $Var2$ is Environment. The top-right panel displays the correlation coefficient and its statistical significance (in this case, significant at 1% level).

for *environment* → *economics* is 0.9746. The causality was significant in the latter case only.

As we mentioned in the previous section, we state that there is a causal link between two variables if the strength index is equal to 100 (i.e., all the criteria point to the same causal direction). For each pairwise evaluation, the remaining variables in the dataset (both remaining topics and sentiments) were added as control variables. The resulting graph depicting all pairwise causality relations can be seen in Figure 8. The first aspect that stands out is that the graph is not fully connected. Another observation is the absence

**Figure 7.** Histograms and correlation between economics and society.



**Note:** $Var1$ is Economics and $Var2$ is Society. The top-right panel displays the correlation coefficient and its statistical significance (in this case, significant at 1% level).

of bidirectional edges, which is a property of any causal graph obtained by the causal structure learning method that we applied.

**Figure 8.** The graph shows causality relationships among topics and sentiments for Twitter posts centered on different G20 summits.
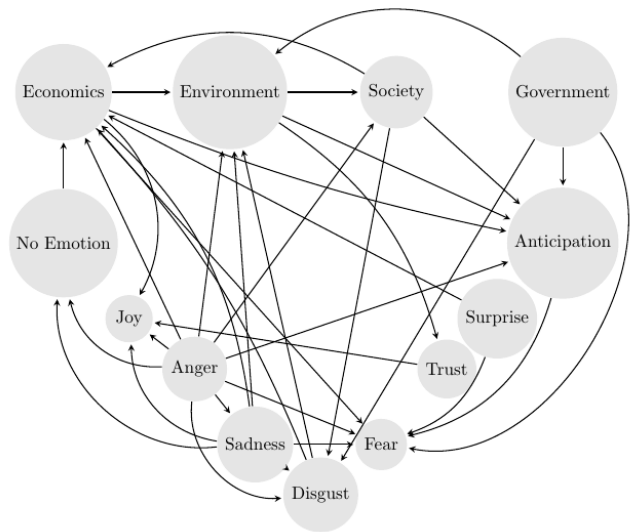
A summary of all causal relations can be found in Tables 3 and 4. The first column of the first table is sorted by how frequent the node is as the cause of other nodes. The first column of the second table is sorted by how frequent the node is as the effect of other nodes acting as the cause. The statistics resulting from these relations can be summarized as out-degree and in-degree values, which can be seen in Table 5.

**Table 3.** Causal relations sorted by frequency of the variable acting as cause of the causal relationship.

| Cause $\rightarrow$ Effect | | $gR^2$ C $\rightarrow$ E | $gR^2$ C $\leftarrow$ E |
|---|---|---|---|
| Anger | Joy | 0.993414 | 0.993058 |
| Anger | Sadness | 0.992291 | 0.988326 |
| Anger | Disgust | 0.998067 | 0.984463 |
| Anger | Anticipation | 0.97910 | 0.970846 |
| Anger | Environment | 0.989731 | 0.984737 |
| Anger | Society | 0.996280 | 0.995460 |
| Anger | No Emotion | 0.963777 | 0.927117 |
| Anger | Economics | 0.976630 | 0.971903 |
| Anger | Fear | 0.972600 | 0.955909 |
| Sadness | Economics | 0.987961 | 0.986517 |
| Sadness | Environment | 0.996444 | 0.992219 |
| Sadness | Disgust | 0.984463 | 0.982098 |
| Sadness | Fear | 0.986808 | 0.985059 |
| Sadness | No Emotion | 0.958496 | 0.956717 |
| Sadness | Joy | 0.993514 | 0.992043 |
| Economics | Environment | 0.974627 | 0.973195 |
| Economics | Anticipation | 0.994376 | 0.992990 |
| Economics | Joy | 0.984088 | 0.982345 |
| Economics | Fear | 0.985095 | 0.982345 |
| Government | Environment | 0.964022 | 0.956268 |
| Government | Anticipation | 0.977168 | 0.969597 |
| Government | Disgust | 0.959444 | 0.945011 |
| Government | Fear | 0.957164 | 0.953282 |
| Environment | Trust | 0.982027 | 0.976416 |
| Environment | Anticipation | 0.982717 | 0.972535 |
| Environment | Society | 0.980642 | 0.974933 |
| Society | Disgust | 0.99223 | 0.990296 |
| Society | Anticipation | 0.982887 | 0.972535 |
| Society | Economics | 0.974627 | 0.973195 |
| Disgust | Economics | 0.970925 | 0.961061 |
| Disgust | Environment | 0.988444 | 0.982021 |
| Surprise | Economics | 0.972045 | 0.967033 |
| Surprise | Fear | 0.963680 | 0.929908 |
| No emotion | Economics | 0.970925 | 0.949100 |
| Trust | Joy | 0.993360 | 0.989945 |
| Anticipation | Fear | 0.991003 | 0.986429 |
| Joy | **No effect** | | |
| Fear | **No effect** | | |

**Note:** $gR^2$ denotes the generalized asymmetric goodness of fit coefficient. It can be observed that values of the third column are always higher than those of the fourth column.

**Table 4.** Causal relations sorted by frequency of the variable acting as effect of the causal relationship.

| Effect ← Cause | |
|---|---|
| Economics | No Emotion |
| Economics | Sadness |
| Economics | Disgust |
| Economics | Surprise |
| Economics | Anger |
| Economics | Society |
| Fear | Sadness |
| Fear | Surprise |
| Fear | Anticipation |
| Fear | Government |
| Fear | Anger |
| Fear | Economics |
| Environment | Economics |
| Environment | Government |
| Environment | Anger |
| Environment | Sadness |
| Environment | Disgust |
| Anticipation | Economics |
| Anticipation | Environment |
| Anticipation | Society |
| Anticipation | Government |
| Anticipation | Anger |
| Disgust | Society |
| Disgust | Government |
| Disgust | Anger |
| Disgust | Sadness |
| Joy | Economics |
| Joy | Anger |
| Joy | Trust |
| Joy | Sadness |
| Society | Environment |
| Society | Anger |
| No Emotion | Anger |
| No Emotion | Sadness |
| Sadness | Anger |
| Trust | Environment |
| Anger | **No cause** |
| Government | **No cause** |
| Surprise | **No cause** |

Some interesting observations are worth highlighting. As we indicated in the previous section, despite the common pattern of high correlation coefficients between different time series pairs, only a small subset of significant causality relations was identified (compared to the full set of possible edges in the graph). The high correlation between each pair of variables is a natural consequence of the fact that whenever there is an increase in the activity associated with any topic or any sentiment the activity associated

**Table 5.** Out-degree and in-degree graph statistics.

| Node | Out-Degree | In-Degree |
|---|---|---|
| **Topics** | | |
| Environment | 3 | 5 |
| Economics | 4 | 6 |
| Government | 4 | 0 |
| Society | 3 | 2 |
| **Sentiments** | | |
| Anger | 9 | 0 |
| Sadness | 6 | 1 |
| Disgust | 2 | 4 |
| Anticipation | 1 | 5 |
| No-Emotion | 1 | 2 |
| Surprise | 2 | 0 |
| Joy | 0 | 4 |
| Trust | 1 | 1 |
| Fear | 0 | 6 |

with any other topic or sentiment also increases. The test applied in our analysis can reveal whether such a high correlation represents a cause-effect relation or not.

As an answer to our research question, we can state that the proposed methodology discovers significant causality relations among topics and sentiments extracted from the G20 discussion on Twitter, as shown on the causality graph of Figure 8 and in Tables 3 and 4. Also, as summarized in Table 5, we observe that some nodes, such as "Anger" and "Sadness" act as causes of multiple variables, while others, such as "Joy" and "Fear" are the effect of several other variables. Despite the absence of bidirectional edges and the presence of variables that never act as a cause and variables that never act as an effect, the graph contains some cycles, such as "Economics $\rightarrow$ Environment $\rightarrow$ Society $\rightarrow$ Economics". Interestingly, posts on government topics act only as cause for posts associated with other topics and sentiments.

To empirically test causality it is necessary to use experimental or quasi-experimental methods that allow the effects of variables to be controlled and that can demonstrate a temporal relationship between the cause and effect variables. Hence, it is important to mention that although the proposed methodology suggests the existence of several causal relations, the actual existence of such relations cannot be empirically tested, and hence remains an open question.

Different from existing proposals that look into the interrelation between topics and sentiments, the proposed methodology allows exploring how social media texts can be studied using methods that go beyond correlation analysis and other forms of interaction, such as those captured by joint sentiment-topic models. This novel methodology offers a valuable tool to understand causal relations among topics and sentiments as indicators for causal mechanisms.

## Conclusions and Future Research

This work proposed a novel methodology aimed at building causal graphs that represent the relations among topics and emotions in social media. The applied causality discovery process uses stochastic dominance, which relies on observational data only and does not require any assumptions of linearity, parametric definitions, or temporal precedence. To illustrate the application of the methodology in a specific scenario this work analyzed topics addressed at different G20 summits to determine whether causality relations exist among topics and sentiments that Twitter users express in their posts.

Recently, there has been considerable interest among researchers about using social media content as a means to monitor and measure how people discuss a variety of topics. Different from most approaches in social media analysis, in which patterns of associations are derived from regressions, correlations or frequent pattern mining, the proposed approach focuses on deriving causality relations by exploiting the asymmetries of explaining a variable from another variable. The proposal departs from the mainstream econometric causality techniques that rely on the Granger causality test and impose strict conditions on how observations are sorted in a time series. Instead, the technique adopted in our analysis applies stochastic causality based on stochastic dominance, which extends Granger's ideas when the data is not necessarily structured as a time series.

The application of the proposed methodology to the G20 domain provided insight into the research question of how to build a causal graph that represents causal relations among topics that were addressed at different G20 summits and the sentiments that social media users expressed in their posts. As a result, we derived a graph that represents the causal relations among four main recurring topics discussed at different G20 editions (government, society, environment & health, and economics), eight primary emotions based on Plutchik's theory (anger, fear, sadness, disgust, surprise, anticipation, trust, and joy), and no emotion. The graph offers a snapshot of the existing causal relationships among all the analyzed topics and sentiments by providing at the same time centrality information of each node both as a cause and as an effect.

The proposed methodology opens a new direction in the analysis of the relationship between topics and emotions in social media. It differentiates from existing work in topic-sentiment modeling in several ways. In terms of generality and scale, it can be applied under different scenarios to massive volumes of text data to identify causal relationships among the topics that are being addressed and the expressed emotions. It also emphasizes how causality analysis can offer a more powerful tool than interaction analysis. Causality analysis not only indicates a relation between topics and sentiments but also helps understand why specific sentiments arise in different scenarios. Understanding the cause of prevailing emotions in social media can help predict future events, and prevent or control them in the future. The proposed approach can replace or supplement other indicators typically used to predict future outcomes, such as opinion polls. Such an approach can naturally complement other existing proposals to derive insights into how citizens view particular government decisions or public policies [57, 58, 59]. For instance, by investigating the Chilean citizens' reactions expressed on social media about the topic "new constitution", it could have been possible to detect that these

reactions were mostly associated with "anger" and "fear" rather than with "joy". This analysis could have led to predict the voters' choice in September 2022 to reject a draft constitution that was due to replace the one drawn up under Augusto Pinochet's military rule. Anticipating this kind of outcome may have helped the Chilean constitutional convention to identify some flaws in the proposal and reformulate it.

Since stochastic causality relies on deciding the causal direction between a pair of variables by exploiting the causal asymmetries, the proposed approach is unable to capture bidirectional causality. This constraint results in discarding some potentially useful causal relationships from the derived causal graph. For instance, we observe that "Anger" is the cause of nine other variables but it is never the effect of any other variable. This is due to the fact that the generalized asymmetric goodness of the fit coefficient is always higher when "Anger" represents a cause than when it represents an effect. We contend that the existing nonbidirectionality constraint gives rise to some limitations that would require extending the proposed technique with additional steps that could complement the criteria adopted by Vinod's causality test. This could be achieved by combining stochastic causality with other causal discovery approaches through the implementation of an ensemble technique. However, this analysis is beyond the scope of the present work and will be considered for future research.

Another area for future work is to investigate the interrelation between information derived from social media data and socio-economic variables. While the former is more abundant the latter is typically more precise. We believe that the integration of both kinds of data will offer greater explainability, resulting in richer and more informative causal graphs.

## Data Availability

We made available as research data (1) the list of users and keywords used each year to filter tweets related to different editions of the G20 summit, (2) G20 outcome documents with relevant phrases highlighted by an expert with a background in international politics, (3) a list of expert-selected phrases related to four main G20 recurring topics (government, society, environment & health and economics), (4) topic-based queries (using Elasticsearch syntax) generated from expert-selected phrases for years 2018, 2020, and 2021, and (5) a file containing all the Twitter ids associated with the G20 tweets used in this analysis. The research data can be downloaded from `http://ir.cs.uns.edu.ar/downloads/assessing-causality-G20-supplementary-material-and-data.zip`.

## Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

## References

[1] Conrad FG, Gagnon-Bartsch JA, Ferg RA et al. Social media as an alternative to surveys of opinions about the economy. *Social Science Computer Review* 2019; : 0894439319875692.

[2] Granger CW. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* 1969; : 424–438.

[3] Rudin C and Radin J. Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition. *Harvard Data Science Review* 2019; 1(2).

[4] Poell T. Social media and the transformation of activist communication: exploring the social media ecology of the 2010 toronto g20 protests. *Information, Communication & Society* 2014; 17(6): 716–731.

[5] Sen I, Flöck F and Wagner C. On the reliability and validity of detecting approval of political actors in tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 1413–1426.

[6] Smailović J, Grčar M, Lavrač N et al. Predictive sentiment analysis of tweets: A stock market application. In *International workshop on human-computer interaction and knowledge discovery in complex, unstructured, big data*. Springer, pp. 77–88.

[7] Li J, Xu Z, Yu L et al. Forecasting oil price trends with sentiment of online news articles. *Procedia Computer Science* 2016; 91: 1081–1087.

[8] Chan SW and Chong MW. Sentiment analysis in financial texts. *Decision Support Systems* 2017; 94: 53–64.

[9] Dehkharghani R, Mercan H, Javeed A et al. Sentimental causal rule discovery from twitter. *Expert Systems with Applications* 2014; 41(10): 4950–4958.

[10] Preethi PG, Uma V et al. Temporal sentiment analysis and causal rules extraction from tweets for event prediction. *Procedia computer science* 2015; 48: 84–89.

[11] Baumann A, Hofmann K, Kern B et al. Exploring causal relationships among emotional and topical trajectories in political text data. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[12] Schreiber T. Measuring information transfer. *Phys Rev Lett* 2000; 85: 461–464. DOI:10.1103/PhysRevLett.85.461. URL `https://link.aps.org/doi/10.1103/PhysRevLett.85.461`.

[13] Sims CA. Macroeconomics and reality. *Econometrica* 1980; 48(1): 1–48. URL `http://www.jstor.org/stable/1912017`.

[14] Nicholson W, Matteson D and Bien J. Bigvar: Tools for modeling sparse high-dimensional multivariate time series. *arXiv preprint arXiv:170207094* 2017; .

[15] Chiquet J, Smith A, Grasseau G et al. SIMoNe: Statistical Inference for MOdular NEtworks. *Bioinformatics* 2008; 25(3): 417–418. DOI:10.1093/bioinformatics/btn637. URL `https://doi.org/10.1093/bioinformatics/btn637`. `https://academic.oup.com/bioinformatics/article-pdf/25/3/417/674790/btn637.pdf`.

[16] Radinsky K, Davidovich S and Markovitch S. Learning causality for news events prediction. In *Proceedings of the 21st international conference on World Wide Web*. pp. 909–918.

[17] Glymour C, Zhang K and Spirtes P. Review of causal discovery methods based on graphical models. *Frontiers in genetics* 2019; 10: 524.

[18] Nogueira AR, Gama J and Ferreira CA. Causal discovery in machine learning: Theories and applications. *Journal of Dynamics & Games* 2021; 8(3): 203.

[19] Maisonnave M, Delbianco F, Tohme F et al. Causal graph extraction from news: a comparative study of time-series causality learning techniques. *PeerJ Computer Science* 2022; 8: e1066.

[20] Pearl J. *Causality*. 2 ed. Cambridge University Press, 2009. DOI:10.1017/CBO9780511803161.

[21] Bareinboim E and Pearl J. Causal inference from big data: Theoretical foundations and the data-fusion problem. Technical report, DTIC Document, 2015.

[22] Peters J, Janzing D and Schölkopf B. *Elements of Causal Inference : Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 9780262037310. URL `https://library.oapen.org/bitstream/id/056a11be-ce3a-44b9-8987-a6c68fce8d9b/11283.pdf`.

[23] Meinshausen N, Peters J, Richardson TS et al. Foundations and new horizons for causal inference. *Oberwolfach Reports* 2020; 16(2): 1499–1571.

[24] Dong X, Xu J, Bu Y et al. Beyond correlation: Towards matching strategy for causal inference in information science. *Journal of Information Science* 2021; : 0165551520979868.

[25] Nogueira AR, Pugnana A, Ruggieri S et al. Methods and tools for causal discovery and causal inference. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2022; 12(2): e1449.

[26] Vinod HD. New exogeneity tests and causal paths. In *Handbook of statistics*, volume 41. Elsevier, 2019. pp. 33–64.

[27] Pearl J and Mackenzie D. *The Book of Why: The New Science of Cause and Effect*. Hachette UK, 2018.

[28] Meng X, Wei F, Liu X et al. Entity-centric topic-oriented opinion summarization in twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 379–387.

[29] Zhao Y, Qin B, Liu T et al. Social sentiment sensor: a visualization system for topic detection and topic sentiment analysis on microblog. *Multimedia Tools and Applications* 2016; 75(15): 8843–8860.

[30] Rahman MM and Wang H. Hidden topic sentiment model. In *Proceedings of the 25th international conference on world wide web*. pp. 155–165.

[31] Lin C and He Y. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*. pp. 375–384.

[32] Dermouche M, Velcin J, Khouas L et al. A joint model for topic-sentiment evolution over time. In *2014 IEEE international conference on data mining*. IEEE, pp. 773–778.

[33] Yang Q, Rao Y, Xie H et al. Segment-level joint topic-sentiment model for online review analysis. *IEEE Intelligent Systems* 2019; 34(1): 43–50.

[34] Sengupta A, Roy S and Ranjan G. Ljst: a semi-supervised joint sentiment-topic model for short texts. *SN Computer Science* 2021; 2(4): 1–16.

[35] Zhou T, Law K and Creighton D. A weakly-supervised graph-based joint sentiment topic model for multi-topic sentiment analysis. *Information Sciences* 2022; 609: 1030–1051.

[36] Liang H, Ganeshbabu U and Thorne T. A dynamic bayesian network approach for analysing topic-sentiment evolution. *IEEE Access* 2020; 8: 54164–54174.

[37] Pathak AR, Pandey M and Rautaray S. Topic-level sentiment analysis of social media data using deep learning. *Applied Soft Computing* 2021; 108: 107440.

[38] Yin H, Yang S and Li J. Detecting topic and sentiment dynamics due to covid-19 pandemic using social media. In *International conference on advanced data mining and applications*. Springer, pp. 610–623.

[39] Godovykh M, Ridderstaat J, Baker C et al. Covid-19 and tourism: Analyzing the effects of covid-19 statistics and media coverage on attitudes toward tourism. *Forecasting* 2021; 3(4): 870–883. DOI:10.3390/forecast3040053. URL `https://www.mdpi.com/2571-9394/3/4/53`.

[40] Silahtaroğlu G, Dinçer H and Yüksel S. *How Is the Stock Exchange Index Affected by the Disclosures of Politicians?* Cham: Springer International Publishing. ISBN 978-3-030-74176-1, 2021. pp. 129–144. DOI:10.1007/978-3-030-74176-1_6. URL `https://doi.org/10.1007/978-3-030-74176-1_6`.

[41] Usher J, Morales L and Dondio P. Brexit: A granger causality of twitter political polarisation on the ftse 100 index and the pound. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. pp. 51–54. DOI:10.1109/AIKE.2019.00017.

[42] Scaramozzino R, Cerchiello P and Aste T. Information theoretic causality detection between financial and sentiment data. *Entropy* 2021; 23(5): 621.

[43] Gradu P, Zrnic T, Wang Y et al. Valid inference after causal discovery. *arXiv preprint arXiv:220805949* 2022; .

[44] Plutchik R. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist* 2001; 89(4): 344–350.

[45] Hasan A, Moin S, Karim A et al. Machine learning-based sentiment analysis for twitter accounts. *Mathematical and Computational Applications* 2018; 23(1): 11.

[46] Hassan SU, Saleem A, Soroya SH et al. Sentiment analysis of tweets through altmetrics: A machine learning approach. *Journal of Information Science* 2021; 47(6): 712–726.

[47] Thelwall M, Buckley K and Paltoglou G. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology* 2011; 62(2): 406–418.

[48] Khoo CS and Johnkhan SB. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science* 2018; 44(4): 491–511.

[49] Song C, Wang XK, Cheng Pf et al. Sacpc: A framework based on probabilistic linguistic terms for short text sentiment analysis. *Knowledge-Based Systems* 2020; 194: 105572.

[50] Mohammad S and Turney P. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. pp. 26–34.

[51] Suppes P. A probabilistic theory of causality. *British Journal for the Philosophy of Science* 1973; 24(4).

[52] Salmon WC. An" at-at" theory of causal influence. *Philosophy of Science* 1977; 44(2): 215–224.

[53] Vinod HD. *Hands-On Intermediate Econometrics Using R: Templates for Extending Dozens of Practical Examples (With CD-ROM).* World Scientific Publishing Company, 2008.

[54] Anderson G. Nonparametric tests of stochastic dominance in income distributions. *Econometrica: Journal of the Econometric Society* 1996; : 1183–1193.

[55] Vinod HD. Generalized correlations and kernel causality using r package generalcorr. *Available at SSRN 2782223* 2016; .

[56] Alfons A. robusthd: An r package for robust regression with high-dimensional data. *Journal of Open Source Software* 2021; 6(67): 3786.

[57] López-Chau A, Valle-Cruz D and Sandoval-Almazán R. Sentiment analysis of twitter data through machine learning techniques. In *Software engineering in the era of cloud computing*. Springer, 2020. pp. 185–209.

[58] Hubert RB, Estevez E, Maguitman A et al. Analyzing and visualizing government-citizen interactions on twitter to support public policy-making. *Digital Government: Research and Practice* 2020; 1(2): 1–20.

[59] Sandoval-Almazan R and Valle-Cruz D. Sentiment analysis of facebook users reacting to political campaign posts. *Digital Government: Research and Practice* 2020; 1(2): 1–13.