Annals of Operations Research
https://doi.org/10.1007/s10479-024-06422-6

ORIGINAL RESEARCH

A network-based matching design for text mining of controversial online reviews

Giulio Giacomo Cantone[1,2] · **Venera Tomaselli**[1]

Received: 4 March 2024 / Accepted: 26 November 2024

Abstract
Online reviews provide users with the opportunity to rate various types of items such as movies, music, and video games using a combination of numeric scores and textual comments. The study proposes a novel method that applies statistical matching on network-based covariates, with the aim to improve the estimation of the association between words and highly controversial items in online reviews. The application of this method on a sample of 40,665 items from the website Metacritic detects 218 highly controversial items. The application supports the theory that controversies on Metacritic are driven with a sense of self-awareness of participating of an online controversy ('review bombing'). Typical controversial topics (sexual identities, religious morality, politics) are associated with controversial reviews, too.

Keywords Review bombing · Polarisation · Bipartite networks centrality · Statistical matching · Text mining

1 Introduction

Through online reviews, users rate *items* such as consumer goods, services, or other subject of a personal judgement with a numeric score or a textual comment, or both (Stöckli and Khobzi, 2021; Watson and Wu, 2022; Sharkey et al., 2023). Given that typically platform websites of online reviews offers large catalogues of item to review, researchers have noticed consistent empirical patterns diverging from typical cases, and tried to associate these to relevant human phenomena implied in the process of consumption of these items. In other words, by observing the distribution of numerical ratings, some items stand out as statistical outliers. For example, Extremely Bipolar Items (EBIs) exhibit anomalous inflation in the frequency of maximum and minimum scores of the rating scale, an empirical occurrence that is not observed outside the cluster of EBIs, and that is related to a source of controversy regarding these items (Li and Hitt, 2008; Hu et al., 2009; Anderson and Simester, 2014; Fu et al., 2015; Amendola et al., 2015; Hu et al., 2017; Zhuang et al., 2018; Santos et al., 2019;

✉ Venera Tomaselli
venera.tomaselli@unict.it

Giulio Giacomo Cantone
g.g.cantone@sussex.ac.uk

[1] Department of Economics and Business, University of Catania, 55, Corso Italia, 95129 Catania, IT, Italy

[2] Science Policy Research Unit - Business School, University of Sussex, Brighton, UK

Published online: 23 December 2024

Janosov et al., 2020; Lu et al., 2020; Schoenmueller et al., 2020; Day and Kim, 2022; Ziser et al., 2023; Li et al., 2023; Cantone et al., 2024).

Observing the most frequent words in these reviews can help to understand the underlying phenomenology. Nevertheless, to simply identify a cluster of outliers and to detect the most frequent words may reach a perfunctory or misleading result. Aside from removing 'stop-words', there may be semantics that are frequent in the cluster, but only because they are frequent in the catalogue of items. If, for example, in a catalogue of video games a majority has violent content, it is reasonable that the semantics of violence ('killing','blood','guns', etc.) would be frequent in the reviews of the most hyper-bipolarised (ergo, controversial) video games. As convincing as it may sound at first sight, to call for an association between violent semantics and polarisation would be an error, since violent semantics would be frequent even in those video games where the opinions are not polarised. The correct method to assess the association between semantics and the cluster should consider words which are frequent in the cluster but rare outside of it. This study proposes that, instead of considering, a more parsimonious technique would be to match every element in the cluster to an element out of the cluster by minimising the distance in their covariates (Ho et al., 2007; Imai et al., 2008; Seawright and Gerring, 2008; Aral et al., 2009; Stuart, 2010; Morgan and Winship, 2014; Steiner et al., 2015; Dong et al., 2021). While this proposal can be applied to many collections of small text (there are many small texts for each items), online reviews have the property that can be modelled as bipartite networks os items and users of the platform. Therefore, by attributing network-based statistics as covariates of the items, it is possible to improve the quality of the matching algorithm (Aral et al., 2009; Charkhabi, 2014; Dewan et al., 2017).

The study is roughly divided into four main sections, plus a final comment on the generalisation of the proposed observational design. The first is an overview of the theoretical and methodological issues involved in the statistical modelling of online reviews. Here is also organically presented the issue of the operative definition of bipolarisation of rating score as inherently relevant to the general statistical theory of online ratings. The second section details how the cluster is identified, the specification of the matching procedure, and the metrics of text mining to finalise the comparison. The third section reports the application of the methods. 1, 552, 750 public reviews are collected from 40, 665 items of the platform Metacritic (Kasper et al., 2019; Santos et al., 2019), where 218 EBIs are detected and matched. Some topics are found particularly related to EBIs: politics and morality, commercial brands, and sexuality. The most salient semantic concept is the mention of the review bombing, a phenomenon documented in Cantone et al. (2024), where a large number of users try to sabotage the reputation of an item. Limitations of the validity of these results are presented in the fourth section.

2 Network model of online reviews

2.1 Classical models of online reviews

Let the vector

$$\mathbf{i} : \{i_1, i_2, \ldots, i, \ldots, \}$$

be the catalogue of items, and let the vector

$$\mathbf{u} : \{u_1, u_2, \ldots, u, \ldots, \}$$

Let a generic review be a vector $\mathbf{j}$ conveying information of different natures (text, score, time, etc.). Consider this example where:

$$\mathbf{j} : \{i, u, y, \texttt{text}, \texttt{time}, \ldots\}$$

where y is the numeric score that u assigned to i. In other words, a $\mathbf{j}$ concerns the interaction between a sender node u, the reviewer or the user of the platform, and the receiver node i, the item in the catalogue of the platform. The interaction counts as a directed link in a bipartite network. The number of links sent by a user is the sender degree k_u, and the number of links received by an item is the receiver degree k_i (Latapy et al., 2008; Graham and Paula, 2019).

Given a $J : \{\mathbf{j}, \ldots\}$ set of reviews, a relevant covariate (or, attribute) for the review $\mathbf{j}$ is the aforementioned rating score

$$Y : \{y, \ldots\}$$

that the user assigns to the item within an interval scale ($\underset{\min}{y} : \underset{\max}{y}$). Only for the sake of simplification, in the rest of the study it will be assumed $\underset{\min}{y} = 0$, but everything else would hold even if $\underset{\min}{y} \neq 0$ since scales can be re-normalised subtracting $\underset{\min}{y}$. The $y_{u,i}$ assigned to i by u acts as the numeric equivalent of his expression of a feedback, hence a positive measure of the sentiment. The vector $\mathbf{y}_i$ contains all the ratings received by i.

The following is an example of a model with a simple functional form for the determination of the rating score:

$$\begin{aligned} y &= \mathcal{B}(n; p) \\ n &:= \underset{\max}{y} \\ p &:= p_{u,i} = \theta \cdot \alpha_i + (1 - \theta) \cdot \beta_u \end{aligned} \tag{1}$$

where α_i is a fixed latent attribute of the item that can be easily assumed to be its 'quality'; β_u is a fixed value for the user u, which is the general tendency of the user to rate items positively; θ is the global tendency to let the quality of the item prevail however the user's inclination for the determination of the rating score and $\mathcal{B}$ is the symbol for the Binomial distribution of n binary draws with p probability.

Assumptions of Eq. 1 are that α_i exists as latent propriety of the item and that β_u is stochastically stable over time: users do not rate more or less positively the more they review items. Minor assumptions of the model regard the modellisation of sources of errors (e.g. the moodiness of the user at the time of the review) as a noise ϵ with an ignorable effect on the average case.

The message conveyed in presenting Eq. 1 is that even a very simple Binomial model with strong assumptions (e.g. on the stability over time of β_u) would still require the estimation of three parameters. Nevertheless, the most common practices of modellisation of online reviews adopt the sample mean of the received rates $\bar{y}_i$ as an estimator for α_i (de Langhe et al., 2016; Santos et al., 2019; Janosov et al., 2020), instead. Compared to Eq. 1, the implicit assumptions behind this practice is that β_u is just another source of unbiased error of $\bar{y}_i$, equivalently to set $\theta = 1$ and $\bar{\epsilon} = 0$.

Such a set of assumptions is not baseless: according to Hu et al. (2009) these assumptions are satisfied in experimental studies on rating behaviours (e.g. focus groups). However, they

do not hold for online platforms, which are typically characterised by a bi-modal, concave, shape (J-shape). Hu et al. (2017) explain the J-shape as determined by:

1. Cognitive availability. The probability of receiving reviews is higher for well-known, famous, items.
2. Sentimental commitment. If u knows i, he will provide feedback on i more likely if he feels a strong sentiment towards it. This bias also regards k_i because items with extremely high or low α_i should reach a higher degree k_i.

According to Hu et al. (2009) k_i and $\mathbb{E}(y_i)$ should be correlated because availability implies also a favourable opinion of the item. However, in the large dataset of ratings observed in Electronic Supplement of Janosov et al. (2020) there is no Kendall correlation between k_i and $\bar{y}_i$.

Finally, this model is disputed by Brandes et al. (2022), too. According to their alternative, the bipolarity of online reviews is also explained by users with moderate inclinations towards rating products being more likely to leave the pool of active reviewers.

2.2 Inflated bipolarity in online reviews

Going beyond the characterisation of Hu et al. (2009) and of Brandes et al. (2022), there are other reasons to observe a J-shape in $\mathbf{y}_i$. It is often unclear if for the user the feedback sentiment regards a personal experience of the item or a more general opinion. Indeed, the latter can be socially induced even in the absence of a direct experience. For example, a user can think: "I dislike this movie because it offends my religion, and I do not need to watch it, because someone else told me enough details on it."

If not differently specified, a review does not imply necessarily that u purchased or actually used i or that u truly holds a sentiment towards i, since he can just follow the request of a third party (Anderson and Simester, 2014; Lee et al., 2021). Actions may be performed under disguise (a 'sock-puppet', i.e. a fake account), or by an artificial agents, i.e. *bots* (Ferrara et al., 2016; Kumar et al., 2017). These are nuances of the deception and disinformation involved in online reviews. According to Wu et al. (2020), fake reviews could range from 10% to 30%. Fake reviews are associated with a high frequency of $y_i = m$ (Anderson and Simester, 2014; Mayzlin et al., 2014), and this empirical pattern is explained by the practice of corporate brands to artificially inflate their own ratings by various means (Ratkiewicz et al., 2011; Petrescu et al., 2022).

In other cases, disgruntled users organise themselves to push lower scores, to sabotage an item. This behaviour is commonly called "review bombing" (Cantone et al., 2024). The documented existence of astroturfing and review bombing implies that those who perform such acts believe in being influential over others with their actions. However, consequences can be chaotic and misaligned with the original intents. Hardly a simple parametric model as Eq. 1 is sufficient to infer associations with the collective behaviour of users, because different populations (or, classes) of users follow different behavioural models.

These phenomena lead the interest for expanded theories behind the patterns of bipolarity. Cantone et al. (2024) hypothesise that EBIs have peculiar latent characteristics that make them targets of bipolarisation, dealing with naturally controversial topics. These latent characteristics could be semantically mirrored in the words of the reviews of EBI, hence the relevance of unbiased estimation of the association of keywords to EBI.

3 Network-based matching design

Without an explicit strategy to adjust for covariates, it is not possible to assess if the observed frequent words in a cluster of EBIs mirror an intrinsic driving features of the controversy. A design for overcoming this issue is presented in three steps:

- An operative definition for a measure of controversy that is based on robust feature of bipolarity. The resulting index $\Cup(i)$ would identify the EBIs.
- Z covariates must be identified too. Each EBI is matched to its non-EBI nearest neighbour, minimising the global distance among the covariates.
- In this context, text mining involves fundamentally the procedures to assess an η coefficient of association between a word and the cluster.

3.1 Index of controversy

Intuitively, bi-polarity in a multipoint scale is defined as inflation of extreme scores (Fisher et al., 2018; Schoenmueller et al., 2020). Schoenmueller et al. (2020) define bi-polarity with a simple nonparametric index starting by observing the $f_i(y)$ frequencies of extremes y for the i. A generalisation of their method is:

$$\Cup_0(i) = f_i(y = \underset{\min}{y}) + f_i(y = \underset{\max}{y}) \tag{2}$$

This indicator is misleading for cases of only one inflated between minimum or maximum. A robust alternative is the following:

$$\underset{\text{NP}}{\Cup}(i) = \min\left[f_i(y = \underset{\min}{y}) + f_i(y = \underset{\max}{y})\right] \cdot 2 \tag{3}$$

which is more conservative and easier to interpret: for example, differently from the index o Schoenmueller et al. (2020) it has a univocal interpretation of $\underset{\text{NP}}{\Cup}(y) = 1$, since it happens only for $f_i(y = \underset{\min}{y}) = f_i(y = \underset{\max}{y}) = .5$.

From the last consideration a semi-parametric alternative is proposed, that is the ratio between the sample variance within $\mathbf{y}_i$ over its bounded (considering $n(y_i) = k_i$) theoretical maximum:

$$\underset{\text{SP}}{\Cup}(i) = \frac{\hat{\sigma}^2(y_i)}{\arg\max[\hat{\sigma}^2(y_i)]} \tag{4}$$

which, for $\underset{\min}{y} := 0$ by assumption or by re-normalisation of y, depends exclusively on $\underset{\max}{y}$. Equation 4 satisfies that if

$$f(y = \underset{\min}{y}) = f(y = \underset{\max}{y}) = .5$$

then

$$\underset{\text{SP}}{\Cup}(i) = 1$$

Finally, since $\underset{\text{NP}}{\Cup}(y)$ and $\underset{\text{SP}}{\Cup}(y)$ are both in the unit interval scale and they share the same conditions for minima and maxima, they can be composite into a singular index of controversy

through their harmonic mean:

$$\mathbb{U}(i) = \frac{2}{[\underset{\text{NP}}{\mathbb{U}}(i)]^{-1} + [\underset{\text{SP}}{\mathbb{U}}(i)]^{-1}} \tag{5}$$

3.2 Matching algorithm and covariates

The matching algorithm pairs each element in the cluster with the most similar element out of the cluster through a Nearest Neighborhood algorithm that aims at minimisation of the distance between the item and its 'potential twin' out of the cluster. The distance is measured over a $\mathbf{Z}$ set of observable covariates of the items. Among these covariates, one must make a distinction between covariates which are direct attributes of the item from those that are indirectly computed from the network structure, or structural covariates. The firsts are accessed in the phase of data collection while the second are defined within the analysis. Follows the list of structural covariates.

3.2.1 Measures of quality

- $\bar{y}_i$ is the sample mean of scores from collected public reviews. As aforementioned, this is usually employed as an estimator of α_i (see Eq. 1).
- $\mathbb{E}(y_i \mid u)$, simplified as $e(i)$ is the prior for $\bar{y}_i$. It assumes knowing (only) the vectors of scores $\mathbf{y}_u$ submitted before the review, for all the u who reviewed i, and it is estimated as follows:

  1. For each i, consider only those u such that $\exists y_{u,i}$ timed before $\mathbf{j}$ are listed, that means who reviewed other items before reviewing i Let $\underset{t<t_i}{\mathbf{y}_u}$ be the set of ratings expressed by u before $y_{u.i}$
  2. then set

  $$e_{u,i} := \underset{t<t_i}{\bar{y}_u} \tag{6}$$

  that is the average score expressed in past reviews by users who then reviewed i.
  3. The estimator of $e(i)$ is just the average previous scoring trends of the users who reviewed i:

  $$e(i) := \bar{e}_{u,i} \mid i \tag{7}$$

3.2.2 Measures of centrality

The network centrality of the node is a concept associated with the relevance of an item conditional to the cognitive availability ('fame') of it among the users.

- k_i is the degree centrality of the item. In conventional applications, degree centrality is sometimes, but not always, interpreted as a measure of fame.
- The median $med_{.i}(k_u)$ of the users who reviewed i is an *indirect* measure of the centrality of the item, measuring the excess degree distribution for i.
- $\underset{k_u=1}{f_i}$ is the relative frequency of $k_u = 1$ for i. It is a relevant non-parametric statistic because it is a spurious measure (it is correlated to) of the share of agents of disinformation targeting the item, given that an agent of disinformation can always sign up with a new

disguise (sock puppet, botnets) and push more scores and reviews (Kumar et al., 2017; Cantone et al., 2024).

Finally, there are measures expressly designed to measure centrality in bipartite networks. The reason to adopt ad-hoc measures is that conventional interpretations of centrality do not hold entirely for models of bipartite networks. For example, it is controversial to assert that a sender, being an agent, has proprieties isomorphic to topological proprieties of a place with no *agency*. Often methods for measuring the centrality of only a class of nodes (senders or receivers) involve redefining the bipartite network as a "one-mode projection" of that class, that is the network where one of the two original classes (senders or receivers) is kept as nodes, and the other is redefined as links. It is debated if one-mode projections bring a relevant loss of structural information through the suppression of the nodes (Lehmann et al., 2008).

- In this study, the `Birank` score (He et al., 2017; Yang et al., 2020) is chosen as third measure of centrality. Birank's algorithm has been developed expressly to avoid the one-mode projection and it has been evaluated as the best-performing algorithm for the centrality of nodes in bipartite networks (Yang et al., 2022). The interpretation of `Birank` scores for items is analogue to the PageRank algorithm. Higher `Birank` is associated with items that receive many reviews from users who review many items with high k_i.

3.3 Text mining

Text mining involves procedures for converting textual comments into statistical objects, then associating words to the cluster of outliers. A classical technique called 'bag-of-words' is proposed: reviews are tokenised and stopwords are filtered out (Silge and Robinson, 2017; Gentzkow et al., 2019). The corpus of token has a structure on four levels: tokens; reviews; items; and finally the two groups emerging from the matching algorithm, the cluster and the matched control group. Even if the two groups have the same number of items by definition, the numbers of reviews and is not forced to be equal, although the matching algorithm will keep these of the same magnitude.

To account for this structure in the definition of the coefficient of association $\eta_{\Cup}(token)$ between a word and the EBI cluster, the following metric is adopted:

1. For each item within the cluster or within the set of the matched counterparts, a positive number $c_+(i,a)$ is associated to each token. $c_+(a,i)$ is the number of times the token a occurs at least one times in a review about i, plus 1.

$$c_+(a,i) = \#(\text{reviews of} \quad i \quad \text{containing} \quad a) + 1 \tag{8}$$

2. $c_+(a,i)$ is divided by the whole number of reviews received by i, plus one. Consider the *logit* function[1] of this division:

$$l(a,i) = logit\left[\frac{c_+(a,i)}{k_i + 1}\right]$$

   This quantity is a normalisation of a pseudo-frequence of the word in how the users would describe their experience with the item.
3. Items are paired, so $i_{X=1}$ would represent the item within the cluster, and $i_{X=0}$ its matched item. Then the following coefficient of association between tokenised words (a) and the

[1] $logit(x) = \ln \frac{x}{1-x}$

condition to be in the cluster (X) is proposed:

$$\eta(a, X) = \sum_{(pair)} [l(a, i_{X=1}) - l(a, i_{X=0})] \cdot \frac{100}{\#(pairs)} \tag{9}$$

where the multiplication for 100 act just to improve the readability of the η index.

With this method to score a high η a word must appear frequently in the cluster, and rarely in the matched item, but it also must be frequent in many reviews of the same EBI and not be concentrated in few reviews repeating always the same word (e.g. "THIS IS GOOD GOOD GOOD"). The formula for η, by being an average difference, and not a difference of averages, fully accounts for the matched structure among the considered items.

4 Application

4.1 Data

A sample of $1,552,750$ public reviews has been collected across $40,665$ items from the catalogues of the platform Metacritic. In Metacritic, users can review an item privately, too; these reviews cannot be collected. The multipoint scale for Metacritic $(0:10)$ and items are classified by the website as movies ($n = 10,617$), music albums ($n = 8,431$), serial shows ($n = 3,318$ seasons) and video games ($n = 18,299$). Reviews come with a textual comment, a score, and the day of submission. The oldest sampled review was submitted on January 2001 while the last one on November 2021.

The average score in the sample is $\bar{y} = 7.14$. The median number of reviews *per* item is $Med(k_i) = 7$, with 22 being the 75^{th} and 65 the 95^{th} percentile. Confirming the findings of Janosov et al. (2020), the Kendall correlation between k_i and $bar y_i$ is trivial (-0.02).

$635,781$ unique users are detected. Summed to $40,665$ items they constitute a bipartite network of $676,446$ nodes. Of these users, $453,359$ (.713 of total) submitted only a public review, $82,909$ (.13) submitted two public reviews, 22.691 (.05) submitted three, and $17,250$ four. Only 8% of users submitted more than four public reviews on Metacritic.

The Theory of Attrition (Brandes et al., 2022) does not explain the Metacritic dataset, while the assumption of Eq. 1 on the stochastic stability of β is not rejected: users with a large number of reviews use less extreme scores but overall the rating behaviour seems time-independent (see Fig. 1).

Collected items are directly associated with attributes (covariates) provided natively by Metacritic. These are:

- The UserScore or `US`, is a number that represents a native score for the quality of the item, based on both public and private scores. This information is displayed through browsing the website Metacritic.
- The MetaScore or `MS` is another score assigned by Metacritic to items. It is a summary score of the judgement of expert journalists only. MetaScore represents the consensus of the experts.
- The `year` of publication of the item.

In the sample, the average bias of public scores adopting UserScore as ground truth is trivial: -0.06. For comparison, the mean difference between the users and the experts (MetaScore) is five times larger in absolute values: 0.3, which is coherent with previous findings (Santos et al., 2019). Given the negligible bias, public scores are assumed as representative of the whole population of public and private scores.

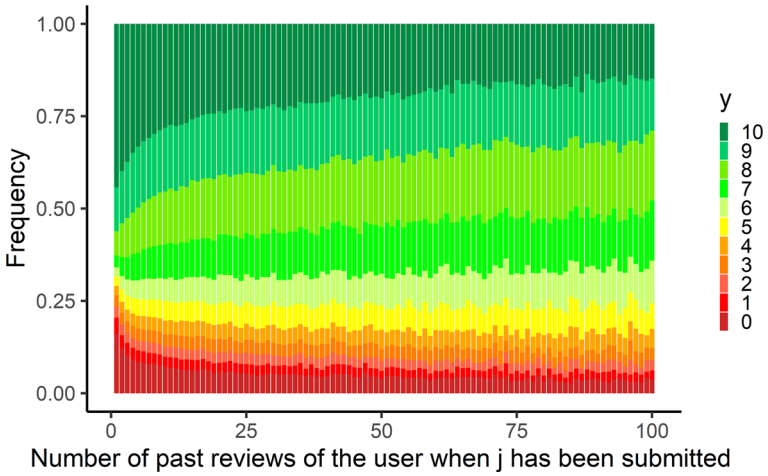

Fig. 1 Less extreme scores are observed when users submit more public reviews on Metacritic. On the vertical axis is represented the relative frequency of the scores. Extreme values are much more likely for users for the first reviews of users. Scores are less extreme after users already reviewed some items in the past. Consider that $\sim 41\%$ of the reviews is concentrated in the first bar of the plot ($k_u = 1$) alone. It is plausible that these frequencies for low k_u shows higher extreme scores due to the high number of users submitting only a review. This statistical behaviour is coherent with the hypothesis that a part of such users with $k_u = 1$ are agents of disinformation (astroturfers, saboteurs, etc.)

4.1.1 Criteria of exclusion from the sample

In the analysis, 90% of items in the sample received less than 66 public reviews. They are under-sampled for robust text mining and possibly they lack relevance too, so they will be excluded from the analysis. 42 additional items are excluded because Metacritic did not assign a MetaScore to them. The video game "The Last of Us Part 2" is excluded from this selection because it is an extreme outlier: it is the item with the largest $k_i = 78,219$ (Cantone et al., 2024), by far much more than any other, and by removing it the matching algorithm improves its diagnostic statistics.

4.2 Classification of EBIs

In Table 1, indicators of polarity are always negatively correlated with indicators of quality. This correlation is likely mediated by k_i: the considered selection of the 3,978 items in the top 10% of k_i does not show this correlation as strong anymore. The full range of correlations within the covariates is in Appendix (Table 5 and Table 6).

Among these 3,978 candidates with $k > 65$ public reviews, 218 EBIs are identified as the items with a $\mathbb{U}(i)$ over the 95^{th} percentile ($\mathbb{U}(i) > .42$), see Fig. 2. Of these 218 EBIs, 26 are movies, 37 are music albums, 19 are seasons of serial shows, and 136 are video games.

4.3 Matching

The 218 EBIs are paired with other 218 items not classified as EBI. The matching algorithm follows these rules:

Table 1 Kendall correlations of measures of polarity with other indicators

	Total Sample		$k_i > 65$	
	$\Cup_{NP}$	$\Cup_{SP}$	$\Cup_{NP}$	$\Cup_{SP}$
US	−0.14	−0.29	−0.44	−0.49
MS	−0.02	−0.14	−0.16	−0.19
$\bar{y}_i$	−0.26	−0.48	−0.49	−0.57
$e(i)$	−0.12	−0.17	−0.25	−0.28
k_i	0.38	0.19	0.07	0.05
$med._i(k_u)$	−0.18	−0.13	−0.24	−0.24
f_i $k_u=1$	0.16	0.14	0.24	0.27
Birank	0.37	0.17	0.03	0.03
$\Cup_{NP}$	1.00	0.52	1.00	0.76
$\Cup_{SP}$	0.52	1.00	0.76	1.00

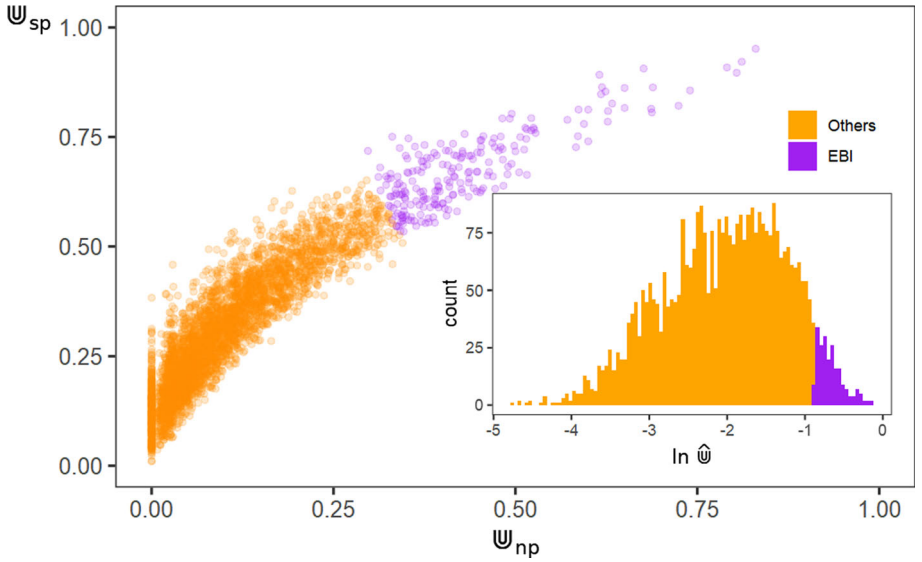

Fig. 2 Densities of polarity scores across items with $k_i > 100$ **in the two clusters**. Most of the probability mass of $\hat{\Cup}(y_i)$ is concentrated between $\exp(-1) = .36$ and $\exp(-3) = .05$

- EBIs can only be matched with items of the same class (movies with movies, etc.).
- Nearest Neighbour Search (NNS) is performed among the eligible items, aimed at minimising the multivariate Mahalanobis distances for 9 control covariates: US, MS, $\bar{y}_i$, $e(i)$, k_i, $med._i(k_u)$, f_i $k_u=1$, Birank, and the year of publication

The NNS algorithm converged towards satisfying results, except for $med._i(k_u)$ which is persistently lower in EBIs (Table 2). In the 218 matched items, the index of controversy is higher than in the pool of 3,759 candidates, but it is still lower than in EBIs, which are characterised as extreme cases of bi-polarity.

Table 2 Evaluation of the matching results

		All	EBIs	Matched
	#	3,759	218	218
$\sum$	k_i	944,769	93,162	85,962
$avg.$	$\underset{\text{NP}}{\cup}(i)$	0.10	0.43	0.20
$avg.$	$\underset{\text{SP}}{\cup}(i)$	0.30	0.67	0.47
$avg.$	$\cup(i)$	0.14	0.52	0.27
$avg.$	US	7.35	5.45	5.79
$avg.$	$\bar{y}_i$	7.22	5.23	5.59
$avg.$	MS	7.35	5.45	5.79
$avg.$	$e(i)$	7.18	6.56	6.66
$avg.$	$med._i(k_u)$	13.97	3.63	4.75
$avg.$	$\underset{k_u=1}{f_i}$	0.21	0.34	0.31
$avg.$	$-\ln(\text{Birank})$	10.73	10.70	10.70
$med.$	year	2013	2017	2016

4.4 Findings

Compared to the numbers of k_i reported in Table 2, only 78, 062 reviews for EBIs and 74, 782 reviews for the matched items are in English[2]. Only these are considered, further reducing the discrepancy between the two groups. In the aggregate *corpus* of 152, 844 reviews, the tokens consisting of less than three symbols have been filtered out, alongside stopwords. As a result, η has been estimated for 100, 952 tokens.

Patterns emerge looking at the top scoring 220 token semantic (Table 3): the second token most associated with EBIs is “bomb", which is linked to the aforementioned concept of review bombing, as confirmed by the presence of “bomber" in the first column of Table 3. It is a signal that users commenting on highly controversial items are aware of participating in a controversial discussion and are often semantically reactive in their reviews, mentioning others’ behaviours. Reactive social behaviour emerges from the concepts related to main tokens, too. All tokens with the highest η point to concern towards the veracity of content in EBI: “haters", “propaganda", “troll", and “fake". These words signal the concern against correct information being altered (“misinformation, 4^{th} column, $\eta = 4.61$") by someone else (a ‘bomber’, a ‘hater, a ‘troll’, a ‘fake’); in this sense, the user reviews the EBI having the past actions of someone else in mind (Cantone et al., 2024). The presence of tokens for social media“Twitter” and “reddit” and “journalist” is pairwise noteworthy.

Other topics can be identified:

- USA Politics and religions: tokens such as “propaganda", “liberal", “leftist", “conservative", “republican" are part of the USA political jargon. Other words are political in nature, such as “election", “vote", “agenda", and “politic". “Hillary" “Clinton" is directly mentioned among the tokens of Table 3. “Russian" could be related given the high η, alongside “west". Other words that are less political but are allusive of a semantic of religious morality: “christian", “saint", “abortion", “Jesus", “atheist", “muslim", “Christ"; but also “abortion", which is actually a relevant topic of debate in current USA politics.

[2] Identified with the *R* package `cld3`, Google Compact Language Detector.

Table 3 Top 220 token associated with EBI

Token	η	Token	η	Token	η	Token	η	Token	η
haters	17.29	law	7.31	arkham	5.77	dmc	4.88	chord	4.32
bomb	15.37	agenda	7.30	captain	5.75	who've	4.87	chick	4.31
propaganda	15.33	pve	7.13	innocent	5.73	Charlotte	4.87	God	4.30
troll	14.61	pretend	7.00	accessory	5.72	april	4.85	atheist	4.29
fake	13.15	goty	6.99	laugh	5.71	outrage	4.84	poser	4.29
child	12.78	loser	6.97	Lily	5.67	unfair	4.83	airport	4.29
ban	12.73	tasteless	6.96	unoriginal	5.63	entitle	4.80	rap	4.27
theory	12.38	support	6.92	pathetic	5.62	punk	4.79	boob	4.26
house	11.64	Keanu	6.92	footage	5.52	study	4.76	Ariana	4.23
cute	11.06	whine	6.90	journalist	5.52	whore	4.76	language	4.22
sex	10.70	argument	6.72	anti	5.50	idiot	4.75	muslim	4.21
overpriced	10.69	leftist	6.67	baby	5.50	debate	4.75	ward	4.21
gta	10.22	debut	6.62	sane	5.43	Jesus	4.74	Clinton	4.19
activity	10.16	Twitter	6.62	museum	5.41	dinosaur	4.72	nazis	4.19
looter	10.16	vote	6.56	streamer	5.39	motorsport	4.72	orc	4.19
liberal	10.15	riot	6.51	conservative	5.39	freely	4.68	anthem	4.17
fanboys	10.10	sexual	6.51	republican	5.37	butthurt	4.67	politic	4.17
disgust	9.95	Justin	6.51	election	5.33	abortion	4.64	gran	4.17
west	9.87	trash	6.48	racism	5.33	misinformation	4.61	Mario	4.16
article	9.69	premium	6.46	platformer	5.33	racist	4.58	promote	4.16
documentary	9.65	bow	6.38	platforming	5.29	nay	4.57	submit	4.16
zero	9.64	flop	6.35	artstyle	5.26	inform	4.56	Portuguese	4.16

Table 3 continued

Token	η	Token	η	Token	η	Token	η	Token	η
Sony	9.63	cat	6.32	skew	5.26	provoke	4.55	Cornell	4.14
girl	9.45	educate	6.30	bill	5.24	driver	4.54	funny	4.14
Russian	9.26	democratic	6.29	service	5.24	toxicity	4.53	burger	4.13
terrorist	9.24	lol	6.25	sweetener	5.22	childish	4.51	deny	4.13
grindy	9.16	exp	6.24	ignorant	5.16	democrat	4.50	Russia	4.12
bomber	9.12	beg	6.22	truth	5.15	wannabe	4.50	voter	4.12
island	8.55	reeve	6.21	Minaj	5.15	Dante	4.48	chopper	4.11
Christian	8.49	offend	6.16	Ashlee	5.15	breathtaking	4.48	homophobic	4.10
cyberpunk	8.44	coaster	6.12	mmr	5.15	theft	4.48	traffic	4.10
saint	8.27	lie	6.08	mobas	5.14	aircraft	4.47	resin	4.09
parent	8.21	moron	6.07	pedestrian	5.09	combo	4.45	lmao	4.08
forza	8.05	Hillary	5.94	hardware	5.07	kat	4.44	warner	4.08
medium	7.93	expose	5.91	cry	5.05	balan	4.43	spec	4.08
hat	7.92	moore	5.91	singer	5.05	animal	4.43	shave	4.05
controversy	7.90	nickelback	5.90	gay	5.03	company	4.42	hbo	4.04
ticket	7.84	data	5.89	president	5.01	metacritic	4.41	wii	4.02
hunt	7.71	predatory	5.84	greed	4.97	cooperate	4.40	Christ	4.01
batman	7.70	penis	5.83	sensitive	4.95	woman	4.39	Billie	4.01
immature	7.63	offensive	5.81	reaper	4.91	shower	4.38	didn't	4.00
Bieber	7.59	geforce	5.78	hate	4.88	bias	4.37	gaiden	4.00
rockstar	7.57	adult	5.77	blizzard	4.88	porn	4.35	Reddit	4.00
Nicki	7.34	endgame	5.77	someday	4.88	turismo	4.34	distort	4.00

Table 4 Elements for a sensitivity analysis

Issue	Alternative
How to measure controversy?	Alternative, more complex measures
$Med(k_i)$ in the sample is very low	Consider a lower filter for k_i
How to define extreme bi-polarity?	A different threshold
Matching's diagnostic are not perfect	Alternative distances
Topics are not automatically inferred	Pre-trained topic modelling

- Brands: "gta", "sony", "cyberpunk", "batman", "bieber", "rockstar", "nicki", "geforce", "blizzard", "balan", "metacritic", "hbo", "wii" are tokens for brands.
- Sexuality: a minor but robust topic, it connects words related to sexual activities or sexual slurs, examples are "sex", "penis", "adult", "gay", "whore", "porn", "homophobic", "boob". In this context, it is peculiar that all of "girl", "woman" and "chick" are in the top 220, but not "boy" or "man".

5 Discussion and limitations

This study presents a method and its application. Given any clustering criterion it can be easily extended to provide insightful descriptive statistics concerning the semantic association to any identifiable small cluster of collections of short texts with large corpus of short texts, and in this regard it adapt very well for a large array of application in computational approaches to social sciences and new media.

Nevertheless, three notable limitations should be addressed. Firstly, the method's ability to be helpful for causal assessment of the highlighted semantic association is uncertain. While methods of statistical matching try to mirror procedures of experimental control, the quality of results depends on the availability of key covariates. High-scoring tokens suggest significant semantic differences between EBI and non-EBI, η alone is not enough to make a causal claim about the relationship between these topics and EBI's phenomenology. This is because the causal direction of the relationship is not uniquely identifiable since scores and reviews are submitted concurrently. To address this, further learning sub-procedures could be implemented for the validation of tokens' along the time span of the item's review history (Egami et al., 2022).

A second limitation of the application concerns the external validity of its findings, specifically, the generalisability of the semantics associated with EBI to other online platforms of user reviews beyond Metacritic. The user demographics on Metacritic primarily consist of young males residing in the USA with interests in science fiction and video games. As a result, the emerging topics in this study may not necessarily hold for EBIs across different cultures and languages, even if the variation in terminological expression were taken into account (e.g. *izquierda* instead of *liberal*). However, given the convergence with previous literature in EBIs (Cantone et al., 2024), some of the emerging topics in this study could still hold general validity for EBIs, but this must be further investigated in future studies.

A third limit concerns the irreducible degrees of freedom of the method, that could condition the validity of the findings in the application. These are listed in Table 4 and shortly commented.

5.1 On the measurement of controversy

The application developed an operative definition for an index controversy which emerged by an evolution of the specific literature Schoenmueller et al. (2020). This definition is slightly different from the parametric characterisation of bi-modality prevalent in Psychometrics (Knapp, 2007; Pfister et al., 2013; Tang et al., 2022). With alternative parametric methods, statisticians derive a parameter of overdispersion for a mixture model of the score (Iannario, 2014), which in Piccolo and Simone (2019) is conceptually equated as a measure of the statistical entropy in the decision making. Econometrics has another different parametric approach that does not assume the duality of polarity and allows multi-polarity (Esteban and Ray, 1994; Duclos et al., 2004; Deutsch et al., 2013).

5.2 Filtering low-reviewed items and clustering EBIs

Filtering out $\sim$ 37,000 may seem like a huge loss of information, but it improves the reliability of η avoiding accounting for items that technically are EBIs, but there is no real public involvement in them - so they do not qualify as truly controversial item.[3] The main issue in filtering is that k_i may potentially follow a scale-free distribution: it does not grow up linearly and does not distribute around a central value, so even non-parametric indexes as the median are only relatively informative (Barabási, 2009; Holme, 2019).

More concerning is the determination of the exact boundaries of the EBIs cluster. Figure 2 shows that polarity follows a logarithmic bell curve and EBIs are the right tail (95^{th} percentile). But not all EBIs are equally bi-polar. For example in Fig. 2 can be noticed that a small micro-cluster of 26 items are much more bi-polar than others, and well-separated from the centre of the EBI cluster. Inference could have been restricted to only those 26, but a larger sample helps for an accurate assessment.

5.3 Alternative matching algorithms

Compared to more elaborated alternatives such as the Optimal Matching algorithm (Hansen and Klopfer, 2006) or Coarse Exact Matching (Iacus et al., 2012) which preserve global optima of distances through matching multiple controls or pruning cases out of the sample, NNS is considered a “greedy" algorithm: and it does not condition a match on the expected effects of reducing the pool of available items for the subsequent EBIs. In this application, greediness is not an issue because there is a large pool of 3,759 items to pair with only 218 EBIs, hence the effect of reducing the pool of candidates for each subsequent matching is negligible.

The Mahalanobis distance is preferred to the alternative Propensity Scores because in literature Mahalanobis is considered a less biased approach for a low number of control covariates (Stuart, 2010; King and Nielsen, 2019).

[3] For example, if an item received only 20 reviews from regular users, it is relatively easy to astroturf other 20 reviews to improve $\bar{y}_i$. $\Doublecup$ is relatively robust to these phenomena, but in general, bi-polarity is associated with centrality, see Table 1.

5.4 Topic modelling

The application still requires a human interpreter of the adjusted association, who understands the hidden semantic patterns behind the findings of the text mining procedure. Table 3 shows 220 tokens, but the model estimated η for more than $100,000 tokens$, of which more than 99% are unrelated to the phenomenology of EBIs. Such richness of results could still be processed automatically by large pre-trained models of language (Lee et al., 2020; Qiang et al., 2022).

6 Final Comments

This study concerns the general feasibility of a methodological design to control for the effects of covariates in the estimation of the association between the semantics implied in online reviews and numeric properties as the polarity of their scores. Results confirm and expand the validity of preliminary results in (Cantone et al., 2024), establishing an incontrovertible association of review bombing to hyper bi-polarisation of reviews in Metacritic, over alternative explanation as astroturfing.

This design suits analysis on data retrieved from platforms of online reviews, where all of these features recur; but it can be extended or slightly adjusted to account for similar applications. For example, just adjusting the strategy for identification of the covariates, this methodological design would suit the semantic analysis of the content of nodes of direct networks, like citation networks, whereas nodes would be textual documents (Light et al., 2021).

Appendix: Correlation matrixes

These 8 indicators are separated by the latent dimensions that they aim to define: quality and centrality of the items. It is expected that indicators of the same latent concept are correlated, and indicators of different concepts are not. With minor exceptions, the latter hypothesis is verified. Establishing if indicators within the same group are measuring the same latent concept is less straightforward. Noteworthy is that direct centrality k_i and indirect centrality

Table 5 Kendall correlations on the whole sample ($n = 40,665$)

	US	MS	$\bar{y}_i$	$e(i)$	k_i	$med._i(k_u)$	$f_{i\,k_u=1}$	Birank
US	–	0.36	0.54	0.22	0.09	−0.06	0.03	0.11
MS	0.36	–	0.31	0.10	0.10	−0.07	0.00	0.10
$\bar{y}_i$	0.54	0.31	–	0.29	−0.02	−0.09	0.03	0.01
$e(i)$	0.22	0.10	0.29	–	−0.03	−0.10	0.03	0.00
k_i	0.09	0.10	−0.02	−0.03	–	−0.18	0.22	0.82
$med._i(k_u)$	−0.06	−0.07	−0.09	−0.10	−0.18	–	−0.49	−0.27
$f_{i\,k_u=1}$	0.03	0.00	0.03	0.03	0.22	−0.49	–	0.33
Birank	0.11	0.10	0.01	0.00	0.82	−0.27	0.33	–

Table 6 Kendall correlations when $k_i > 65$ ($n = 3,978$)

	US	MS	$\bar{y}_i$	$e(i)$	k_i	$med._i(k_u)$	f_i $k_u=1$	Birank
US	–	0.37	0.78	0.39	0.03	0.07	−0.10	0.07
MS	0.37	–	0.34	0.16	0.15	−0.04	−0.03	0.12
$\bar{y}_i$	0.78	0.34	–	0.43	0.02	0.02	−0.08	0.05
$e(i)$	0.39	0.16	0.43	–	−0.03	0.05	0.06	0.04
k_i	0.03	0.15	0.02	−0.03	–	−0.16	0.12	0.77
$med._i(k_u)$	0.07	−0.04	0.02	0.05	−0.16	–	−0.68	−0.09
f_i $k_u=1$	−0.10	−0.03	−0.08	0.06	0.12	−0.68	–	0.11
Birank	0.07	0.12	0.05	0.04	0.77	−0.09	0.11	–

$med._i(k_u)$ are negatively correlated, while Birank index (third order centrality) is correlated to k_i. These correlations are likely a side effect of the prevalence users with $k_u = 1$.

Acknowledgements Giulio Giacomo Cantone declares the present as a product of the Research Project of National Relevance of the Italian Ministry of University and Research, titled "POPSHPERE - Post-truth politics and the resilience of the public sphere in Europe". CINECA code: 2022FA5YPL. ID code: E53D23006690006. Venera Tomaselli thanks the European Union - NextGenerationEU, Mission 4, Component 2, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 - CUP E63C22002120006).

Funding The authors declare that no funds, grants, or other support were received during the preparation of the manuscript.

Declarations

Conflict of interest The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

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

References

Amendola, L., Marra, V., & Quartin, M. (2015). The evolving perception of controversial movies. *Palgrave Communications, 1*(1), 1–9. https://doi.org/10.1057/palcomms.2015.38

Anderson, E. T., & Simester, D. I. (2014). Reviews without a Purchase: Low Ratings, Loyal Customers, and Deception. *Journal of Marketing Research, 51*(3), 249–269. https://doi.org/10.1509/jmr.13.0209

Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proceedings of the National Academy of Sciences 106(51):21,544–21,549. https://doi.org/10.1073/pnas.0908800106

Barabási, A. L. (2009). Scale-Free Networks: A Decade and Beyond. *Science, 325*(5939), 412–413. https://doi.org/10.1126/science.1173299

Brandes, L., Godes, D., & Mayzlin, D. (2022). Extremity Bias in Online Reviews: The Role of Attrition. *Journal of Marketing Research, 59*(4), 675–695. https://doi.org/10.1177/00222437211073579

Cantone, G. G., Tomaselli, V., & Mazzeo, V. (2024). *Review bombing: ideology-driven polarisation in online ratings?* The case study of The Last of Us Part II: Quality & Quantity. https://doi.org/10.1007/s11135-024-01981-z

Charkhabi, M. (2014). Adjustments to propensity score matching for network structures. In: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), pp 628–633, https://doi.org/10.1109/ASONAM.2014.6921651

Day, K., & Kim, J. M. (2022). Investigating polarisation in critic and audience review scores via analysis of extremes, medians, averages, and correlations. *International Journal of Environment, Workplace and Employment, 1*(1), 1. https://doi.org/10.1504/IJEWE.2022.10051729

Deutsch, J., Fusco, A., & Silber, J. (2013). The BIP Trilogy (Bipolarization, Inequality and Polarization): One Saga but Three Different Stories. Economics 7(1). https://doi.org/10.5018/economics-ejournal.ja.2013-22

Dewan, S., Ho, Y. J. I., & Ramaprasad, J. (2017). Popularity or Proximity: Characterizing the Nature of Social Influence in an Online Music Community. *Information Systems Research, 28*(1), 117–136. https://doi.org/10.1287/isre.2016.0654

Dong, X., Xu, J., Bu, Y., et al. (2021). Beyond correlation: Towards matching strategy for causal inference in Information Science. *Journal of Information Science*. https://doi.org/10.1177/0165551520979868

Duclos, J. Y., Esteban, J., & Ray, D. (2004). Polarization: Concepts, Measurement. *Estimation. Econometrica, 72*(6), 1737–1772. https://doi.org/10.1111/j.1468-0262.2004.00552.x

Egami, N., Fong, CJ., Grimmer, J., et al. (2022). How to make causal inferences using texts. Science Advances 8(42):eabg2652. https://doi.org/10.1126/sciadv.abg2652

Esteban, J. M., & Ray, D. (1994). On the Measurement of Polarization. *Econometrica, 62*(4), 819–851. https://doi.org/10.2307/2951734

Ferrara, E., Varol, O., Davis, C., et al. (2016). The rise of social bots. *Communications of the ACM, 59*(7), 96–104. https://doi.org/10.1145/2818717

Fisher, M., Newman, G. E., & Dhar, R. (2018). Seeing Stars: How the Binary Bias Distorts the Interpretation of Customer Ratings. *Journal of Consumer Research, 45*(3), 471–489. https://doi.org/10.1093/jcr/ucy017

Fu, J. R., Ju, P. H., & Hsu, C. W. (2015). Understanding why consumers engage in electronic word-of-mouth communication: Perspectives from theory of planned behavior and justice theory. *Electronic Commerce Research and Applications, 14*(6), 616–630. https://doi.org/10.1016/j.elerap.2015.09.003

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature, 57*(3), 535–574. https://doi.org/10.1257/jel.20181020

Graham, B., & Paula, A. (Eds.). (2019). *The Econometric Analysis of Network Data* (1st ed.). United States: Academic Press, London, United Kingdom San Diego, CA.

Hansen, B. B., & Klopfer, S. O. (2006). Optimal Full Matching and Related Designs via Network Flows. *Journal of Computational and Graphical Statistics, 15*(3), 609–627. https://doi.org/10.1198/106186006X137047

He, X., Gao, M., Kan, M. Y., et al. (2017). BiRank: Towards Ranking on Bipartite Graphs. *IEEE Transactions on Knowledge and Data Engineering, 29*(1), 57–71. https://doi.org/10.1109/TKDE.2016.2611584

Ho, D. E., Imai, K., King, G., et al. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis, 15*(3), 199–236. https://doi.org/10.1093/pan/mpl013

Holme, P. (2019). Rare and everywhere: Perspectives on scale-free networks. *Nature Communications, 10*(1), 1016. https://doi.org/10.1038/s41467-019-09038-8

Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped distribution of product reviews. *Communications of the ACM, 52*(10), 144–147. https://doi.org/10.1145/1562764.1562800

Hu, N., Pavlou, PA., & Zhang, J. (2017). On self-selection biases in online product reviews. MIS Quarterly 41(2):449–471. https://doi.org/10.25300/MISQ/2017/41.2.06

Iacus, S. M., King, G., & Porro, G. (2012). Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis, 20*(1), 1–24. https://doi.org/10.1093/pan/mpr013

Iannario, M. (2014). Modelling Uncertainty and Overdispersion in Ordinal Data. *Communications in Statistics - Theory and Methods, 43*(4), 771–786. https://doi.org/10.1080/03610926.2013.813044

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 171*(2), 481–502. https://doi.org/10.1111/j.1467-985X.2007.00527.x

Janosov, M., Battiston, F., & Sinatra, R. (2020). Success and luck in creative careers. *EPJ Data Science, 9*(1), 9. https://doi.org/10.1140/epjds/s13688-020-00227-w

Kasper, P., Koncar, P., Santos, T., et al. (2019). On the Role of Score, Genre and Text in Helpfulness of Video Game Reviews on Metacritic. In: 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), pp 75–82, https://doi.org/10.1109/SNAMS.2019.8931866

King, G., & Nielsen, R. (2019). Why Propensity Scores Should Not Be Used for Matching. *Political Analysis, 27*(4), 435–454. https://doi.org/10.1017/pan.2019.11

Knapp, T. (2007). Bimodality Revisited. Journal of Modern Applied Statistical Methods 6(1). https://doi.org/10.22237/jmasm/1177992120

Kumar, S., Cheng, J., Leskovec, J., et al. (2017). An Army of Me: Sockpuppets in Online Discussion Communities. In: Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, WWW '17, pp 857–866, https://doi.org/10.1145/3038912.3052677

de Langhe, B., Fernbach, P. M., & Lichtenstein, D. R. (2016). Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings. *Journal of Consumer Research, 42*(6), 817–833. https://doi.org/10.1093/jcr/ucv047

Latapy, M., Magnien, C., & Vecchio, N. D. (2008). Basic notions for the analysis of large two-mode networks. *Social Networks, 30*(1), 31–48. https://doi.org/10.1016/j.socnet.2007.04.006

Lee, H. A., Choi, A. A., Sun, T., et al. (2021). Reviewing Before Reading? An Empirical Investigation of Book-Consumption Patterns and Their Effects on Reviews and Sales. *Information Systems Research, 32*(4), 1368–1389. https://doi.org/10.1287/isre.2021.1029

Lee, L. W., Dabirian, A., McCarthy, I. P., et al. (2020). Making sense of text: artificial intelligence-enabled content analysis. *European Journal of Marketing, 54*(3), 615–644. https://doi.org/10.1108/EJM-02-2019-0219

Lehmann, S., Schwartz, M., & Hansen, LK. (2008). Biclique communities. Physical Review E 78(1):016,108. https://doi.org/10.1103/PhysRevE.78.016108

Li, R., Li, Y. Q., Ruan, W. Q., et al. (2023). Sentiment mining of online reviews of peer-to-peer accommodations: Customer emotional heterogeneity and its influencing factors. *Tourism Management, 96*(104), 704. https://doi.org/10.1016/j.tourman.2022.104704

Li, X., & Hitt, L. M. (2008). Self-Selection and Information Role of Online Product Reviews. *Information Systems Research, 19*(4), 456–474. https://doi.org/10.1287/isre.1070.0154

Light, R., Theis, N., Edelmann, A., et al. (2021). Clouding climate science: A comparative network and text analysis of consensus and anti-consensus scientists. *Social Networks*. https://doi.org/10.1016/j.socnet.2021.11.007

Lu, L., Wu, L., & He, Z. (2020). Is Your Restaurant Worth the Risk? A Motivational Perspective on Reviews' Rating Distribution and Volume. *Journal of Hospitality & Tourism Research, 44*(8), 1291–1317. https://doi.org/10.1177/1096348020944537

Mayzlin, D., Dover, Y., & Chevalier, J. (2014). Promotional Reviews: An Empirical Investigation of Online Review Manipulation. *American Economic Review, 104*(8), 2421–2455. https://doi.org/10.1257/aer.104.8.2421

Morgan, S. L., & Winship, C. (2014). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Analytical Methods for Social Research: Cambridge University Press, Cambridge.

Petrescu, M., Kitchen, P., Dobre, C., et al. (2022). Innocent until proven guilty: suspicion of deception in online reviews. *European Journal of Marketing, 56*(4), 1184–1209. https://doi.org/10.1108/EJM-10-2019-0776

Pfister, R., Schwarz, K., Janczyk, M., et al. (2013). Good things peak in pairs: a note on the bimodality coefficient. *Frontiers in Psychology,4*. https://doi.org/10.3389/fpsyg.2013.00700

Piccolo, D., & Simone, R. (2019). The class of cub models: statistical foundations, inferential issues and empirical evidence. *Statistical Methods & Applications, 28*(3), 389–435. https://doi.org/10.1007/s10260-019-00461-1

Qiang, J., Qian, Z., Li, Y., et al. (2022). Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. *IEEE Transactions on Knowledge and Data Engineering, 34*(3), 1427–1445. https://doi.org/10.1109/TKDE.2020.2992485

Ratkiewicz, J., Conover, M., Meiss, M., et al. (2011). Truthy: mapping the spread of astroturf in microblog streams. In: Proceedings of the 20th international conference companion on World wide web. Association

for Computing Machinery, New York, NY, USA, WWW '11, pp 249–252, https://doi.org/10.1145/1963192.1963301

Santos, T., Lemmerich, F., Strohmaier, M., et al. (2019). What's in a Review: Discrepancies Between Expert and Amateur Reviews of Video Games on Metacritic. Proceedings of the ACM on Human-Computer Interaction 3(CSCW):140:1–140:22. https://doi.org/10.1145/3359242

Schoenmueller, V., Netzer, O., & Stahl, F. (2020). The Polarity of Online Reviews: Prevalence, Drivers and Implications. *Journal of Marketing Research, 57*(5), 853–877. https://doi.org/10.1177/0022243720941832

Seawright, J., & Gerring, J. (2008). Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options. *Political Research Quarterly, 61*(2), 294–308. https://doi.org/10.1177/1065912907313077

Sharkey, A., Kovács, B., & Hsu, G. (2023). Expert Critics, Rankings, and Review Aggregators: The Changing Nature of Intermediation and the Rise of Markets with Multiple Intermediaries. *Academy of Management Annals, 17*(1), 1–36. https://doi.org/10.5465/annals.2021.0025

Silge, J., & Robinson, D. (2017). Text Mining With R: A Tidy Approach. Oreilly & Associates Inc, Beijing ; Boston

Steiner, P. M., Cook, T. D., Li, W., et al. (2015). Bias Reduction in Quasi-Experiments With Little Selection Theory but Many Covariates. *Journal of Research on Educational Effectiveness, 8*(4), 552–576. https://doi.org/10.1080/19345747.2014.978058

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1–21. https://doi.org/10.1214/09-STS313

Stöckli, D. R., & Khobzi, H. (2021). Recommendation systems and convergence of online reviews: The type of product network matters! *Decision Support Systems, 142*(113), 475. https://doi.org/10.1016/j.dss.2020.113475

Tang, T., Ghorbani, A., Squazzoni, F., et al. (2022). Together alone: a group-based polarization measurement. *Quality & Quantity, 56*(5), 3587–3619. https://doi.org/10.1007/s11135-021-01271-y

Watson, F., & Wu, Y. (2022). The Impact of Online Reviews on the Information Flows and Outcomes of Marketing Systems. *Journal of Macromarketing, 42*(1), 146–164. https://doi.org/10.1177/02761467211042552

Wu, Y., Ngai, E. W. T., Wu, P., et al. (2020). Fake online reviews: Literature review, synthesis, and directions for future research. *Decision Support Systems, 132*(113), 280. https://doi.org/10.1016/j.dss.2020.113280

Yang, KC., Aronson, B., Ahn, YY. (2020). BiRank: Fast and Flexible Ranking on Bipartite Networks with R and Python. Journal of Open Source Software 5(51):2315. https://doi.org/10.21105/joss.02315

Yang, KC., Aronson, B., Odabas, M., et al. (2022). Comparing measures of centrality in bipartite patient-prescriber networks: A study of drug seeking for opioid analgesics. PLOS ONE 17(8):e0273,569. https://doi.org/10.1371/journal.pone.0273569

Zhuang, M., Cui, G., & Peng, L. (2018). Manufactured opinions: The effect of manipulating online product reviews. *Journal of Business Research, 87*, 24–35. https://doi.org/10.1016/j.jbusres.2018.02.016

Ziser, Y., Webber, B., & Cohen, S. B. (2023). Rant or rave: variation over time in the language of online reviews. *Language Resources and Evaluation, 57*(3), 1329–1359. https://doi.org/10.1007/s10579-023-09652-5