Reproduction and Replication of the Paper: "Rant or Rave: Variation Over Time in the Language of Online Reviews"

Project Report for the Lecture
“Social Media Data Analysis”

submitted by
Gina-Maria Unger
[deleted]

at the

Universität Konstanz

Faculty of Law, Economics and Politics

Department of Politics and Public Administration

Examiner: Professor Doctor David Garcia

Konstanz, 2023

Table of Content

List of Figures

List of Tables

1. Motivation

“The only languages which do not change are dead ones.”

is a striking quote made by the British linguist and author David Crystal (2007), that refers to the constantly changing nature of languages. Simply as societies evolve, their languages do too. This fact seems to hold true irrespective of the context, it might be alluding to the language we use in day-to-day conversation, the one we stick to in writing, or the language we use on the internet. That implies that given the context a language can evolve also in different directions, for instance, “gamer-slang” like “AFK” (abbreviation for “away from keyboard”) is not commonly used in day-to-day conversations.

Unsurprisingly, there are many ways to observe this evolutionary process, whereby Ziser et al. (2023) chose to concentrate their efforts on language change trends affecting the sentiment intensity, lexical diversity, and comprehensiveness of online reviews. For matters of definition, an online review is an expressed opinion about “a product or service made by a consumer who has experienced a service or purchased a product” (D'Acunto et al., 2020). According to Ziser et al. (2023) studying these reviews in the context of language change is especially promising since they take a growing role in a digitalized consumer society, express the present but also allow for observing long-term trends due to the long existence of review platforms. In addition, review data often comes with further information that allows to couple it to an underlying sentiment and hence permits researchers to track language changes down more narrowly by controlling for this factor. In summary, Ziser et al.’s (2023) findings suggest that reviews have become sentimentally more intense, less lexically diverse, and less comprehensive.

Following the author’s footsteps, the underlying work will focus on the research questions:

1. *Are the Ziser et al.’s (2023) findings actually reproducible?*
2. *If so, can the found trends of language change be replicated on the basis of different more contemporary data and do they have a continuing character?*

For tackling the first question, one of the three by Ziser et al. (2023) selected data sets, containing reviews of the online movie database IMDb, will be used. Further, to address the second question, a more recent data set containing reviews from the business review website Yelp will be used. Here the answering of this second question is motivated by its scientific and practical relevance. The scientific motivation is given by the argument of good research practice, that incorporates the continuous testing of the reproducibility of scientific results to prevent advancing replication crises like the one in the field of psychology (see Wiggins & Christopherson, 2019). Whereas the practical motives are given by the fact that the correctness and possible continuing nature of Ziser et al.’s (2023) results have implications for the readers and writers of online reviews, as well as for the platforms hosting them. For example, if older reviews of a product are less one-sided a mindful reader should consider reading them more attentively. Giving a small spoiler, this work indeed confirms the reproducibility of Ziser et al.’s (2023) results about language trends and that these trends also show and likely continue in other, more recent, data.

To get an overview of this work, the next section will first give a short introduction to the topic of language change trends in the online context. Then the two subsequent sections introduce the data selected for addressing the research questions, its preparation, as well as the analysis' methodological foundation. Consequently, after this, the results relying on this data and methodology will be presented and discussed in the light of the research questions while pointing also out possible limitations induced due to the gone through research process.

2. Background

Since the topic of trends in the online language over the past decades is quite broad the subsequent section will present a small literature outline addressing them to further a general understanding of the matter and give impressions of what concrete trends could look like. Additionally, a small selection of literature studying online reviews by relying on sentiment analysis will be displayed to show that Ziser et al.'s (2023) approach, thus the one of this work, is rather unique.

Remembering David Crystal's (2007) quote from the beginning of this work, one can read an other noteworthy in his 2001 book "Language and the Internet", that is "if the Internet is a revolution, therefore, it is likely to be a linguistic revolution" (Crystal, 2001). Considering that the World Wide Web was theoretically made accessible to everyone in 1993, one could say that it was still in its childhood in 2001, and yet, its potential to influence our language was noticeable from the start. Consistent with this the "online language" seems to have developed its own peculiarities, for instance, it is generally less grammatical (Eisenstein, 2013) and more uncivil (Santana, 2014). It also delivers the fertile soil for many neologisms, for example, Irfan's (2021) findings suggest that the communication on social media can lead to semantic changes in established words. In addition to the fact that general grammatical rules, norms, and the lexicon differ in the online space, the language itself also varies strongly depending on what part of the net community is observed. As an example, Bokányi et al. (2016) could relate language use in social networks to variables like "slang use, urbanization, travel, religion and ethnicity" that correlate highly with demographics.

Shedding more light on the research applying sentiment analysis in the context of online reviews shows that Ziser et al.'s (2023) research approach for studying language trends is quite unique since a lot of the related research seems to concentrate on the evaluation of customer satisfaction or preference identification. For instance, Bian et al. (2022) used a convolutional neural network with sentiment-based features to highlight which categories of a hotel review are of importance for reviewers. Logically such research is of great interest for platforms to optimize their offers respectively. Moreover, when taking preference extraction one step further and coming to the actual proposition of products to possible customers, sentiment analysis procedures also play an important role. Likewise, Zhang et al. (2022) developed a product selection model that incorporates measures such as the sentiment direction or intensity, and combines them with an intuitionistic fuzzy TODIM method to determine ranking results of alternative products. Meanwhile, Awajan et al. (2021) rely in addition to sentiment measures on neutrosophic set theory to rank alternative products. However, there does exist some literature that comes a bit closer to Ziser et al.'s (2023) work, which studies attributes associated with online reviews. For instance, Tripathi et al. (2021) extracted from online reviews attributes like the sentiment intensity to analyze patterns over the reviews' temporal order and one of their findings was that sentiment intensity diminishes with an increasing review order. At first glance, this finding seems to be contradicting to Ziser et al.'s (2023) findings, but one should consider that an increase in the review order must not be comparable to an actual long-time trend.

Relating Ziser et al.'s (2023) research on language trends in online reviews to the topics of the "Social Media Data Analysis" lecture and its tutorial, it can be connected mainly to the third lecture block "text in social media". This block introduced dictionary and supervised methods, as well as large language models, that can assist in solving tasks like emotion or sentiment detection for text documents. Moreover, the perquisites of some of these methods were discussed in this block, namely the generation of text or word-level embeddings. Being more precise, Ziser et al. (2023) heavily relied on dictionary methods for sentiment analysis by using the in the lecture introduced Vader sentiment lexicon (Hutto & Gilbert, 2014) in multiple ways to retrieve sentiment intensity, lexical diversity, and sentiment dichotomy metrics. Whereat depending on the metric of interest, the Vader lexicon gets used with or without modifiers or just assists as a summary of sentiment words. Besides that, the authors also fine-tuned the language model BERT (Devlin et al., 2019) for sentiment analysis and determined the accuracy score for the Vader-based sentiment predictions and BERT's to compare them.

3. Data

Before addressing the research questions of reproduction and replication suitable data for doing so must be found or collected. More precisely, the data requirements to make an investigation of the outlined questions possible are: (1) the data must incorporate textual reviews, (2) besides the textual reviews the data must include some form of ordered ratings that allow to control for sentiment, (3) the data must contain at least yearly information about the review publication dates, and (4) the data must include reviews of consecutive years.

For their work, Ziser et al. (2023) relied on three freely available data sets from the online review platforms Amazon (Books)[1], the Internet Movie Database (IMDb)[2], and Yelp[3]. As for the matters of reproduction working with just one of these data sets should be sufficient to show that the authors' results are reproducible, hence it will be focused only on the IMDb data set[4]. This data set is available as a Kaggle data set (Biswas, 2021) and comes split into six JSON files due to its total size of 7.78 GB. It consists of 5,571,499 units and 9 variables, where these variables are the review ID, the reviewer's name, the reviewed movie's name, the review's associated numerical rating, a summary of the review, the review publishing date, a binary tag if a review contains spoilers, the textual review itself, and an indicator if the review is helpful. Given the data requirements, the variables needed for further analysis steps are the ones stating the textual reviews, the publishing dates, and the numerical ratings. These numerical ratings range from 1.0 to 10.0. Timewise the data spans over the years 1998 to 2021, although

[1] Amazon Books refers to a retail division of the Amazon company and primarily focuses on selling various genres of books through the online platform. See https://www.amazon.de/

[2] IMDb, namely the Internet Movie Database, is an Amazon-owned online database that provides information about movies, television shows, actors, directors, and other industry professionals. See https://www.imdb.com/

[3] Yelp is an online platform and mobile app that enables users to discover and review local businesses, such as restaurants, cafes, shops, and services. See https://www.yelp.de/

[4] Accessible at https://www.kaggle.com/datasets/ebiswas/imdb-review-dataset

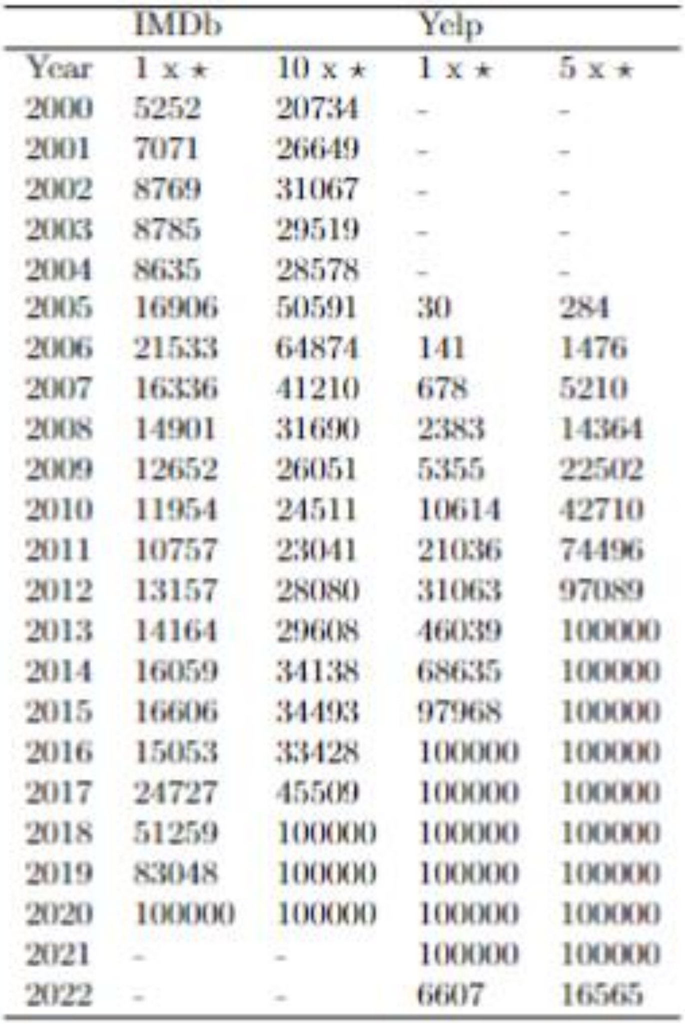
Ziser et al. (2023) indicated a 2000 to 2020 span, strangely that cannot be explained with any updates in the data as the data sets update frequency states “never”. Thus, it can be assumed that the authors had a reason to reduce the years represented in the dataset, which was probably the observance of just the last 20 years, nevertheless stating that more directly would have been preferable.

Moreover, when it comes to the research objective of result replication a newer data set fulfilling the outlined data requirements is a necessity. Here the choice fell on a more recent version of the Yelp review data set[5], which can be downloaded together with five other Yelp data sets from Yelp’s website. The review data set comes as a JSON file with a size of 5,216,669KB, and contains 6,990,280 units and 9 variables. These 9 variables state the review ID, the reviewer’s ID, the reviewed business’s ID, the review’s numerical star rating, the review posting date, the textual review, the number of useful votes a review got, the number of funny votes it received, and the number of cool votes it collected. Again, for matters of analysis, only the variables stating the textual reviews, the publishing dates, and the numerical star ratings are of importance. Giving further insights about the data, all the in the data set included reviews are recommended by Yelp, which means they have a greater credibility in terms of their authenticity than regular reviews. Further, the review’s numerical ratings can rank from 1.0 to 5.0 stars and the data set covers the more current period of 2005 to 2022.

Before the analysis, several data preparation steps were conducted. This preparation process implies in a first step that the columns containing the publishing date information get converted to some form of datetime format as they come initially as strings. Consequently, in a second additional step, a new column for each data set gets generated that stores the yearly publishing information extracted from the publishing dates. This procedure is necessary, as in the preceding analysis steps the data must be grouped based on these yearly values. Moreover, the data sets get reduced to those reviews that carry the numerically smallest or greatest rating scores, where the smallest score is for both 1.0 and the greatest either 10.0 or 5.0. Such a kind of data reduction permits one to control for the review sentiment when studying the overall effect of time on language use since one can observe long-term effects for homogenous review sentiment groups. Thus, to prevent any misunderstandings, in the following one should think of reviews with a numerical rating of 1.0 as reviews carrying a negative sentiment and those with a rating of 10.0 or 5.0 as carrying a positive. In addition, the IMDb data set gets minimized to reviews that were posted from 2000 to 2020 to be in line with Ziser et al. (2023).

Although these preparation processes reduce the data sets greatly, both are still of a considerable size afterwards, which would make the subsequent analysis very computationally demanding when proceeding simply without any adjustments. Thus, for each possible sentiment and year combination a random sample with replacement of 100,000 reviews gets drawn if the review frequency count for such a combination is greater than 100,000, if the respective count is smaller all reviews of the group end up in the final data. On this occasion, it should be stated, that when considering the reduction of data loss one should have preferred a random sampling procedure without replacement but that would not have been in line with Ziser et al. (2023), who refrained from these considerations. The final review frequency counts of the so-reduced IMDb and Yelp data sets per year and sentiment combination can be viewed in ***Table 1***:

[5] Accessible at https://www.yelp.com/dataset



	IMDb		Yelp	
Year	1 x ⋆	10 x ⋆	1 x ⋆	5 x ⋆
2000	5252	20734	-	-
2001	7071	26649	-	-
2002	8769	31067	-	-
2003	8785	29519	-	-
2004	8635	28578	-	-
2005	16906	50591	30	284
2006	21533	64874	141	1476
2007	16336	41210	678	5210
2008	14901	31690	2383	14364
2009	12652	26051	5355	22502
2010	11954	24511	10614	42710
2011	10757	23041	21036	74496
2012	13157	28080	31063	97089
2013	14164	29608	46039	100000
2014	16059	34138	68635	100000
2015	16606	34493	97968	100000
2016	15053	33428	100000	100000
2017	24727	45509	100000	100000
2018	51259	100000	100000	100000
2019	83048	100000	100000	100000
2020	100000	100000	100000	100000
2021	-	-	100000	100000
2022	-	-	6607	16565

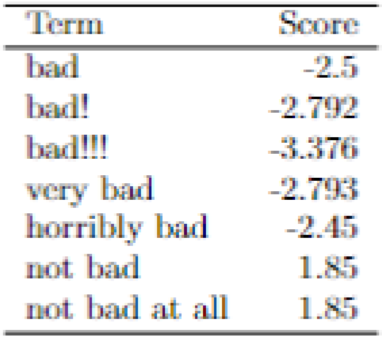
Table 1: The number of negative (one star) and positive (five stars for Yelp, ten stars for IMDb) reviews for each data set by year.

Here one can notice that both data sets seem to have similar tendencies in terms of review frequency counts, that is, the counts of the earlier years are sparser compared to the one of latter, which mostly maxed out the 100,000 sample size. The 2022 Yelp data seems to be an exception to this. Moreover, in general, it can be noted that there are typically fewer positive reviews than negative ones on average and on a yearly basis. On top of that the frequency counts of the earlier years of the Yelp data set are always smaller than their IMDb counterparts, hereby one should especially keep the small sample sizes of 2005 and 2006 in mind, which might induce some form of numerical outliers in the data.

4. Methods

Next, the methodological foundation of Ziser et al.'s (2023) work, which is due to the research objectives of reproduction and replication recycled in this work, will be outlined in the subsequent section. Thereby the first thing to tackle is to choose how to derive sentiment scores for reviews to make the observation of changes in the intensity for a fixed sentiment over time possible. This task can be addressed by relying on the Vader sentiment lexicon

(Hutto & Gilbert, 2014), which contains sentiment scores between −4 (highly negative) and 4 (highly positive) for more than 7.500 words. But instead of just relying on the simple lexicon one can also obtain sentiment intensity scores by using Vader’s enhanced version that implements heuristics which consider negation, punctuation, and degree modifiers. Generally speaking, given that word-level sentiment methods might not perform well on real-life examples since sentences get their semantic meaning from their coherent parts, which are often not just single words, that seems to be a reasonable extension. For instance, to get an impression of how the sentiment intensity score of a word like “bad” changes when introducing further textual information and working with the enhanced lexicon, one can have a glance at ***Table 2***.

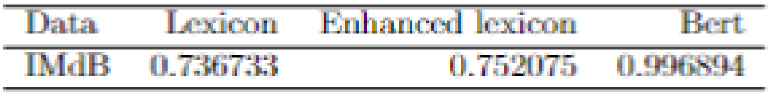


Term	Score
bad	-2.5
bad!	-2.792
bad!!!	-3.376
very bad	-2.793
horribly bad	-2.45
not bad	1.85
not bad at all	1.85

Table 2: Enhancement effects of the improved Vader lexicon for the word “bad”.

Not surprisingly adding a negation changes the score direction, while adding an exclamation increases the intensity. Hence, given Vader’s suitability for observing sentiment intensity changes, Ziser et al. (2023) and so this work will rely on it for the derivation of sentiment scores.

Nevertheless, one might ask how accurately Vader predicts a review’s actual sentiment or consequently the overall review sentiment since there might be a discrepancy between Vader’s prediction capability and performance. Therefore, one can exemplarily calculate the accuracy metric for both Vader versions for the IMDb data. The accuracy metric, as noted in the “Social Media Data Analysis” lecture, states the fraction of predictions a model got right, means closer values to 1 indicate a better model. Further, to grab an idea of how the lexicon-based method performs compared to supervised methods, one can also fit a latter. A reasonable comparison choice would be the natural language model BERT (Devlin et al., 2019), which is a semi-supervised learning method and a long-held gold standard when it comes to natural language processing tasks. For this reason, a BERT model gets fine-tuned for sentiment classification with 1,000 reviews and tested with 4,000, while the training and testing data set were randomly sampled with replacement from the IMDb data in such a manner that they contain equal-sized year-sentiment groups. Important to note, before starting the model fine-tuning the textual reviews must be converted to lowercase as it is a necessary preprocessing step for using BERT-like models. Given these methods, one can find their respective accuracy scores for the testing data in ***Table 3***:



Data	Lexicon	Enhanced lexicon	Bert
IMdB	0.736733	0.752075	0.996894

Table 3: Accuracy of different sentiment classifiers for the IMDb data set.

Unsurprising one can see that the enhanced lexicon predicts the sentiments better, as it includes context-aware features. In addition, BERT beats both variants of the lexicon-based method by far. That kind of method ranking conforms to Ziser et al.'s (2023) findings, however, the authors argue that supervised methods are not an optimal choice when it comes to addressing their and thus the outlined research questions, because they are highly dependent on the distribution of the labelled data used in the training process which makes answering time-related questions difficult.

Metrics

To approach the research questions the used metrics for measuring language trends should be defined. Namely, the metrics try to grab the reviews' sentiment intensity, language comprehensiveness and richness. Further, to evaluate possible trends for the groups of reviews with positive or negative sentiment over time, they get calculated on a yearly sentiment basis.

Word-level sentiment

The word-level sentiment metric gives a weighted average of the sentiment scores for each sentiment term that is part of a review r and included in a lexicon L, here the Vader lexicon. The metric can be derived as follows:

$$S_r = \left(\sum_{i=1}^{|L|} C_{rw_i} \times L_{w_i} \right) \Bigg/ \left(\sum_{i=1}^{|L|} C_{rw_i} \right)$$

Whereby S_r represents the overall sentiment score for review r, L_{w_i} word w_i's sentiment score given by lexicon L, and C_{rw_i} the count of the lexicon word w_i in review r. Since $|L|$ refers to the total size of the lexicon L, in each summation step the count of lexicon word w_i in the review r gets weighted by word w_i's sentiment score, means sentiment scores of words that are not part of the review do not influence its overall sentiment score S_r.

Enhanced word-level sentiment

The enhanced word-level sentiment metric introduces in the simple word-level sentiment metric some form of context awareness by considering the related parts of review r for the metric calculation. Ziser et al. (2023) state that it can be calculated by replacing $L_{w_i}$with $\bar{L}_{w_i}$, a score that reflects lexicon-based enhancements, what is unfortunately a very misleading discription as it suggests the reliance on some word-level enhanced sentiment scores. Viewing instead Ziser et al.'s (2023) coding files shows that the authors simply based their calculations on the 'compound' sentiment score provided by the Vader package, a metric that gives the normalized sum of all word sentiment scores in a text corrected by rule-based adjustments (Hutto & Gilbert, 2014).

Absolute sentiment

The absolute sentiment metric changes the simple metric in such a way, that it can be easily interpreted in terms of an absolute increase or decrease in the general sentiment intensity. This means higher values imply the use of stronger sentiment words independent of the underlying sentiment class. Therefore, the metric can be calculated by replacing word w_i's sentiment score given by lexicon L, L_{w_i}, with its absolute value $|L_{w_i}|$.

Informativeness

For deriving an estimate of the informativeness of a review one can approximate it with the review length, as a long review is expected to contain more details and information (Schindler & Bickart, 2012). Hence, one can assess the informativeness of a set of reviews by calculating their average number of words.

Dichotomy

For obtaining an estimate of the dichotomous character of reviews one can approximate it by the percentage of one-sided reviews, which means by the percentage of reviews that have "at least one sentiment term corresponding to their numerical sentiment rating and none of the opposite sentiment" (Ziser et al., 2023). For this endeavour, one can use the enhanced Vader lexicon. Specifically, someone can calculate the percentage of reviews that get assigned by Vader correctly to their underlying sentiment while it does not assign a sentiment score greater than 0 for their opposing sentiment. For example, an IMDb review with a negative sentiment should get assigned by the Vader package a "neg" score greater than 0 and a "pos" score no greater than 0, hereby the "neg" score gives the ratio of text proportions that fall into the negative sentiment category and "pos" into the positive (Hutto & Gilbert, 2014).

Lexical diversity

The lexical diversity of reviews can be approximated according to Ziser et al. (2023) by how often frequent sentiment words occur in them based on the reasoning that using fewer sentiment words leads to less diverse texts. More precisely, for studying the lexical diversity one can have a look at the "head" of a Zipfian distribution (Zipf, 1942) of the corpus words. Whereby, a Zipfian distribution is a distribution that describes a pattern where the frequency of elements in the data decreases as their rank increases, hence it is a distribution with a narrow "head" and a long "tail". Such a kind of distribution can be seen in ***Figure 1***:

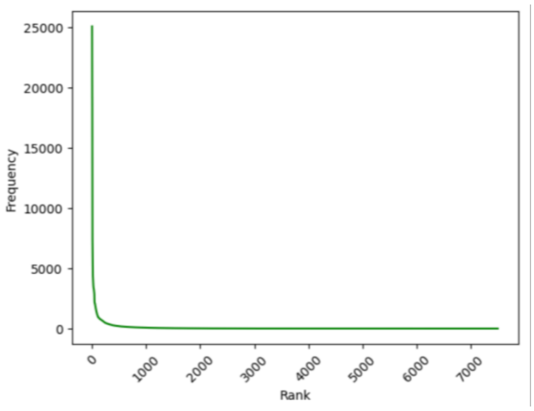


Figure 1: Zipfian distribution for the positive sentiment corpus of the 2020 IMDb data.

When considering the situation where the elements of the distribution are sentiment words, a narrower "head" implies that a small set of words gets repeated frequently over the corpus, which implies that it is less lexically diverse. For calculating a distribution's "head width" for a corpus of reviews, here the top 1% of the most frequent sentiment words, one can proceed as follows:

$$S_y = \left(\sum_{i=1}^{\left\lceil \frac{|L|}{100} \right\rceil} C_{w_i,y} \right) \Bigg/ \left(\sum_{i=1}^{|L|} C_{w_i,y} \right)$$

Whereat S_y represents the percentage of frequency counts of the top 1% most frequent sentiment words of a review corpus in a year y. L states the lexicon and $C_{w_i,y}$ the count frequency of lexicon word w_i in the yearly review corpus, where the $C_{w_i,y}$ values are sorted in decreasing order. Concordantly, smaller values of S_y indicate that the yearly reviews are more diverse with respect to the lexicon L.

Qualitative analysis

For having a qualitative look at the data one can retrieve the most intense reviews, which are the shortest reviews that contain at least one intense sentiment term. For this purpose, intense reviews are defined as reviews that include at least one word with a sentiment score higher than 3 or lower than -3. Consequently, one can fetch the n most intense reviews by reducing them to those with at least one intense sentiment term, sorting the remaining reviews according to their length, and returning the n shortest ones.

Statistical significance analysis

For testing statistically whether observed upward or downward trends do likely exist, the Mann-Kendall statistical test (Kendall, 1948; Mann, 1945) can be used. This non-parametric test examines if a time series shows a consistent upward or downward trend, with a null hypothesis of no distinct trend. To be more accurate, the method ranks the data points, tests whether preceding values of a data point are always increasing or decreasing, calculates a variance-based test statistic on this information and compares it to a critical value that depends on the number of data points of the series. If the test statistic is significantly different from the critical value, a trend is identified. Since in this work views multiple time series, one should control - according to Ziser et al. (2023) - for a false discovery rate at the level $q = 0.05$ using the algorithm of Benjamini and Yekutieli (2005). This algorithm does not assume test independence and helps to prevent excessive false positive findings when analysing trends across multiple time series.

5. Results

In the subsequent section the analysis results will be displayed, again hereby it is important to note that the analysis itself focuses on groups of reviews for a certain year and sentiment in the stated data sets. Beginning with the sentiment metrics by having a glance at ***Figure 2*** showing the development of the average word-level sentiment metric, one can notice that the sentiment strength increased over the last decades. That means positive reviews got more positive and negative reviews more negative. This finding is in concordance with Ziser et al. (2023), more accurately, the time series of the IMDb data set looks nearly identical to the one presented by the authors. Furthermore, the series of the Yelp data set indicates that this trend also continues for the years 2021 and 2022. Contrary to Ziser et al. (2023), who just indicated a tipping point for the IMDb data in 2016 but not their used version of the Yelp data set, ***Figure 2*** also displays a tipping point for the latter in the year 2007. Nonetheless, it should be kept in mind that the years 2005 and 2006 of the Yelp data set seem to be outliers as earlier stated, and thus this tipping point might be just a consequence of the erroneous data.

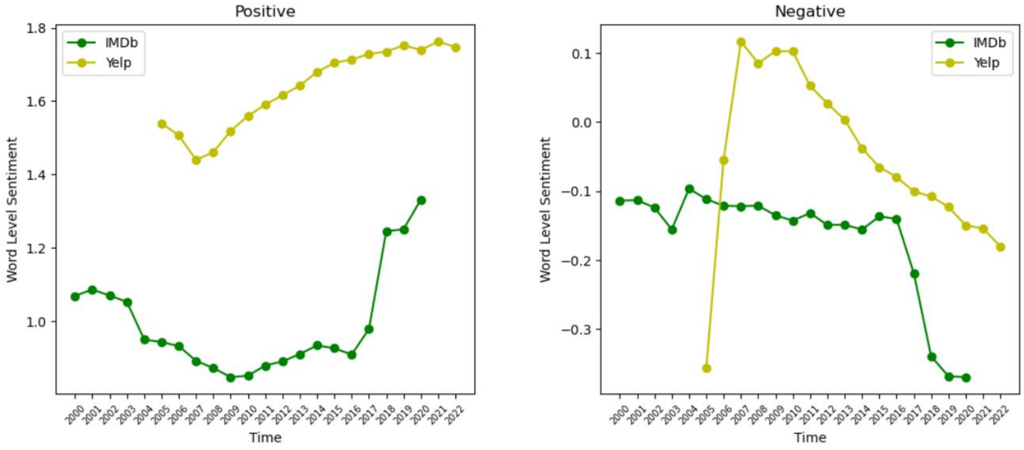


Figure 2: The average yearly word-level sentiment for the positive and negative reviews in IMDb and Yelp data set.

Moreover, the same timely strengthening trends of the sentiment intensity are also visible when viewing the changes in the average enhanced word-level sentiment over time for positive and negative reviews, which are shown in ***Figure 3***. However, the incorporation of lexicon-based enhancements leads to more extreme metric scores compared to the simpler sentiment metric, as also concluded by Ziser et al. (2023). At this juncture, given the workings of the rule-based enhancements on the final sentiment score, someone could conclude that the studied reviews probably do not contain many negations since that would have reduced the sentiment strength and more likely include a lot of exclamation marks or adverbs.

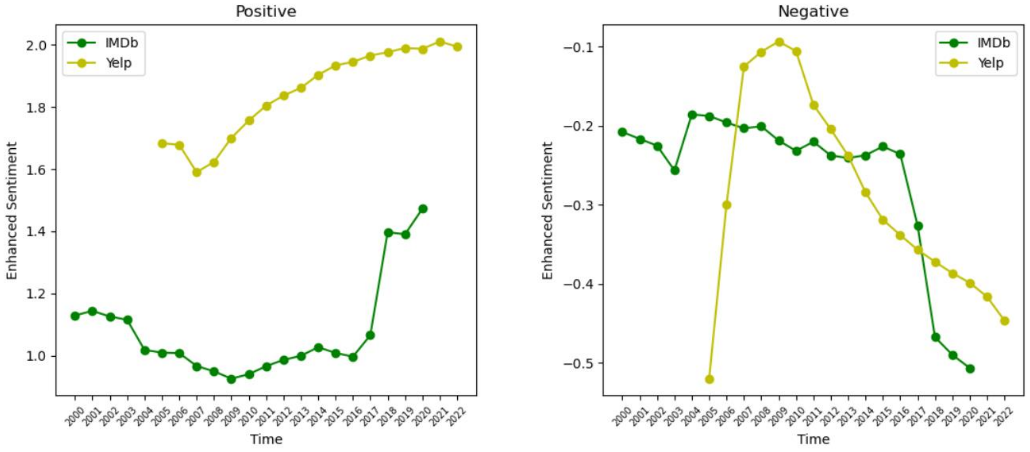


Figure 3: The average yearly enhanced sentiment for the positive and negative reviews in IMDb and Yelp data set.

Considering the last remaining sentiment metric in ***Figure 4***, the yearly average absolute sentiment, it is viewable that its development is not completely identical to Ziser et al.'s (2023) results. The authors found that the absolute sentiment intensity increased over the last decade although ***Figure 4*** indicates that this is not the case for negative Yelp reviews. Nevertheless, the authors also state that trends in their version of the Yelp data are generally smoother than those of the IMDb, thus one can say in line with Ziser et al. (2023) that people started to use stronger sentiment words for reviews over the last decades. When additionally just concentrating on the more recent years in the Yelp series, it shows inconsistent results if these trends do continue.

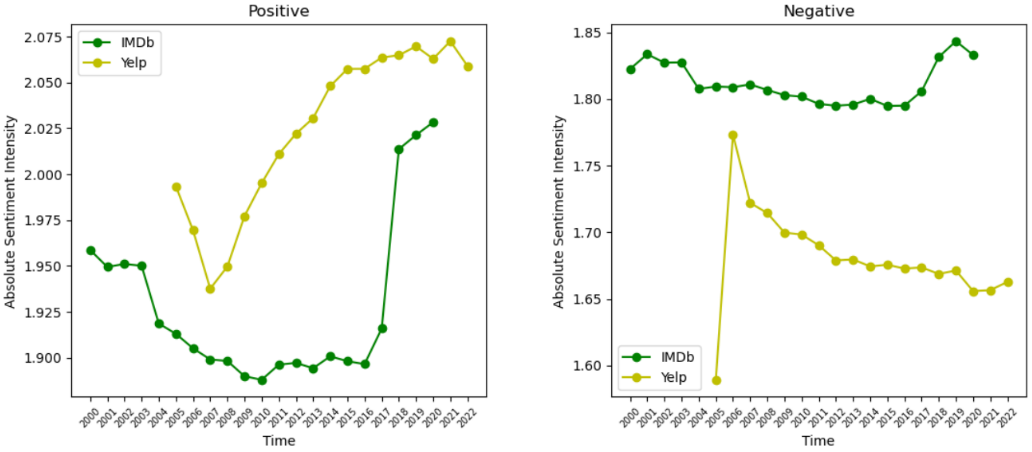


Figure 4: The average yearly absolute sentiment for the positive and negative reviews in IMDb and Yelp data set.

Besides the sentiment metrics, also those describing the language comprehensiveness and richness should be viewed in the course of this analysis. Thus, for evaluating trends in language comprehensiveness, ***Figure 5*** displays the average yearly review length and percentage of dichotomous reviews. By taking a glance, one can remark that the average review length for positive and negative reviews decreased after the years 2009 and 2010. Further, having a look at the newer Yelp data shows that the trend might have reached its valley. In addition, it can be seen that the percentage of one-sided reviews increased steadily over the last decades regardless of the sentiment. More precisely, the IMDb data set shows a sharp increase in the percentage of one-sided reviews after the year 2016 and the more recent years of the Yelp data suggest a continuation of this trend. Taken together this implies that reviews have become less comprehensive over the last decades as they got shorter and concentrate either on positive or negative aspects, which again is in line with Ziser et al.'s (2023) findings. Nonetheless, the possible continuation of this trend is a bit vague.

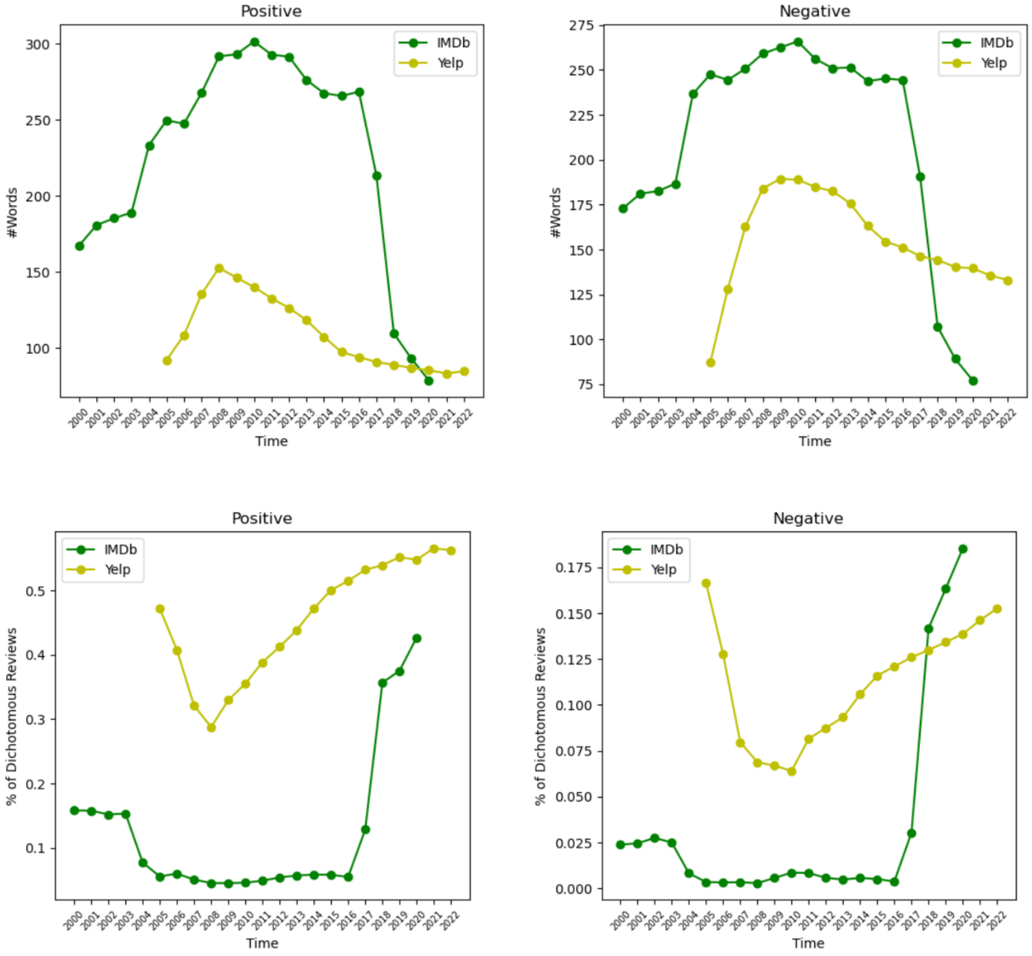


Figure 5: The average yearly review length and percentage of one-sided reviews for the positive and negative reviews in IMDb and Yelp data set.

Additionally, when being interested in the language richness one should gaze at ***Figure 6***, which gives the percentage of the frequency count of the most frequent sentiment words per year for positive and negative reviews. Here it can be observed that the reliance on frequent sentiment words instead of on more rare ones increased over the years, although like earlier, the more recent Yelp data might hint that it has reached a plateau. Consistent with Ziser et al. (2023) these circumstances allude to the fact that the language of reviews has become less diverse over time, however, this trend does not necessarily continue.

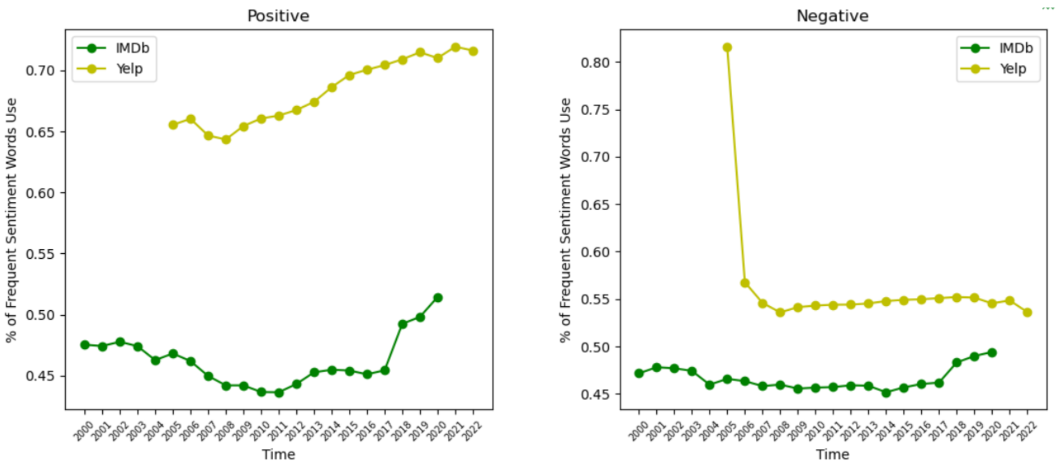
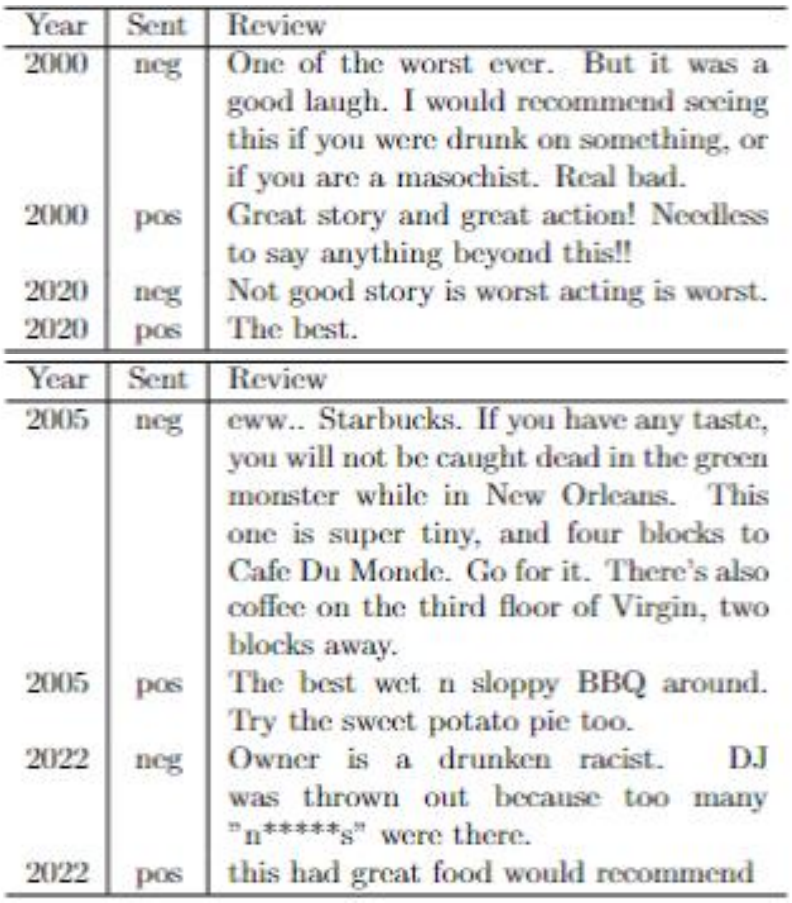


Figure 6: The percentage of the frequency count of the top 1% most frequent sentiment words per year for the positive and negative reviews in IMDb and Yelp data set.

For getting an impression of the observed trends instead of just purely viewing the statistics, ***Table 4*** compromises examples of intense positive and negative reviews for the earliest and latest years of the IMDb and Yelp data set. When having a glance at it one can see that the reviews of the recent years are on average shorter, incorporate less nuanced statements, and do not provide as much additional information as their older counterparts. On top of that, although they are shorter they often include intense sentiment words like “good” and “worst”.



Year	Sent	Review
2000	neg	One of the worst ever. But it was a good laugh. I would recommend seeing this if you were drunk on something, or if you are a masochist. Real bad.
2000	pos	Great story and great action! Needless to say anything beyond this!!
2020	neg	Not good story is worst acting is worst.
2020	pos	The best.

Year	Sent	Review
2005	neg	eww.. Starbucks. If you have any taste, you will not be caught dead in the green monster while in New Orleans. This one is super tiny, and four blocks to Cafe Du Monde. Go for it. There's also coffee on the third floor of Virgin, two blocks away.
2005	pos	The best wet n sloppy BBQ around. Try the sweet potato pie too.
2022	neg	Owner is a drunken racist. DJ was thrown out because too many "n*****s" were there.
2022	pos	this had great food would recommend

Table 4: Samples of some of the shortest reviews containing at least one intense term of the (a) IMDb and (b) Yelp data set.

Lastly, to check if the observed trends represent indeed an overall increase or decrease over time Mann-Kendall statistical tests can be applied to the time series, while controlling for a false discovery rate of $q = 0.05$. The tests point out that the null hypothesis of no trend cannot be rejected for most of the IMDb time series. But when focusing only on the last 10 years of the data, all null hypotheses except those for the absolute sentiment and dichotomy of the negative reviews can be rejected. Again, that is similar to what Ziser et al. (2023) found. Furthermore, when doing the same kind of test for the Yelp time series the null hypothesis for all tests except for those of the average review length of negative reviews can be rejected. Even when concentrating on the last 10 years of the data that does not change, which strongly suggests an interpretation in favour of trend continuations. This should be especially the case, as the data reduction excludes the likely outliers of 2005 and 2006 and gives the more recent data a stronger influence on the test statistic. Yet it should be noted that the empirical p-values for a lot of the series increased with the data reduction.

6. Discussion

In light of what was learned due to the proceeding of the data analysis, this final section concentrates on answering the research questions. Approaching the first research question about the reproducibility of Ziser et al.'s (2023) findings when it comes to the language trends in online reviews, one can say that they are indeed quite well reproducible. That means when relying on the same IMDb data set as the authors, the trends of an increase in the sentiment intensity, a reduction in the lexical diversity, and a decrease in the general comprehensiveness do show. Furthermore, these trends are not just visually visible but to a high degree statistically significant when carrying out statistical tests for the last 10 years covered by the data set. Here the trends for the absolute sentiment and dichotomy of the negative reviews build a non-significant exception.

Although it is nice to know that Ziser et al.'s (2023) findings are reproducible, it would be also nice to know if they are replicable when relying on a different data basis and if the found trends are actually of a lasting nature, therefore the second research question should be answered too. When it comes to the trend of an increasing sentiment intensity, it can be detected in most time series of the Yelp data set, where the absolute sentiment for negative reviews works as an exception. Nevertheless, this finding is not completely counterintuitive, as this time series is comparingly smooth and does not indicate an opposing trend. That all implies that on a general basis, the intensifying trend can be found in the Yelp data. Furthermore, the analysis of the Yelp data also indicates a decrease in the average review length and an increase in the percentage of one-sided reviews. To put it in a nutshell, the analysis also speaks in favour of the trend that reviews have become less comprehensive over the last decades. Besides that, the Yelp data also displays a stronger usage of frequent sentiment words over the last decades, which means the trend of a decreasing language richness can also be confirmed.

Nonetheless, while all these trends are present in the Yelp data their continuation is less sure. To be more precise, the development of the Yelp time series in the years 2021 and 2022 either indicates a trend continuation for just positive or negative reviews or suggests, that a trend starts to stagnate at its current level. Nevertheless, the null hypothesis of the Mann-Kendall statistical tests can be rejected for all time series except for the one displaying the average review length of negative reviews. Further, this fact does not change when reducing the time series to the 10 most recent years included in the data. Thus, although the empirical p-values for most time series increase this can be interpreted as evidence for a future trend continuation.

Summarizing the answers to the research questions it can be said, that Ziser et al.'s (2023) results about language trends in online reviews are indeed reproducible, they show also in more recent data, and have a continuing character. Namely, the sentiment intensity of reviews increased over the last years, and the lexical diversity as well as the general comprehensiveness decreased.

Albeit this work comes to similar results as Ziser et al. (2023) it underlies some limitations that should be kept in mind when regarding its results. For instance, the underlying reproduction was only conducted using one of three data samples from Ziser et al. (2023) and therefore cannot make assumptions about the other samples. Even so, it seems highly unlikely that by chance the only data basis that delivers reproducible results or on which behalf the authors do not make any errors in the analysis was selected in the course of this work. In addition, it might be possible that the found trends are strongly bound to the origin of the data, which means they just show in data related to Amazon, Yelp, or IMDb. If that were true, the selection of a newer version of the Yelp data set to test the replicability of Ziser et al.'s (2023) work would have just enforced the false results. Still, there are no valid reasons why someone would assume that might be the case.

Likewise, as this work underlies some limitations the original work of Ziser et al. (2023) can be also criticised in some of its aspects. Especially the misleading explanation of the retrieval of the enhanced sentiment made the reproduction and replication objective difficult, more precisely, without viewing the author's coding files the implementation of the enhanced sentiment metric would have been done wrongly. Besides this critique point, it can be stated that Ziser et al. (2023) mostly relied in their sentiment metric calculations on sentence parts that were generated by splitting a review at the beginning, end, and after white spaces. Albeit the implementation of further preprocessing steps like stemming and lemmatisation would have been favourable to match better the Vader lexicon more precisely or to access more fine-grained sentiment information.

7. Bibliography

Awajan, I., Mohamad, M., & Al-Quran, A. (2021). Sentiment analysis technique and neutrosophic set theory for mining and ranking big data from online reviews. *IEEE Access, 9, 47338*-47353.

Benjamini, Y., & Yekutieli, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association, 100*(469), 71-81.

Bian, Y., Ye, R., Zhang, J., & Yan, X. (2022). Customer preference identification from hotel online reviews: A neural network based fine-grained sentiment analysis. *Computers & Industrial Engineering, 172*, 108648.

Bokányi, E., Kondor, D., Dobos, L., Sebők, T., Stéger, J., Csabai, I., & Vattay, G. (2016). Race, religion and the city: twitter word frequency patterns reveal dominant demographic dimensions in the United States. *Palgrave Communications, 2*(1), 1-9.

Crystal, D. (2001). Language and the Internet. *Cambridge*, CUP.

Crystal, D. (2007). How language works: How babies babble, words change meaning, and languages live or die. *Penguin.*

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.*

D'Acunto, D., Tuan, A., & Dalli, D. (2020). Are online reviews helpful for consumers?: Big data evidence from services industry. In *Exploring the power of electronic word-of-mouth in the services industry* (pp. 198-216). IGI Global.

Biswas, E. (2021). <i>IMDb Review Dataset - ebD</i> [Data set]. *Kaggle.* https://doi.org/10.34740/KAGGLE/DSV/1836923

Eisenstein, J. (2013, June). What to do about bad language on the internet. In *Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies* (pp. 359-369).

Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).

Irfan, J. J. D. H. (2021). Semantic Change in English Language: Social Media Neologisms.

Kendall, M. G. (1948). Rank correlation methods.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica: Journal of the econometric society,* 245-259.

Santana, A. D. (2014). Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism practice, 8*(1), 18-33.

Schindler, R. M., & Bickart, B. (2012). Perceived helpfulness of online consumer reviews: The role of message content and style. *Journal of Consumer Behaviour, 11*(3), 234-243.

Tripathi, S., Deokar, A. V., & Ajjan, H. (2021). Understanding the order effect of online reviews: a text mining perspective. *Information Systems Frontiers,* 1-18.

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology, 39*(4), 202.

Zhang, Z., Guo, J., Zhang, H., Zhou, L., & Wang, M. (2022). Product selection based on sentiment analysis of online reviews: An intuitionistic fuzzy TODIM method*. Complex & Intelligent Systems, 8*(4), 3349-3362.

Zipf, G. K. (1942). The unity of nature, least-action, and natural social science. *Sociometry, 5*(1), 48-62.

Ziser, Y., Webber, B., & Cohen, S. B. (2023). Rant or rave: variation over time in the language of online reviews. *Language Resources and Evaluation*, 1-31.