Language Resources and Evaluation
https://doi.org/10.1007/s10579-023-09652-5

ORIGINAL PAPER

Rant or rave: variation over time in the language of online reviews

Yftah Ziser[1] · **Bonnie Webber**[1] · **Shay B. Cohen**[1]

Accepted: 27 February 2023

Abstract

We examine how the language of online reviews has changed over the past 20 years. The corpora we use for this analysis consist of online reviews, each of which is paired with a numerical rating. This allows us to control for the perceived sentiment of a review when examining its linguistic features. Our findings show that reviews have become less comprehensive, and more polarized and intense. We further analyzed two subgroups to understand these trends: (1) reviews labeled "helpful" and (2) reviews posted by persistent users. These trends also exist for helpful reviews (albeit in a weaker form), suggesting that the nature of reviews perceived as helpful is also changing. A similar pattern can be observed in reviews by persistent users, suggesting that these trends are not simply associated with new users but represent changes in overall user behavior. Additional analysis of Booking.com reviews indicates that these trends may reflect the increasing use of mobile devices, whose interface encourages briefer reviews. Lastly, we discuss the implications for readers, writers, and online reviewing platforms.

Keywords Online reviews · Language variation · Language sentiment · Language polarization · Corpus analysis

✉ Yftah Ziser
yftah.ziser@ed.ac.uk

Bonnie Webber
bonnie.webber@ed.ac.uk

Shay B. Cohen
scohen@inf.ed.ac.uk

[1] School of Informatics, University of Edinburgh, 10 Crichton St., Edinburgh EH89AB, Scotland, UK

Published online: 31 March 2023

1 Introduction

Online communities are virtual spaces where one can participate in lively discussions, read and write reviews, ask questions in multiple domains and meet people with similar interests, among other uses. In the last few decades, the use of online communities has grown rapidly, and they are now an integral part of our lives. According to Global Web Index,[1] 76% of internet users participated in an online community in 2019, and 64% of visitors to online community sites said they were visiting community platforms more often than they did a few years ago.

The growing role that online communities play in our lives has led to a corresponding academic interest in understanding online user behavior in such communities. User behavior research covers many aspects, including user motivation to participate in online communities (Lampe et al., 2010), user loyalty towards a community (Hamilton et al., 2017) and user reaction to community norms at different stages of their membership (Danescu-Niculescu-Mizil et al., 2013).

In this work, we aim to identify developing trends in online users' language to express their opinion. Where changes in the language may or may not imply changes in the strength of people's opinions or their commitment to those opinions, it may simply be a matter of how expression has evolved. For this purpose, we have focused on online reviews for four main reasons: (1) Reviews inform us in our day-to-day consumption of goods and use of services, with their role growing as e-commerce becomes more popular and more products and services become subject to review; (2) People have opinions. Reviews provide an opportunity to express them. Hence, reviews are a natural choice for tracking changes in how people express themselves; (3) Some long-standing review platforms have reviews that go back decades, allowing us to have a broad look over observable phenomena; (4) Several online review platforms, e.g., Amazon, ask their users to provide numerical ratings alongside their textual reviews to share their impressions as they perceive it, allowing us to collect data that couples language with sentiments.

Even if language changes over time, review numerical ratings can serve as stable anchors to examine how people express their opinion when they are fully content (e.g., 5-star review on Amazon) or utterly disappointed (e.g., 1-star review on Amazon) throughout the years. These anchors may eliminate some of the noise coming from changes in people's opinions over time. We want to investigate how people's expression evolved, not if a sentiment shift for a specific product or service has occurred. Sentiments may shift in a time span of years, months, and even days (Alattar & Shaalan, 2021; Jiang et al., 2011; Tan et al., 2013), thus contaminating our measurements. For example, using the most positive reviews helps us reduce the noise coming from negative sentiment shifts as we examine the language the users use when utterly pleased, even if the portion of positive reviews for a specific product or service is lower in a given year. In addition, Schoenmueller et al. (2020) recently showed that review scores tend to be polarized in most reviewing

[1] https://www.gwi.com/reports/online-communities-reddit.

platforms,[2] making the most positive and negative reviews highly representative of the overall review population. Throughout our research, the following questions emerged:

- **RQ1** How has the way users express themselves in online reviews evolved over the years?
- **RQ2** Do new users solely drive the changes? Or do existing users change their habits as well?
- **RQ3** How are helpful reviews, which get high exposure, affected by these changes?
- **RQ4** What could be causing such changes?

We emphasize that in our research questions, we aim to analyze shifts over time in *language expression* for a fixed sentiment level rather than shifts in the *sentiment* itself (as indicated, for example, by the number of stars selected in a review). We mitigate the effect of a possible sentiment shift on our analysis (which arguably, could exist) because we focus on a subset of the data with extremely positive or negative reviews.

We analyze data from three review sources that differ in their domain and their review platform. Using data from popular long-standing review platforms enables us to collect a sufficient number of reviews for earlier years. Each dataset covers 15 or more years of reviews, allowing us to observe trends over time. In general, we have observed similar trends in all three datasets: (1) Users have come to use stronger words to convey both negative and positive sentiments towards the reviewed product or service; (2) The diversity in the language used in reviews has decreased over the years; (3) The reviews have become briefer and more one-sided, hence less comprehensive.

To shed light on the trends mentioned above, we compare all users to a segment of users who submit reviews regularly to understand if the overall trends stem solely from new users. Our analysis shows that while the trends are weaker for regular users, they still exist, demonstrating that such users have been actively changing their habits over the years. We elaborate on this analysis in Sect. 5.

All three review platforms allow users to mark a review as *helpful*, positively reflecting on the review's usefulness to future readers and its capability to aid them in making more informed decisions. Helpful reviews are of high importance, as they normally appear first in the list of reviews of a product or a service, and in many cases, due to the hosting platform interface organization, they are the only ones users encounter when accessing reviews. Again, we find that helpful reviews show similar trends. Although these trends tend to be weaker, their existence means that the reviews that a community perceives as helpful are also changing over the years. We elaborate on those analyses in Sect. 6.

[2] Schoenmueller et al. (2020) measured how polarized the reviews are by dividing the number of extreme reviews (reviews with the maximal or minimal score available in the platform) by the number of total reviews. They also found that most datasets are highly skewed toward positive reviews.

In our conclusion, we describe an analysis of reviews on Booking.com which revealed a discernible difference in language quality and sentiment expression between reviews posted from personal computers and those posted from mobile devices. Reviews originating from mobile devices exhibited lower language quality and greater use of strong sentiment words. Given the proliferation of mobile device usage, this finding may potentially account for the trends we observed in decreased language diversity and heightened intensity of opinions.

2 Related work

We begin with a high-level discussion of the language in the Internet and its relation to an extreme form of expression. We continue by reviewing the use of extreme language in online platforms. While we do not explicitly examine this angle in our research, we discuss it to distinguish this line of work from ours. Both are closely related as people often use profanities or racial and sexist slurs to convey strong sentiments. We then describe academic capitalism, as it is related to our work, focusing on the increasing sentiment intensity over time researchers use in their academic papers in presenting their results. We continue by discussing language variation in social media. Finally, we discuss the phenomenon of fake reviews, as we initially considered it to be a driving force behind the trends we observe. In Sect. 3, we see that this phenomenon is not likely to significantly affect our results and conclusions.

2.1 Language use on the internet

In his seminal book, Crystal (2001) stated, “if the Internet is a revolution, therefore, it is likely to be a linguistic revolution.” The emergence of Internet communities, where individuals are free to express their opinions with minimal moderation or editing, has been a subject of interest in studies of cultural and linguistic developments since the early days of the Internet (Wilson & Peterson, 2002). The democratization of publishing content on the internet (for example, through social media) has brought about new forms of communication, creating unique language varieties. In general, the language of the Internet is less grammatical (Eisenstein, 2013), and noisier than traditional text language (Baldwin et al., 2013), posing new challenges for NLP researchers. Previous studies have addressed these challenges by building customized models (DeLucia et al., 2022; Qudar & Mago, 2020; Severyn & Moschitti, 2015) or adapting models that were trained using high-quality data to perform well on noisy user-generated data (Ben-David et al., 2022; Meftah et al., 2021; Schnabel & Schütze, 2014).

However, online and offline expression differences are broader and deeper than noisier grammar and slang vocabulary alone. Internet anonymity, the lack of physical presence, and the absence of social cues lead to more varied and sometimes more extreme expressions of personality and emotions online. This dissonance between online and offline behavior is often referred to as *the online disinhibition effect*. Suler (2004) distinguished between two opposite sides of the same coin: (a) *benign disinhibition*, where one acts more kindly and generously online and (b) *toxic disinhibition*, where people use rude language, write harsh criticisms, and generally, show more anger than usual. This unique relationship between the Internet and language use with the increasing amount of time Internet users spend on social media[3] and perceived increase in the polarization of public discourse on a variety of issues motivated us to ask: Can we characterize such trends through a test-case study, aiming to separate language **use** and language **intent**?

2.2 Extreme language in online discourse

Understanding and detecting extreme language on online platforms has recently become a research focus. Kenski et al. (2020) used two surveys to better understand general public perceptions of incivility. Their findings show that name-calling and vulgarity are perceived as more uncivil than other speech acts. Santana (2014) show that user anonymity plays a significant part in incivility, almost doubling the probability of an uncivil submission to a news site platform. Detecting uncivil and toxic behavior on online platforms is of great importance and hence is widely studied in the natural language processing (NLP) community. Hua et al. (2018) released a Wikipedia talk page corpus containing a complete moderation history of each page in the dataset. They show that the prevalence of personal attacks on moderators is more extensive than estimated in previous research, using only the final version of the talk page. Waseem and Hovy (2016) provide a dataset for detecting hate speech in the form of racist and sexist remarks. While extreme use of language through incivility, sexism, or racism, might correlate with conveying strong sentiment, we do not explicitly investigate this connection.[4]

2.3 Positivity bias and academic capitalism

It was observed and hypothesized early by Boucher and Osgood (1969) that people tend to use positive words more frequently than negative ones. Known as the *Pollyanna hypothesis*, more comprehensive research later reinforced this hypothesis using large-scale multilingual corpora, showing the words in natural languages possess a universal positivity bias (Dodds et al., 2015). Recently, Wen and Lei (2022) showed that positivity bias is common in academic papers, supporting the results of previous studies in this field (Holtz et al., 2017; Vinkers et al., 2015). Examining paper abstracts from the last five decades presents an upward trend towards more positive abstracts over the years. This phenomenon is also known as *academic capitalism*, and is believed to result from an increasing pressure to publish work ("publish or perish") and the competitiveness of academic environments. Our own research is connected to this work through the examination of changes in language use over time.

[3] https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/.

[4] The sentiment lexicon we use contains profanity and uncivil words. These words have highly negative scores.

2.4 Language variation in social media

Language in social media evolves at a rapid pace. As such, it is a fruitful testbed for studying language change and variation. Zanzotto and Pennacchiotti (2012) studied the emergence of new words (neologisms) in social media and whether crowdsourced dictionaries can help interpret such words. They proposed a setup where Twitter served as the social media platform, and Urban Dictionary,[5] a community dictionary that, through user contributions, presents slang words and phrases and their meaning in plain English. Their findings demonstrate that most novel words are introduced on Twitter before they are added to Urban Dictionary. They concluded that crowdsourced dictionaries alone would not provide a comprehensive solution to the problem of interpreting new words.

Similarly, Grieve et al. (2017) examined the properties of newly emerged words on Twitter, including their semantic classes, grammatical parts of speech, and word formation processes. In addition to time, language variation in social media has been studied across other dimensions, such as the author's location and demographics (Blodgett et al., 2016; Jurgens et al., 2017; Pavalanathan & Eisenstein, 2015). For example, Eisenstein et al. (2010) used cascaded topic models to show that Twitter users' language varies with their geographic location. Schwartz et al. (2013) presented Differential Language Analysis (DLA) to find words, phrases, and topics for accurately characterizing the demographic and psychological attributes of Facebook users. Eisenstein et al. (2014) examined how language change propagates on Twitter. Their findings show that linguistic changes are more likely to be transmitted between areas with similar demographics.

2.5 Review manipulation

Fake reviews (FR) (Lappas, 2012), where "bad faith" reviewers intentionally harm or boost a product or service reputation, is a well-researched phenomenon. Online review platforms invest significant resources to mitigate FR, as it harms users' trust and often results in poor market efficiency (Wu et al., 2020). For example, Yelp, a business review platform (see Sect. 3), presents users with *recommended reviews*, i.e., reviews that were classified by the Yelp algorithm to be trustworthy and *not recommended reviews*, for the rest. Yelp retroactively filters reviews using the latest version of its detection algorithm. Luca and Zervas (2016) inspect Yelp algorithm predictions for reviews of businesses that were caught in the act of soliciting FR. Their findings show that Yelp marked almost 80% of the FR as *not recommended reviews*, where the average rate for other businesses in this area was below 20%, thus reinforcing the algorithm's reliability.

In contrast to Yelp, a user desiring to post a review on Amazon's e-commerce platform needs to meet several payment conditions.[6] This is designed to discourage the posting of FR. Recently, He et al. (2021) studied the nature of FR on Amazon,

[5] https://www.urbandictionary.com/.

[6] See the Eligibility section in https://www.amazon.com/gp/help/customer/display.html?nodeId=GLHXEX85MENUE4XF.

showing that "Beauty & Personal Care," "Health & Household," and "Home & Kitchen" categories are more prone to FR than other categories. Our search for evidence of FR frequency or attempts to prevent them on the Internet Movie Database platform (IMDb; see Sect. 3) failed at finding any relevant research studies.[7]

3 Data and tools

This section discusses the rationale behind using online reviews to test language variation. Then, we describe how we mitigate the issue of fake reviews and to what extent it influences our conclusions. We continue by describing the sources, attributes and statistics of our datasets. Finally, we present the sentiment lexicon we use to score each review and the more advanced classifier that harnesses this lexicon to get more accurate sentiment scores.

3.1 Rationale behind review analysis

We have focused our study on reviews of products and services because, for years, consumers have been encouraged to express their views on the options that e-commerce platforms make available to them. Moreover, platforms such as Amazon ask users to give a numerical rating alongside their textual review to quantify their impression of products or services. Such numerical ratings allow us to control and quantify, to some extent, our understanding of the sentiment level conveyed by users towards a product or service. To put it simply, using online reviews allows us to control for sentiment when studying the overall effect of time on language use.

We examine user reviews that have been collected from three prominent online review platforms: Amazon (Books), the Internet Movie Database (IMDb)[8] and Yelp.[9] For Amazon Books, we use a dataset containing reviews from 2000 up to 2018 (Ni et al., 2019). We chose to focus on the books category for two main reasons: (1) Amazon started as an online bookstore, and thus we can obtain an ample amount of reviews from the early years when reviews were usually scarce; (2) In the study of fake reviews that (He et al., 2021) conducted, Amazon's book category is not among the categories they find most prone to fake reviews. For Yelp reviews, we use the Yelp open dataset[10] provided by Yelp for academic purposes, containing reviews from 2007 up to 2020. All the reviews in this dataset are labeled as *recommended reviews* by the Yelp algorithm, thus are more trustworthy (Luca &

[7] While IMDb did publish user review guidelines explicitly stating that reviews that include "advertising, promotions or solicitations of any kind" will be removed, some online discussions suggest that the FR phenomenon exists on IMDb.

[8] IMDb is an online movie, TV shows and other entertainment sources database, now operated by Amazon. The database contains over ten million entries and includes fan and critical reviews, among other user-generated content.

[9] Yelp is an American company that provides community-sourced reviews about businesses and is widely used by users who are interested in finding reliable information about people's experience with these businesses.

[10] https://www.yelp.com/dataset.

Zervas, 2016). For a more thorough discussion about Amazon and Yelp review credibility, see Sect. 2. For film reviews on IMDb, we use Kaggle's dataset,[11] containing reviews from 2000 up to 2020.

The above datasets are extensively used in the NLP and machine learning research communities and are considered the standard benchmark for many tasks. For example, these datasets are used for the research of recommendation systems (Wang et al., 2020; Zhang & Chen, 2020), sentiment analysis (Ben-David et al., 2020; Kumar et al., 2019), aspect-based sentiment analysis (Pontiki et al., 2014; Xu et al., 2019), and even to analyze domain shifts in neural networks (Zhao et al., 2022).[12] This paper aims to examine the change in the reviews of these datasets over time, an aspect that is often overlooked in previous works but may have implications for their use and reliability. In Table 1, we show statistics about the number of reviews in each dataset, both positive and negative, per year. We can see that positive reviews are more common than negative ones, supporting the findings of Schoenmueller et al. (2020), which explored the distribution of review ratings in multiple online platforms.

3.2 Sentiment lexicon

To obtain intensity scores of sentiment words, we use the Vader sentiment lexicon (Hutto & Gilbert, 2014), which contains scores between − 4 (highly negative) and 4 (highly positive) for more than 7500 sentiment words, focusing on microblog-like contexts. The final score for each sentiment word is the average score given by ten independent human raters. For examples, see Table 2a. To further validate our use of the Vader lexicon, we use the AFFIN sentiment lexicon (Nielsen, 2011), which contains scores between − 5 (highly negative) and 5 (highly positive) for more than 3300 sentiment words. The lexicon was manually curated between 2009 and 2011. The AFFIN and Vader intersection contains 2647 terms, almost 80% of the AFINN terms, focusing on microblog-like contexts. The Pearson correlation between the sentiment scores given by the lexicons mentioned above is over 0.92, indicating strong agreement. Moreover, conducting the experiments in Sects. 4, 5, 6, and 7 with the AFFIN lexicon yield nearly identical results to the ones we obtain using the Vader lexicon. We hence report the Vader lexicon results, as it is a more comprehensive lexicon. It is noteworthy that while the lexicons were curated several years apart, which can be a significant time with respect to microblog-like contexts, there is no detectable sentiment shift between the lexicons.

3.3 Lexicon-based enhancements

Using word-level sentiment methods may be sub-optimal for scoring real-world reviews. For example, consider the review statement *The stuffed peppers weren't*

[11] https://www.kaggle.com/ebiswas/imdb-review-dataset.

[12] In some previous work, an older version of these datasets is used, as they are consistently updated each year.

very good—a naïve use of a sentiment lexicon might classify this review as positive, and ignore the negation, rather focusing on the word “good.” To mitigate this problem, Hutto and Gilbert (2014) offered heuristic enhancements to the vanilla lexicon-based approach. These heuristics take into account negation, punctuation, and degree modifiers to alter sentiment intensity and enable a scoring mechanism that contributes to the sentiment intensity score. The scoring mechanism is based on applying rule-based modifications to existing tweets (for example, adding an exclamation mark to the tweet *great news*) and then using human evaluations to assess their impact on the perceived sentiment of the tweets after the heuristics are applied to these tweets. For some examples of the modifications that are applied to the tweets (with the base word *good*), see Table 2b.

3.4 Lexicons as sentiment classifiers

To examine how well lexicon-based approaches align with the overall review sentiment, we experiment with binary sentiment analysis, where the positive samples are the reviews with the highest possible scores, e.g., five stars for Amazon reviews, and the negative samples are the reviews with the minimal score. In addition to lexicon-based methods, we test BERT (Devlin et al., 2019), which can be used as a state-of-the-art text classifier. Unlike lexicon-based methods, BERT is a supervised learning method that needs annotated examples to perform sentiment analysis. To this end, we sampled 1000 examples for training and 4000 for testing. Both have an equal amount of positive and negative samples and consist of an equal amount of samples from each year available for the dataset. Table 3 presents the results. We can see a clear pattern in which the enhanced lexicon is superior to the basic lexicon method. This pattern is expected because the enhanced lexicon uses context-aware features. BERT outperforms both lexicon-based methods, a finding previously observed by Alaparthi and Mishra (2021). The reader might ask: *If supervised methods are significantly better than lexicon-based methods for predicting the sentiment of reviews, why not use them for our analyses?* We use lexicon-based methods instead of supervised methods for several reasons: (1) While consistently providing state-of-the-art results across many tasks, supervised methods, especially recent ones, are often challenging to interpret, rendering them difficult to use in the context of our research questions; (2) Supervised machine learning methods use labeled data to classify and predict. However, these methods are sensitive to changes in the data distribution, such as shifts in the domain or temporal changes, as shown in recent research (Alkhalifa et al., 2022; AL-Sharuee et al., 2021; Bjerva et al., 2019, inter alia). This sensitivity may potentially affect the accuracy of our analysis, particularly since our questions focus on changes manifested in the data over time. To avoid any potential biases introduced by distribution shifts, we choose to use alternative methods, such as lexicon-based approaches, which do not rely on labeled data. This approach helps ensure that our results are reliable and not influenced by external factors; (3) We analyze shifts in language use over time rather than changes in sentiment. To achieve this goal, we have implemented measures to control for

Table 1 The number of negative (one star) and positive (five stars for Amazon and Yelp, ten stars for IMDb) reviews for each dataset by year

Year	Amazon		Yelp		IMDb	
	★ × 1	★ × 5	★ × 1	★ × 5	★ × 1	★ × 10
2000	18,287	100,000	–	–	5252	20,734
2001	17,220	100,000	–	–	7071	26,649
2002	17,658	100,000	–	–	8769	31,067
2003	19,689	100,000	–	–	8785	29,519
2004	26,842	100,000	–	–	8635	28,578
2005	41,262	100,000	–	–	16,906	50,591
2006	40,854	100,000	–	–	21,533	64,874
2007	43,147	100,000	4365	21,534	16,336	41,210
2008	48,953	100,000	9565	43,296	14,901	31,690
2009	61,494	100,000	17,031	65,147	12,652	26,051
2010	70,759	100,000	27,426	100,000	11,954	24,511
2011	90,887	100,000	44,345	100,000	10,757	23,041
2012	100,000	100,000	55,466	100,000	13,157	28,080
2013	100,000	100,000	71,399	100,000	14,164	29,608
2014	100,000	100,000	99,452	100,000	16,059	34,138
2015	100,000	100,000	100,000	100,000	16,606	34,493
2016	100,000	100,000	100,000	100,000	15,053	33,428
2017	100,000	100,000	100,000	100,000	24,727	44,509
2018	100,000	100,000	100,000	100,000	51,259	100,000
2019	–	–	100,000	100,000	83,048	100,000
2020	–	–	100,000	100,000	100,000	100,000

For computational reasons, we limit the number of reviews we use to 100,000 per year

sentiment in our analysis. We do not use classifiers like BERT as they provide sentence-level scores, and require further supervision to obtain word-level scores. We use dictionary-based methods that offer such word-level score annotations. We can then better capture and analyze changes in the use of specific words or phrases over time, providing a more granular and accurate understanding of shifts in language use. For example, consider the review pair *This book is nice, the plot is fine, and the pace is decent* and *This book is fantastic, the plot is incredible, but the pace is horrible*. At the review level, BERT would likely classify the first review as more positive than the latter, masking the use of stronger terms (positive or negative) in the second review.

4 Main analysis

We have analyzed the data with respect to **sentiment**, **language richness** and **comprehensiveness**. Each analysis is conducted separately for positive and negative reviews.

Table 2 The Vader lexicon contains words, emojis, and Internet slang (e.g., 182 means *I hate you*)

(a)

Term	Score	Term	Score
euphoria	3.3	mediocrity	−0.3
great	3.1	apathetic	−1.2
10q	2.1	annoying	−1.7
:)	2.0	)-':	−2.1
good	1.9	bad	−2.5
acceptable	1.3	182	−2.9
OK	1.2	worst	−3.1
compelling	0.9	hell	−3.6

(b)

Term	Score
good!!!	2.776
very good	2.193
so good	2.193
good!	2.192
good	1.9
kind-of good	1.607
not good	−1.406
wasn't very good	−1.622

(a) Terms and their corresponding Vader sentiment scores; (b) Enhancements effects on the sentiment word "good"

4.1 Sentiment

We have used three variants of sentiment analysis metrics to better understand changes in sentiment intensity over time. We start with the word-level sentiment, a weighted average of the sentiment scores for each sentiment term in the review. We then continue with similar metrics, replacing the sentiment scores with more context-aware ones. Our final sentiment metric uses an absolute sentiment score.

4.2 Word-level sentiment

Word-level sentiment (denoted below by S_r for review r) is calculated for a given review as follows:

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

where

- S_r: is the sentiment score for review r,

Table 3 Performance of different sentiment classifiers for each of the datasets we study (measured as accuracy)

Model	Amazon	Yelp	IMDb
Lexicon	0.678	0.756	0.742
Enhanced lexicon	0.709	0.795	0.759
BERT	0.928	0.955	0.936

- L: is the sentiment lexicon,
- C_{rw_i}: is the count of lexicon word w_i in review r,
- L_{w_i}: is the intensity score given to the word w_i by lexicon L.

--> basically loop through lexicon

In the above, the index i ranges over the words in review r that appear in L. The notation $|L|$ refers to the size of L as a set of lexicon words. As such, Eq. 1 is the weighted (by the count of the word in the review) average of the lexicon sentiment score of words that occur in the review.

4.3 Enhanced word-level sentiment

The enhanced word-level sentiment is calculated for a given review by replacing L_{w_i} in Eq. 1 with $\overline{L}_{w_i}$, a score that reflects *lexicon-based enhancements*, such as negations, punctuation, and degree modifiers (cf. Sect. 3).

4.4 Absolute sentiment

For absolute sentiment intensity, we use a variant of Eq. 1, in which L_{w_i} is replaced with its absolute value, to measure intensity independent of sentiment class (positive or negative).

Figure 1 presents the analyses mentioned above. Both enhanced and word-level sentiment analyses show a clear sentiment trend for all our datasets, in which the positive reviews are getting more positive over the years. Likewise, the negative ones are getting more negative. The absolute sentiment analysis suggests that people have used stronger sentiment words for positive and negative reviews in recent years. While Yelp trends are smooth, for Amazon and IMDb, we can observe a tipping point, i.e., a year in which the trends intensify. In absolute terms, in both enhanced and word-level sentiment analyses, negative reviews yield lower values than positive ones, e.g., while word-level sentiment values range from 0.8 to 2 for positive reviews, negative reviews range from − 0.5 to 0.2. The difference in absolute terms shows that the *Pollyanna hypothesis*, i.e., that people tend to use positive words more frequently than negative ones (Boucher & Osgood, 1969), holds in the domain of the reviews as well (for more details, see Sect. 2).[13]

[13] A more comprehensive analysis of the *Pollyanna hypothesis* across a variety of review datasets can be found in Aithal and Tan (2021).

4.5 Language diversity and comprehensiveness

To better understand trends in language diversity, we examine how often people use frequent sentiment words. The rationale is that using only a handful of sentiment words will lead to less diverse opinionated texts. To better understand text comprehensiveness, we examine text length and how one-sided the opinions in it are. Schindler and Bickart (2012) showed that longer and less one-sided opinionated text is often painting a more complete picture of the reviewed object. In the next few paragraphs, we specify how we measure the metrics mentioned above.

4.6 Use of frequent sentiment words

Several measurements have been proposed to assess lexical diversity of language, including Type-Token Ratio (TTR; the ratio between the number of unique words and the overall number of words in a corpus - the closer to 1, the greater the complexity), vocd-D (McKee et al., 2000), and the measure of textual lexical diversity (MTLD; McCarthy 2005). These measures have a shortcoming with short texts: they have a large measurement variance when applied to them. We refer the interested readers to Koizumi (2012), who conclude that such metrics are not reliable for short texts, such as reviews. Hence, we do not use the above methods to measure diversity and instead offer a different method.

The alternative metric we use to assess lexical diversity relies on the observation that such diversity is indicated in the shape of the Zipfian distribution of the underlying corpus. Zipf (1942) showed that word distributions have a narrow "head"[14] and a long "tail." The narrower the head is, the more it indicates the use of a small number of words very frequently, and as a consequence, high rate of repetition of words from the high-frequency part of the word distribution.

This leads us to assess diversity by measuring how "wide" the head of the Zipfian distribution is for a given corpus. More precisely, we calculate for each corpus and each year, the percentage of the most frequent sentiment word types (top 1%) with respect to all the sentiment words that appear in the corpus as follows:

$$S_y = \left(\sum_{i=1}^{\left\lceil \frac{|L|}{100} \right\rceil} C_{w_i,y} \right) \Bigg/ \left(\sum_{i=1}^{|L|} C_{w_i,y} \right), \tag{2}$$

where

- S_y: is the percentage of the most frequent sentiment word types (top 1%) with respect to all the sentiment words that appear in reviews written in year y,
- $C_{w_i,y}$: is the count of lexicon word w_i in a corpus composed of reviews written in year y. They are sorted by their frequency in decreasing order such that $C_{w_1,y}$ is

[14] For example, only 135 terms are needed to account for half the Brown Corpus (Fagan & Gençay, 2010).

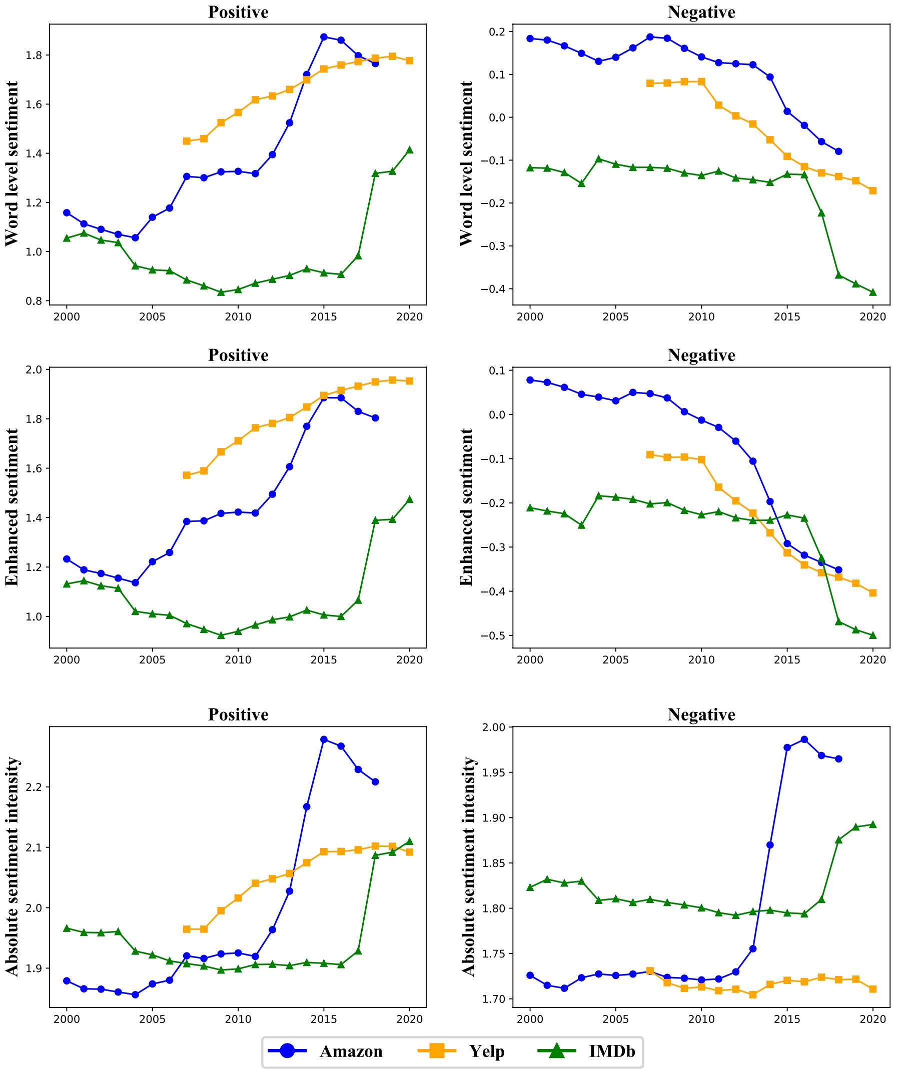

Fig. 1 The sentiment measurement variants. The figures on the left correspond to the positive reviews, while the figures on the right correspond to the negative reviews. Note the *y*-axes in the left plots and right plots are of a different scale

the count of the most frequent word in year y, and $C_{w_2,y}$ is the count of the second most frequent word in that year.

- L: is the sentiment lexicon, and $|L|$ stands for its size. Note that in the numerator, the index i ranges from 1 to $\frac{|L|}{100}$, iterating the top 1% most frequent sentiment words for the year y.

In the above, the numerator is the sum of the top used (top 1%) sentiment word frequencies for a given year. The denominator is the sum of **all** sentiment words frequencies for a given year. A large value of S_y indicates that the reviews that year do not make a diverse use of the sentiment vocabulary.

4.7 Review length

Review length is strongly connected to its informativeness, as a lengthy review is expected to include more details and information about the product or service in hand (Schindler & Bickart, 2012). In addition, review length has shown to be indicative of readership, helpfulness, and product sales (Aggarwal & Aakash, 2020; Salehan & Kim, 2016), which are of high interest to the review readers, writers and the sellers of the reviewed object. To measure the review length. We calculate the average number of words per review.

4.8 Dichotomy in reviews

We examine the percentage of one-sided reviews, i.e., reviews having at least one sentiment term corresponding to their numerical sentiment rating and none of the opposite sentiment. To do so, we use the enhanced word-level sentiment. For example, the review *The actors were good, and the plot was not bad as well,* from IMDb positive reviews is considered one-sided as negative sentiment is not conveyed throughout the entire review. As shown in the example, one-sided reviews do not have to convey extreme sentiment. In terms of trend-tipping points, we observe similar patterns for sentiment, language richness, and comprehensiveness, which demonstrate that Yelp trends are smooth, and Amazon and IMDb trends have clear tipping points, i.e., years in which the trend intensifies.

Figure 2 presents the analyses mentioned above. We observe a sharp decline in review length, for both positive and negative reviews, mainly in Amazon and IMDb.[15] The percentage of one-sided reviews, either positive or negative, suggests that reviews have become more one-sided over time. In absolute terms, negative reviews tend to be less one-sided than positive ones, which could stem from a positivity bias. The combination of shorter and more one-sided reviews leads to less comprehensive reviews, as the reviews are less detailed and tell one side of the story. The increased use of frequent sentiment words suggests that the language people use is becoming less diverse. For example, in Amazon positive reviews, the percentage of highly frequent words rose from about 44% in 2000 to about 54% in 2018, which means the use of infrequent words, i.e., words that are not among the top 1% frequent words decreased by about 10% over the last couple of decades.

[15] IMDb is the only platform that requires a minimum length limit of 150 characters from its reviewers.

4.9 Qualitative analysis

To shed some light on the trends we observed, we look at the most intense reviews, i.e., the shortest reviews containing at least one intense sentiment term. We focus on IMDb reviews for this analysis, as they provide us with the longest period, from 2000 up to 2020. We consider terms with a sentiment score higher than three as intense terms for positive reviews, resulting in 35 terms, less than 5% of the lexicon words. Similarly, terms with minus-three scores are considered intense for negative reviews, resulting in 114 terms, about 15% of the lexicon words. Table 4 presents a representative sample out of the ten shortest intense reviews from 2000 and 2020 for both positive and negative reviews. We see that for reviews from early years, both positive and negative, the reviews are more informative, discussing different aspects of the movies. For example, reviews from earlier years often mention main characters, actors, and directors when conveying strong sentiment towards a movie. In addition, they are less one-sided, e.g., mentioning the positive aspects in highly negative reviews. In contrast, many current reviews do not contribute new information the reader could not infer from the reviewer's numerical rating—the review just echos the sentiment score.

4.10 Statistical significance analysis

To confirm that the trends we observe indeed indicate an overall increase or an overall decrease, we use the Mann-Kendall statistical test (Kendall, 1955; Mann, 1945) on each of the plots in Figs. 1 and 2. This test checks whether a time series represents an overall consistent increasing or decreasing trend, with a null hypothesis of no clear trend. Since we perform multiple tests (36 tests in total), we control the false discovery rate (FDR; at level $q = 0.05$) using the algorithm of Benjamini and Yekutieli (2005). This correction does not assume independence between the tests. We find that except for the absolute sentiment intensity for Yelp's negative reviews, all Amazon and Yelp plots have a significant trend with the FDR-adjusted p-values ($p < 0.01$). As expected, we do not reject the null hypothesis for most of the IMDb plots. However, if we focus our analysis on the last ten years of the IMDb data, then except for the plots describing the sentiment intensity and percentage of dichotomous reviews with IMDb negative reviews, all IMDb plots present a clear trend according to the Mann-Kendall test with an FDR correction ($p < 0.01$).

5 Persistent reviewers analysis

So far, our analysis of reviews has shown them to have become less comprehensive, less lexically rich, and more sentiment-intense over the years (see Sect. 4). For Amazon and IMDb, we found tipping points, i.e., specific years in which most of the trends intensify. We found that the tipping points are roughly shared across analyses and ratings for a given platform, e.g., 2013 is a tipping point for both positive and

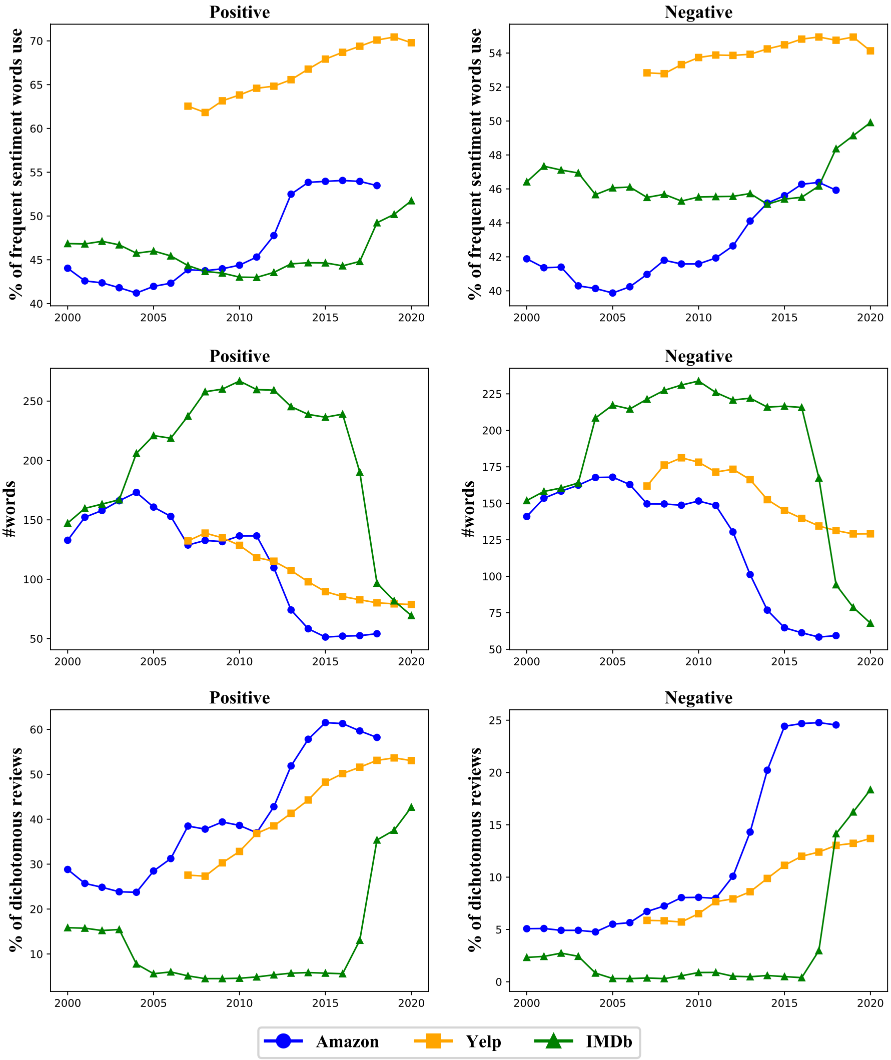

Fig. 2 The language diversity and review comprehensiveness metrics. The figures on the left correspond to the positive reviews; similarly, the figures on the right correspond to the negative reviews. Note the *y*-axes in the left plots and right plots are of a different scale

negative Amazon reviews, in which most of the trends mentioned above intensify.[16] Similarly, 2017 is a tipping point for IMDb. The trends we observed are calculated over the entire community of reviewers, thus not providing any evidence about individual behavior. To shed light on the nature of behavioral changes, we conduct

[16] We choose the years in which the number of words is rapidly decreasing, as we observe the rest of the trends often align with this metric. We exclude Yelp from this analysis, as Yelp trends are smooth and do not have an obvious tipping point. We comment on that in the end of the section.

Table 4 A sample from the shortest reviews containing at least one intense term

Year	Sent	Review
2000	Pos	Batman meets Beetlejuice Really nice... **Fantastic** story... As usual a **masterpiece** of Tim Burton and Johnny Depp
2000	Pos	B. Monkey was a **great** film! Johnny and Asia were **great**. I liked B's strong character and attitude. Good storyline. A must see.
2020	Pos	These girls are **awesome** how can u not like this???
2020	Pos	A **wonderful** movie i love it I cannot stop focusing
2000	Neg	I ***hated*** it!!!!!! I loved the first one cause it made you think. This one was just a shoot um up movie. While I do give credit for the great special effects. Who cares if the storyline sucks.
2000	Neg	The story was simply wrong, the actors were bad... I ***hated*** this Movie! Depardieu could do better, just have a look at his version of Cyrano...! Try the older version of 1979 with Jaques Weber instead!
2020	Neg	A complete cinematic ***disaster*** avoid it at all costs
2020	Neg	If you don't want to get ***cancer*** choose better movie

The terms marked with high intensity are in bold, with a **bold-underline** face for a positive sentiment and a ***bold-italic*** face for a negative one

the same analyses as before, considering only users who wrote a review each year for three consecutive years, where the second year is the tipping point. By comparing persistent users' behavior to that of the general users population, we hope to understand whether the change entirely stems from new users or whether existing users are also changing their writing habits. Figure 3 presents our analyses only for positive reviews for brevity, as their trends are similar to the negative ones. Our analysis shows that persistent users' reviews tend to be longer than those of general reviewers. For example, the length of Amazon's persistent users' reviews decreased from 237 to 190 words in two years. Similarly, the overall population reviews length decreased from 107 to 41 words over the same period. Persistent users' reviews are less positive and intense than those of general reviewers in both Amazon and IMDb sentiment analyses.[17] For example, the enhanced sentiment metric of IMDb's persistent users' reviews increased from 0.83 to 0.97 in two years. Similarly, the overall population reviews enhanced sentiment metric increased from 0.99 to 1.37 over the same period. Persistent users' reviews are also less one-sided and more lexically rich than those of the general reviewing population. While more moderate in terms of trends, the persistent users' reviews demonstrate similar behavior to those of general reviewers. These findings suggest that while not-persistent users are the main contributors to our observed trends, existing users are also actively changing their habits over the years.

While we did not find in the Yelp dataset a clear point in time with an intense change, we included it in our analyses for completeness. We chose 2014 as the baseline year for this dataset. In that year, there was the largest change in the number

[17] We observe similar patterns for the negative reviews, in which persistent users' reviews tend to be less negative and less intense.

of words written in the reviews. We observe that the shifts for persistent users and overall reviewers for this dataset are mild, as expected, with the smoothness of the Yelp trends.

6 Helpful reviews analysis

Online reviews enable potential customers to understand better the product or service they are interested in, thus assisting them in making better decisions. Going through the entire review section for highly popular products or services is time-consuming and in some cases, infeasible. To mitigate this problem, review platforms often encourage users to mark reviews that they find helpful to showcase those to future readers. Understanding the trends in helpful reviews is of great importance as many platforms present them at the top of the review section, thus increasing their visibility.

To better understand the nature of helpful reviews, we compare their trends with those of the general review population (see Fig. 4). Unlike Amazon and Yelp, which share the same helpfulness voting mechanism, in which you can up-vote a review, both up-votes and down-votes are allowed in IMDb. For up-vote-only platforms, such as Amazon and Yelp, we consider any review with ten or more helpfulness votes as helpful. For IMDb, where both down-votes and up-votes are allowed, we consider reviews with 50 or more helpfulness votes, in which the up-vote down-vote ratio is at least 2:1.[18] For brevity, we present our analyses for positive reviews, as their trends are similar to the negative ones. Our analyses show that helpful reviews tend to be longer compared to the general review population. This finding is not surprising as many previous studies support it (Karimi & Wang, 2017; Lu et al., 2018; Salehan & Kim, 2016). In all three sentiment analyses, helpful reviews are less positive and intense than the general review population.[19] While previous works find "extremeness" is highly indicative of helpfulness (Cao et al., 2011), their definition of polarization is driven by numerical ratings, for which we control in our experiments, thus, not relevant to us.[20] Helpful reviews are more lexically rich and less one-sided than those of the general review population. For all the analyses mentioned above, we see that the IMDb helpful reviews almost perfectly match the IMDb's general reviews trends, suggesting the magnitude of the effect on IMDb users, as helpful reviews are highly visible. In terms of trends, we can see that helpful reviews share similar trends to the general review population, although more moderate ones, for Amazon and Yelp. Due to their high visibility, helpful reviews

[18] If we lower the helpfulness threshold for IMDb, e.g., by considering reviews with fewer than 50 votes or reviews for which the up-votes down-votes ratio is smaller than 2:1, the helpful reviews trends are almost identical to the general reviews trends. While this definition of helpful reviews may seem ad-hoc, we provide full code for this experiment, and invite the readers to explore different thresholds and criteria.

[19] We observe similar patterns for the negative reviews, in which helpful reviews tend to be less negative and less intense.

[20] Cao et al. (2011) define "extremeness" of a review as the absolute value of the difference between the reviewers' rating and the average of all user ratings.

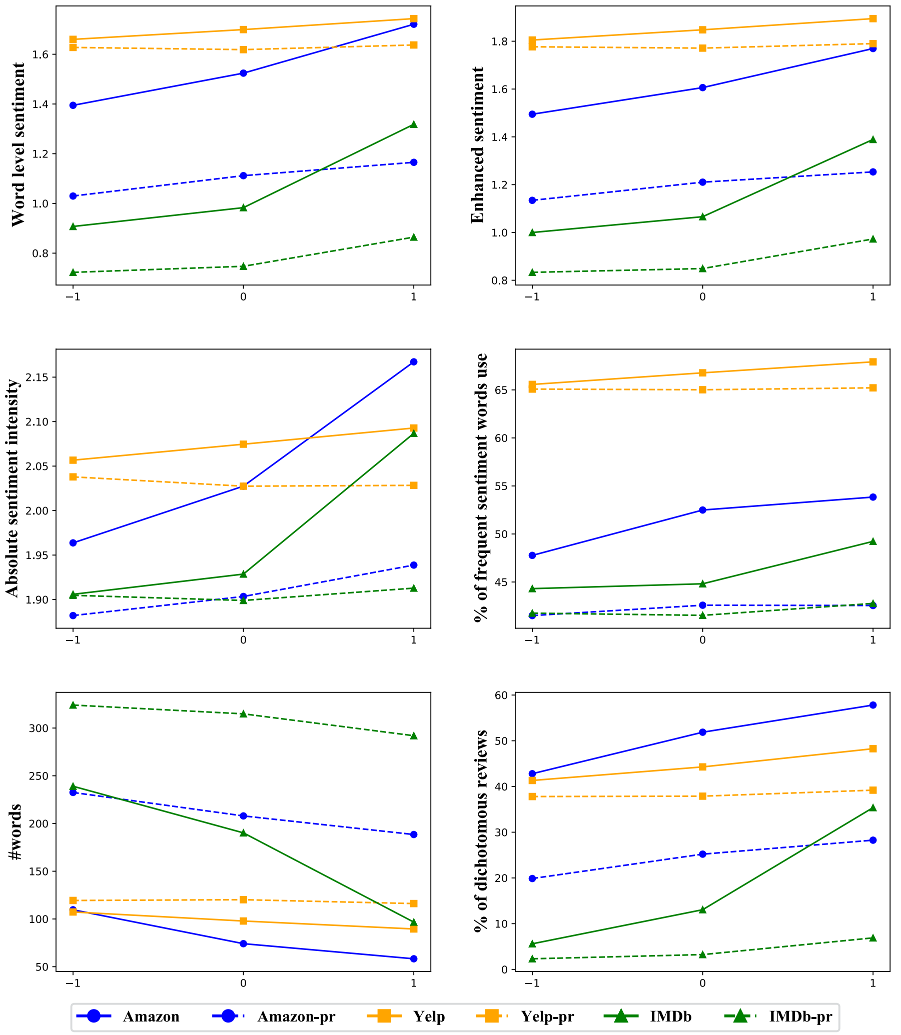

Fig. 3 A comparison between persistent users for positive reviews, who are represented by dashed lines, and the general users population (solid lines) in each platform. As different datasets have different tipping-point years, we align their graphs, where 0 on the x-axis indicates the tipping-point year. The results are similar for negative reviews

play a crucial role in the way people make decisions, and their perception of the entire online community. The analyses mentioned above show that helpful reviews are indeed changing as well over time, thus extending the impact of these trends beyond "picky" readers, as less than 15% of readers read more than 10 reviews before making their decision.[21]

[21] https://tinyurl.com/2b65xync.

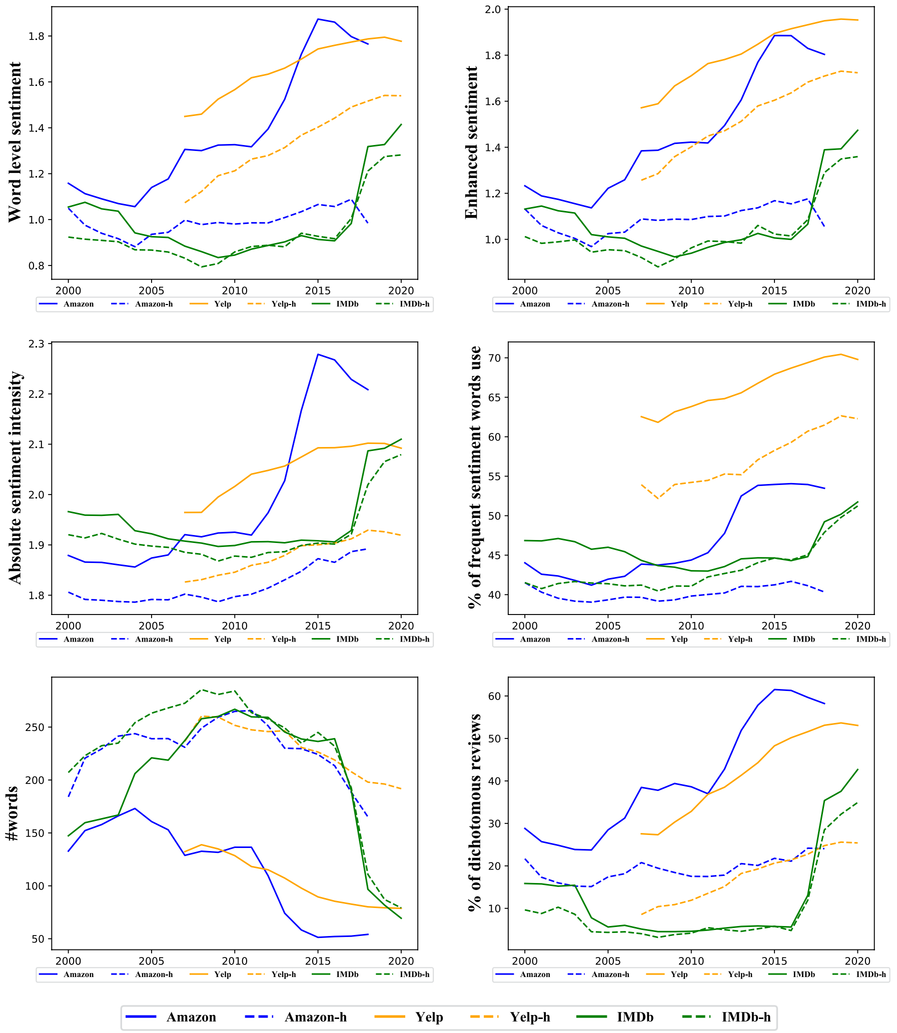

Fig. 4 A comparison between positive-sentiment helpful reviews, which are represented by the dashed lines, and the overall reviews in each platform

7 Analysis of reviewer authoring electronic device type

--> different data set, not necessary

The causes for the behavior we observe in the previous sections cannot be analyzed easily. The way people express themselves can be affected by various factors. Linguistic differences between different people can be a result of cultural differences (House, 1997; Ziv, 1988), gender (Wahyuningsih, 2018), and different political stances (Sylwester & Purver, 2015), to name a few examples. Here we analyze one technical factor that affects this behavior (Nicholas et al., 2013; Sellen et al., 2002)—the type of the electronic device on which a user composes their review.

More specifically, we examine the use of mobile devices. We do so for four main reasons: (1) Mobile device internet traffic accounts for more than 54% of total web traffic,[22] thus making mobile device usage a phenomenon of significant magnitude; (2) Mobile ownership and internet usage are steadily increasing in recent years, which aligns with our observed trends timeline, and are forecast to increase further as mobile technologies are becoming more affordable and available around the world;[23] (3) Online platforms can easily obtain knowledge about users' submission device, enabling those platforms to act upon our findings; (4) There is other work tying mobile device usage and users' online behavior.

The rise in mobile device use has affected the relation between people and online reviews. Okazaki (2009) shows that compared to personal computer users, mobile users are more motivated to take an active part in online review communities. By analyzing almost 300,000 reviews, Lurie et al. (2014) show that compared to personal computer user reviews, mobile device user reviews are less reflective, more focused on the present, less subject to retrospective biases, less cognitive, more one-sided, more negative, and less socially oriented. März et al. (2017) examine the perceived value of reviews that were published via mobile devices, finding that these reviews were perceived as less helpful and of lower value to the reader. Mariani et al. (2019) explore the differences between reviews of London hotels posted on Booking.com from a personal computer and ones posted from a mobile device. Their analyses find that the latter are shorter and appear to be less helpful. Moreover, the share of online reviews submitted by mobile devices has been increasing, reducing the share of reviews submitted by personal computer users.

The question is whether these differences can explain (even in part) the trends we have observed. To this end, we examine a different collection of reviews from Booking.com.[24] The dataset contains reviews from 2015 to 2017, segmented by the device the user used to submit them.[25] Unlike the previous review platforms we examine, Booking.com encourages its reviewers to express both positive and negative impressions in their reviews, aiming for a more balanced discourse.[26] While Booking.com review rating can range between one and ten, our dataset's lowest rating is three, and ratings below five are rare. On the other hand, positive reviews with a score of ten are common, with over 40,000 reviews sent from a personal computer and over 70,000 sent from a mobile device. We hence focus on positive reviews at the edges of the spectrum, enabling us to control for the user-perceived sentiment.

Table 5 compares review attributes with respect to the submission device. The comparison shows that reviews sent via mobile devices are substantially shorter and less diverse (in terms of most frequent sentiment words usage) than reviews sent from a personal computer. While the word-level sentiment analysis shows

[22] https://tinyurl.com/nhzrncry.

[23] https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/.

[24] https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe.

[25] The current version of Booking.com does not include the feature that allows us to segment the reviews by submission device. We therefore use the snapshot from 2015 to 2017.

[26] Booking.com provides two designated text boxes for positive and negative impressions of the hotel. We concatenate them to create one text chunk for our analyses.

that reviews sent from a mobile device are more positive than reviews sent from a personal computer, the enhanced sentiment analysis reveals that their sentiment difference is relatively small. As already noted, Booking.com encourages reviewers to express both negative and positive impressions, resulting in a similar percent of dichotomous reviews for both devices, overcoming mobile device users' tendency to submit one-sided reviews (Lurie et al., 2014). While the Booking.com review platform tamed mobile device users' preference for one-sided reviews and strong positivity compared to personal computer users, in absolute terms, the former is using stronger language. While the overall sentiment score is similar, reviews sent from mobile devices are composed of stronger language for both negative and positive terms.

We cannot directly project the Booking.com analysis to other platforms, as Booking.com differs with respect to the review mechanism and is the only platform for which we can segment reviews by their submission device. That said, it would be worth investigating whether the tendency of users of mobile devices to produce shorter reviews and use stronger sentiment words, can explain the trends we observe in Sect. 4. It would also be worth investigating whether users of personal computers and mobile devices are similar with respect to enhanced sentiment and one-sided reviews simply because Booking.com requires them to report both negative and positive views: Without this requirement, the similarity might disappear.

8 Conclusions and implications

In this paper, we demonstrate the linguistic changes in online reviews over time. Our findings show that, in general, reviews have become less diverse, less comprehensive, and more polarized. By analyzing user behavior over tipping points, we show these trends are not unique for new users but rather happen across the general reviewers' population. Additional analysis reveals that the trends mentioned above are taking place in helpful reviews as well, having a broad effect as these reviews are of high visibility. We offer an optional explanation for these trends, as we observe linguistic differences stemming from the texts submission device. We show that reviews sent from a mobile device are characterized by shorter texts and stronger sentiment terms, on average. The behavioral differences between mobile device users and personal computer users, combined with the growing popularity of mobile devices, might lead to the changes we observe (see Sect. 7 for more details). We believe our findings have multiple implications concerning diverse groups of people.

8.1 Implications for readers of online reviews

Whether it is for healthcare (Jucks & Thon, 2017; Witteman et al., 2016), finance (Li et al., 2019), or purchasing a product (Lackermair et al., 2013; Shihab & Putri, 2019), people often rely on reviews to help them make more informed decisions. As our findings show that reviews have gotten less comprehensive over the years, readers might consider reading older reviews, if available, or reading both negative and positive reviews to get a full picture. We advise taking under consideration a potential change in the expression of identical sentiment when considering a bundle of reviews from different years, and correcting for such possible variation when reviews from older years are compared to reviews from recent years. Given a specific reviewed item, though, there might be other factors in such a change that are relevant to the item (such as reviews becoming outdated).

8.2 Implications for writers of online reviews

Yoo and Gretzel (2008) show that two of the main motivations of people to write a hotel review are helping a travel service provider and their concerns for other consumers. In addition, they found that venting negative feelings is not a significant motivation for posting negative reviews. Later, Rensink (2013) reinforced these findings. For writing a helpful opinionated text, one needs to be mindful of the text length, as it is strongly correlated with its helpfulness (see Sect. 6 for more details). In addition, our findings show that reviews sent from a mobile device are shorter, supporting previous results. Reviews sent from a mobile device are also less helpful and perceived as such by the public (see Sect. 7 for more details). We, therefore, suggest that for submitting a useful opinionated text, one should be mindful of the review length and the device used for submitting the review.

8.3 Implications for platforms hosting online reviews

Whether they are the main service, a platform provides, e.g., in Yelp, or a complementary service for customers, e.g., in Amazon, online reviews are crucial for any platform's value proposition. A testimony to the importance of review quality is the extensive efforts that review platforms invest in moderating them and improving their credibility (see Sect. 2). We believe that review platforms, or any platform for opinionated texts, should monitor changes in the language used and adjust accordingly, increasing reviews' helpfulness. For example, Booking.com encouraging reviewers to articulate the hotel's negative and positive aspects might ease the percentage of highly one-sided reviews. High-quality reviews are crucial for new or less popular products or services where reviews are scarce.

The platforms mentioned above often provide users with artificial intelligence-driven services to enrich their experience and guide their shopping experience. For example, sentiment analysis and opinion mining, i.e., automatically predicting which sentiment a given text conveys in the sentence level (Liu, 2012; Wilson et al., 2005; Ziser & Reichart, 2017) and in the aspect level (Lekhtman et al., 2021; Ruder et al., 2016; Thet et al., 2010), allowing the users to understand better the sentiment conveyed towards each aspect of the product/service at hand. Since the linguistic traits of online reviews are changing over time, in some cases, over a few years, we

Table 5 Comparison between reviews sent by a mobile device (Mobile) and personal computer (PC)

Analysis	PC	Mobile
Word level sentiment	2.09	2.23
Enhanced sentiment	1.95	1.92
Absolute sentiment intensity	2.33	2.47
#words	32.31	24.08
% of dichotomous reviews	65	64
% of frequent sentiment words use	81	85
#reviews	41,222	74,631

The analyses are the same as in Sect. 4. #reviews stands for the number of reviews in the dataset we obtained, not the overall number of reviews sent by PC/Mobile that year

believe that these platforms' science and engineering teams should be mindful of the temporal aspect of their data. To further examine this subject, we trained and applied classifiers on data taken from the earliest and latest years we have available data for each one of the datasets[27] (Table 6). In eight out of the nine setups, the classifiers trained and tested on data from more recent years achieved higher results than their early years' counterparts.[28] Such higher results may indicate an increasing intensity and polarization in reviews, providing more evidence for the conclusions in our paper.

9 Limitations and future work

In our study, we used a lexicon to evaluate the sentiment of a review. As shown in Table 3 and discussed in Sect. 3, the error level of sentiment prediction from such a lexicon-based method can be quite high, around 25%, but sufficient for observing overall trends of sentiment. While neural methods provide higher accuracy, the main issue with using them to perform our assessment is that they need a training set that would represent a specific time period and their supervision granularity. We leave it for future work to adapt unsupervised neural methods, possibly more accurate than our lexicon-based methods, to assess the sentiment.

In addition, we have focused on examining language variation in reviews over time. However, it is essential to note that many factors could potentially contribute to such variation. While we have considered some of these factors in our analysis, there are others that we have not considered. These include changes in the user interface that companies use to collect reviews, which could affect the length and tone of the reviews, as well as demographic and geographic changes in the reviewer

[27] The train and test sizes are identical to the ones used in Sect. 3.

[28] The differences between early years and later years for all results, with the exception of BERT scores on the IMDb dataset, are statistically significant with $p < 0.01$. The early-year results of BERT on IMDb show higher scores than the recent-year results.

Table 6 Performance of different sentiment classifiers for the earliest and latest available years of each dataset

Model	Amazon		Yelp		IMDb	
	2000	2018	2007	2020	2000	2020
Lexicon	0.656	0.695	0.704	0.763	0.739	0.762
Enhanced lexicon	0.681	0.741	0.752	0.816	0.744	0.778
BERT	0.88	0.939	0.944	0.978	0.901	0.9

pool and the moderation guidelines enforced by each platform. Further research is needed to fully explore the impact of these and other factors on language variation in reviews. Despite this, our findings suggest that the language used in reviews has become more extreme in its tone, even when the reported sentiment level remains the same, and this trend is evident in multiple datasets.

In future work, we would like to extend our understanding of the trends we observe, for example, by exploring other factors such as users' age, their fluency in English, and their socio-economic background. Our discussion and study are limited to online reviews. One could apply the analysis in this paper to test whether the trend of language polarization has happened in other domains of language use. For example, recent work (Algan et al., 2017; Dorn et al., 2020; Gentzkow, 2016) shows political views have become more polarized over the recent years. It remains to be conclusively established both whether this is reflected in language use and vice versa—to what extent such polarization is deemed to increase due to changes in language use. While some of the possible factors we mentioned above could be relevant to such domains, other factors would have to be considered as well.

Acknowledgements The authors thank the anonymous reviewers for their thorough feedback. We also thank Mark Steedman and Robin Hill for discussions that improved the paper.

Funding No funding was received to assist with the preparation of this manuscript.

Data availibility Our code and data are available at https://github.com/yftah89/ReviewsOverTime.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Aggarwal, A. G., & Aakash. (2020). Analysing the interrelationship between online reviews and sales: The role of review length and sentiment index in electronic markets. *International Journal of Internet Marketing and Advertising, 14*(4), 361–376.

Aithal, M., & Tan, C. (2021). On positivity bias in negative reviews. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing* (Volume 2: Short Papers, pp. 294-304). Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2021.acl-short.39, https://aclanthology.org/2021.acl-short.39

Alaparthi, S., & Mishra, M. (2021). BERT: A sentiment analysis odyssey. *Journal of Marketing Analytics, 9*(2), 118–126.

Alattar, F., & Shaalan, K. (2021). Using artificial intelligence to understand what causes sentiment changes on social media. *IEEE Access, 9*, 61756–61767.

Algan, Y., Guriev, S., Papaioannou, E., & Passari, E. (2017). The European trust crisis and the rise of populism. *Brookings Papers on Economic Activity, 2017*(2), 309–400.

Alkhalifa, R., Kochkina, E., & Zubiaga, A. (2022). Building for tomorrow: Assessing the temporal persistence of text classifiers. ArXiv preprint. arXiv:2205.05435

AL-Sharuee, M. T., Liu, F., & Pratama, M. (2021). Sentiment analysis: Dynamic and temporal clustering of product reviews. *Applied Intelligence, 51*(1), 51–70.

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., & Wang, L. (2013). How noisy social media text, how diffrnt social media sources? In *Proceedings of the sixth international joint conference on natural language processing* (pp. 356–364).

Ben-David, E., Rabinovitz, C., & Reichart, R. (2020). PERL: Pivot-based domain adaptation for pre-trained deep contextualized embedding models. *Transactions of the Association for Computational Linguistics, 8*, 504–521. https://doi.org/10.1162/tacl_a_00328https://aclanthology.org/2020.tacl-1.33'

Ben-David, E., Ziser, Y., & Reichart, R. (2022). Domain adaptation from scratch. arXiv preprint arXiv:2209.00830

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.

Bjerva, J., Kouw, W., & Augenstein, I. (2019). Back to the future–sequential alignment of text representations. ArXiv preprint. arXiv:1909.03464

Blodgett, S. L., Green, L., & O'Connor, B. (2016). Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1119–1130). Association for Computational Linguistics, Austin, Texas. https://doi.org/10.18653/v1/D16-1120, https://aclanthology.org/D16-1120

Boucher, J., & Osgood, C. E. (1969). The Pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behavior, 8*(1), 1–8.

Cao, Q., Duan, W., & Gan, Q. (2011). Exploring determinants of voting for the helpfulness of online user reviews: A text mining approach. *Decision Support Systems, 50*(2), 511–521.

Crystal, D. (2001). Language and the internet. *Cambridge University Press*. https://doi.org/10.1017/CBO9781139164771

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013). No country for old members: user lifecycle and linguistic change in online communities. In D. Schwabe, V. A. F. Almeida, H. Glaser, et al. (Eds.), *22nd international world wide web conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013* (pp. 307–318). International World Wide Web Conferences Steering Committee / ACM. https://doi.org/10.1145/2488388.2488416

DeLucia, A., Wu, S., Mueller, A., Aguirre, C., Resnik, P., & Dredze, M. (2022). Bernice: A multilingual pre-trained encoder for twitter. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 6191–6205).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423

Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., & Danforth, C. M. (2015). Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences, 112*(8), 2389–2394.

Dorn, D., Hanson, G., Majlesi, K., & Majlesi, K. (2020). Importing political polarization? The electoral consequences of rising trade exposure. *American Economic Review, 110*(10), 3139–83.

Eisenstein, J. (2013). 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).

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1277–1287). Association for Computational Linguistics, Cambridge, MA. https://aclanthology.org/D10-1124

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. *PLoS ONE, 9*(11), e113,114.

Fagan, S., & Gençay, R. (2010). An introduction to textual econometrics. In *Handbook of empirical economics and finance* (pp. 133–153).

Gentzkow, M. (2016). Polarization in 2016. Toulouse Network for Information Technology Whitepaper (pp. 1–23).

Grieve, J., Nini, A., & Guo, D. (2017). Analyzing lexical emergence in modern American English online 1. *English Language & Linguistics, 21*(1), 99–127.

Hamilton, W., Zhang, J., Danescu-Niculescu-Mizil, C., Jurafsky, D., & Leskovec, J. (2017). Loyalty in online communities. In *Proceedings of the International AAAI conference on web and social media*

He, S., Hollenbeck, B., & Proserpio, D. (2021). The market for fake reviews. Available at SSRN 3664992.

Holtz, P., Deutschmann, E., & Dobewall, H. (2017). Cross-cultural psychology and the rise of academic capitalism: Linguistic changes in CCR and JCCP articles, 1970–2014. *Journal of Cross-Cultural Psychology, 48*(9), 1410–1431.

House, J. (1997). *Translation quality assessment: A model revisited*. Gunter Narr Verlag.

Hua, Y., Danescu-Niculescu-Mizil, C., Taraborelli, D., Thain, N., Sorensen, J., & Dixon, L. (2018). WikiConv: A corpus of the complete conversational history of a large online collaborative community. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2818–2823). Association for Computational Linguistics, Brussels, Belgium. https://doi.org/10.18653/v1/D18-1305, https://aclanthology.org/D18-1305

Hutto, C., & Gilbert, E. (2014). 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*

Jiang, Y., Meng, W., & Yu, C. (2011). Topic sentiment change analysis. In *International workshop on machine learning and data mining in pattern recognition* (pp. 443–457). Springer

Jucks, R., & Thon, F. M. (2017). Better to have many opinions than one from an expert? Social validation by one trustworthy source versus the masses in online health forums. *Computers in Human Behavior, 70*, 375–381.

Jurgens, D., Tsvetkov, Y., & Jurafsky, D. (2017). Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th annual meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 51–57). Association for Computational Linguistics, Vancouver, Canada. https://doi.org/10.18653/v1/P17-2009, https://aclanthology.org/P17-2009

Karimi, S., & Wang, F. (2017). Online review helpfulness: Impact of reviewer profile image. *Decision Support Systems, 96*, 39–48.

Kendall, M. (1955). *Rank correlation methods*. Griffin.

Kenski, K., Coe, K., & Rains, S. A. (2020). Perceptions of uncivil discourse online: An examination of types and predictors. *Communication Research, 47*(6), 795–814.

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens. *Vocabulary Learning and Instruction, 1*(1), 60–69.

Kumar, H., Harish, B., & Darshan, H. (2019). Sentiment analysis on IMDb movie reviews using hybrid feature extraction method. *International Journal of Interactive Multimedia & Artificial Intelligence, 5*(5).

Lackermair, G., Kailer, D., & Kanmaz, K. (2013). Importance of online product reviews from a consumer's perspective. *Advances in Economics and Business, 1*(1), 1–5.

Lampe, C., Wash, R., Velasquez, A., & Ozkaya, E. (2010). Motivations to participate in online communities. In Mynatt, E. D., Schoner, D., Fitzpatrick, G., et al. (eds) *Proceedings of the 28th* *international conference on human factors in computing systems, CHI 2010, Atlanta, Georgia, USA, April 10–15, 2010* (pp. 1927–1936). ACM. https://doi.org/10.1145/1753326.1753616

Lappas, T. (2012). Fake reviews: The malicious perspective. In *International conference on application of natural language to information systems* (pp. 23–34). Springer.

Lekhtman, E., Ziser, Y., & Reichart, R. (2021). DILBERT: Customized pre-training for domain adaptation with category shift, with an application to aspect extraction. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 219–230). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic. https://doi.org/10.18653/v1/2021.emnlp-main.20, https://aclanthology.org/2021.emnlp-main.20

Li, Z., Qian, Y., & Yuan, H. (2019). Users' opinions in online financial community and its impact on the market. In *2019 16th international conference on service systems and service management (ICSSSM)* (pp. 1–6), IEEE.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 5*(1), 1–167.

Lu, S., Wu, J., & Tseng, S. L. A. (2018). How online reviews become helpful: A dynamic perspective. *Journal of Interactive Marketing, 44*, 17–28.

Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science, 62*(12), 3412–3427.

Lurie, N. H., Ransbotham, S., & Liu, H. (2014). The characteristics and perceived value of mobile word of mouth. Marketing Science Institute Working Paper Series Report 14.

Mann, H. B. (1945). Nonparametric tests against trend. Econometrica: Journal of the Econometric Society 13: 245–259.

Mariani, M. M., Borghi, M., & Gretzel, U. (2019). Online reviews: Differences by submission device. *Tourism Management, 70*, 295–298.

März, A., Schubach, S., & Schumann, J. H. (2017). Why would I read a mobile review? Device compatibility perceptions and effects on perceived helpfulness. *Psychology & Marketing, 34*(2), 119–137.

McCarthy, P. M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). PhD thesis, The University of Memphis.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing, 15*(3), 323–338.

Meftah, S., Semmar, N., Tamaazousti, Y., Essafi, H., & Sadat, F. (2021). On the hidden negative transfer in sequential transfer learning for domain adaptation from news to tweets. In *Proceedings of the second workshop on domain adaptation for nlp* (pp. 140–145).

Ni, J., Li, J., & McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 188–197). Association for Computational Linguistics, Hong Kong, China. https://doi.org/10.18653/v1/D19-1018, https://aclanthology.org/D19-1018

Nicholas, D., Clark, D., Rowlands, I., & Jamali, H. R. (2013). Information on the go: A case study of Europeana mobile users. *Journal of the American Society for Information Science and Technology, 64*(7), 1311–1322.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903

Okazaki, S. (2009). Social influence model and electronic word of mouth: PC versus mobile internet. *International Journal of Advertising, 28*(3), 439–472.

Pavalanathan, U., & Eisenstein, J. (2015). Audience-modulated variation in online social media. *American Speech, 90*(2), 187–213.

Androutsopoulos, I., Manandhar, S., AL-Smadi, M., & Eryiğit, G. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)* (pp. 27–35). Association for Computational Linguistics, Dublin, Ireland. https://doi.org/10.3115/v1/S14-2004, https://aclanthology.org/S14-2004

Qudar, M. M. A., & Mago, V. (2020). Tweetbert: A pretrained language representation model for twitter text analysis. arXiv preprint arXiv:2010.11091

Rensink, J. (2013). What motivates people to write online reviews and which role does personality play? A study providing insights in the influence of seven motivations on the involvement to write positive and negative online reviews and how five personality traits play a role. Master's thesis, University of Twente.

Ruder, S., Ghaffari, P., & Breslin, J. G. (2016). A hierarchical model of reviews for aspect-based sentiment analysis. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 999–1005). Association for Computational Linguistics, Austin, Texas. https://doi.org/10.18653/v1/D16-1103, https://aclanthology.org/D16-1103

Salehan, M., & Kim, D. J. (2016). Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. *Decision Support Systems, 81*, 30–40.

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.

Schnabel, T., & Schütze, H. (2014). Flors: Fast and simple domain adaptation for part-of-speech tagging. *Transactions of the Association for Computational Linguistics, 2*, 15–26.

Schoenmueller, V., Netzer, O., & Stahl, F. (2020). The polarity of online reviews: Prevalence, drivers and implications. *Journal of Marketing Research, 57*(5), 853–877.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE, 8*(9), e73,791.

Sellen, A. J., Murphy, R., & Shaw, K. L. (2002). How knowledge workers use the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 227–234).

Severyn, A., & Moschitti, A. (2015). Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 464–469).

Shihab, M. R., & Putri, A. P. (2019). Negative online reviews of popular products: Understanding the effects of review proportion and quality on consumers' attitude and intention to buy. *Electronic Commerce Research, 19*(1), 159–187.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior, 7*(3), 321–326.

Sylwester, K., & Purver, M. (2015). Twitter language use reflects psychological differences between democrats and republicans. *PLoS ONE, 10*(9), e0137,422.

Tan, S., Li, Y., Sun, H., Guan, Z., Yan, X., Bu, J., Chen, C., & He, X. (2013). Interpreting the public sentiment variations on twitter. *IEEE Transactions on Knowledge and Data Engineering, 26*(5), 1158–1170.

Thet, T. T., Na, J. C., & Khoo, C. S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science, 36*(6), 823–848.

Vinkers, C. H., Tijdink, J. K., & Otte, W. M. (2015). Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: Retrospective analysis. *Bmj, 351*.

Wahyuningsih, S. (2018). Men and women differences in using language: A case study of students at STAIN Kudus. *EduLite: Journal of English Education, Literature and Culture, 3*(1), 79–90.

Wang, J., Ding, K., Hong, L., Liu, H., & Caverlee, J. (2020). Next-item recommendation with sequential hypergraphs. In J. Huang,Y. Chang, X. Cheng, et al. (Eds.), *Proceedings of the 43rd international ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020* (pp. 1101–1110). ACM. https://doi.org/10.1145/3397271.3401133

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL student research workshop* (pp. 88–93). Association for Computational Linguistics, San Diego, California. https://doi.org/10.18653/v1/N16-2013, https://aclanthology.org/N16-2013

Wen, J., & Lei, L. (2022). Linguistic positivity bias in academic writing: A large-scale diachronic study in life sciences across 50 years. *Applied Linguistics, 43*(2), 340–364.

Wilson, S. M., & Peterson, L. C. (2002). The anthropology of online communities. *Annual Review of Anthropology, 31*(1), 449–467.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing* (pp. 347–354). Association for Computational Linguistics, Vancouver, British Columbia, Canada, https://aclanthology.org/H05-1044

Witteman, H. O., Fagerlin, A., Exe, N., Trottier, M. E., & Zikmund-Fisher, B. J. (2016). One-sided social media comments influenced opinions and intentions about home birth: An experimental study. *Health Affairs, 35*(4), 726–733.

Wu, Y., Ngai, E. W., Wu, P., & Wu, C. (2020). Fake online reviews: Literature review, synthesis, and directions for future research. *Decision Support Systems, 132*(113), 280.

Xu, H., Liu, B., Shu, L., & Yu, P. S. (2019). BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 2324–2335). Association for Computational Linguistics, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1242, https://aclanthology.org/N19-1242

Yoo, K. H., & Gretzel, U. (2008). What motivates consumers to write online travel reviews? *Information Technology & Tourism, 10*(4), 283–295.

Zanzotto, F. M., & Pennacchiotti, M. (2012). Language evolution in social media: A preliminary study. *Linguistica Zero*.

Zhang, Y., & Chen, X. (2020). Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval, 14*(1), 1–101.

Zhao, Z., Ziser, Y., & Cohen, S. B. (2022). Understanding domain learning in language models through subpopulation analysis. In *Proceedings of the Fifth BlackboxNLP workshop on analyzing and interpreting neural networks for NLP*

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

Ziser, Y., & Reichart, R. (2017). Neural structural correspondence learning for domain adaptation. In *Proceedings of the 21st conference on computational natural language learning (CoNLL 2017)* (pp. 400–410). Association for Computational Linguistics, Vancouver, Canada. https://doi.org/10.18653/v1/K17-1040, https://aclanthology.org/K17-1040

Ziv, A. (1988). Teaching and learning with humor: Experiment and replication. *The Journal of Experimental Education, 57*(1), 4–15.