# Sentiment Analysis and Ratings of Professors: A Comparison of Rate-My-Professor and Department Results

**Faruk Guder**
**Loyola University Chicago**

**Mary Malliaris**
**Loyola University Chicago**

*Previous studies employing numeric scores have observed that anonymous postings on the Rate-My-Professor (RMP) site tend to be more negative than ratings within a school. In this study, rather than using numeric evaluations, we compared the official UNIV evaluations with the RMP evaluations at the individual faculty course level, employing sentiment analysis on the text comments in evaluations. We compared positive sentiments in ratings on RMP to those in a specific business school department at a university. Our results show a statistically significant difference, with higher positive ratings at the university. We also analyzed emotions using NRC, finding a significant difference, with RMP having higher levels of negative emotions.*

*Keywords: text mining, teacher course evaluations, Rate-My-Professor, emotions in text*

## INTRODUCTION

Researchers have studied teacher performance in many ways. Initially, performance evaluations took place in a classroom, typically at the end of a course, with the teacher absent from the classroom. Computers made data collection easier and gave us other ways to collect data. Unregulated forums began that collected comments about our interactions with many parts of our daily lives: restaurants, movies, businesses, and classes. About twenty years ago, Rate My Professor was created, and has become a well-known destination site for comments about teachers and their performance. Rate My Professor (RMP) is an online site where students are allowed to post anonymous reviews about professors and classes they have taken. RMP is a business. As with many businesses, word of mouth and the freedom to write negative reviews freely encourage visits. Scholars who have studied online communication have found that many audiences prefer negative news and that negative events are more "contagious" than positive ones (Rozin and Royzman, 2001; Van der Meer et al., 2020; Trussler & Soroka, 2014; Lengauer et al., 2011).

According to the RMP site, they currently have over 19 million posted ratings on 1.7 million professors from 7,500 schools (RateMyProfessors.com, 2021). This site was created in 1999 by software engineer John Swapceinski. Swapceinski sold the site in 2005. It has had several owners, one being the live streaming financial news network Cheddar. In 2019, Cheddar was acquired by Altice USA for $200 million, though the RMP site is still run under the brand of Cheddar. Altice is a broadband communications provider based in New York City. In addition to Cheddar, it owns several cable networks, a digital advertising unit, and

internet, telephone and television services, among others. Altice is publicly traded on the New York Stock Exchange and reports year-over-year total revenue of $2.57 billion (Altice USA Reports Third Quarter 2021 Results, 2021). Online estimates of the revenue brought in annually by RMP vary from $2.5 million to $3.4 million.

A number of researchers have studied various aspects of the evaluations in RMP and examined the relationships between the numeric instructor quality ratings and various characteristics such as gender, quantitative vs humanities and arts, easiness, and sexiness (hotness). Rosen (2018) found positive correlations between ratings of instructor quality and easiness, as well as between instruction quality and easiness. Rosen also observed lower RMP ratings in science and engineering disciplines than in the humanities and arts. Boehmer & Wood (2017) reported gender bias, showing that the male instructors have higher teaching scores than women. Felton et al. (2004) reported high positive correlations between quality and easiness and also observed that students give sexy-rated professors higher quality scores. Otto, et al. (2008) show that the average helpfulness and average clarity are strongly correlated. Katrompas & Metsis (2021) investigated Rate My Professors for bias, inaccuracy, and invalid data. They find evidence supporting the theory that the type of data collection used by RMP is defective and inappropriate as an assessment for faculty evaluation.

Student evaluations of teachers have also been widely studied. Kim & Hodge (2000) show that the student perception of a professor is an important factor in student evaluations. Balkin et al., (2021) find that women are more likely to experience inequalities when they teach management in U.S. business schools. The recent introduction of online evaluations has made the evaluation process more efficient. There have been extensive studies comparing the results of in-class and online evaluations conducted by universities (Guder & Malliaris, 2010; Donovan et al., 2006; Morrison, 2013). One advantage of evaluations administered by a school is that student participation can be higher, and therefore, more representative of the actual student's experience in a class than the sporadic comments posted on RMP.

The studies mentioned above, which explore various aspects of evaluations on RMP, have exclusively utilized the numerical values from RMP and compared them across different groups (e.g., gender, discipline, etc.). However, as of now, no study has undertaken a direct comparison between the official UNIV evaluations and the RMP evaluations at the individual faculty course level.

In this study, we utilize the official teaching evaluations (UNIV) provided by the university, focusing on professors within a specific department in the school of business. We then compare these evaluations with the postings on RateMyProfessors (RMP) using Sentiment Analysis. Specifically, we conduct and analyze the following comparisons:

- Sentiment scores of the evaluations by UNIV and RMP
- Emotion scores of the group of evaluations by UNIV and RMP

The paper is structured as follows. Section 2 describes Sentiment Analysis and emotion scores with the bing and nrc lexicons. Section 3 states our hypotheses and discusses the data and methodology that will be used for testing. This is followed by a discussion of the results.

## SENTIMENT AND EMOTION SCORES

Sentiment analysis, a methodology within the domain of text mining, is used to understand and analyze opinions and feelings of the customers using textual reviews. This subset of text mining focuses on analyzing the feelings conveyed in the text.

Text mining is the process of deriving meaningful information from unstructured text data. The text data could be from online reviews, social networks, emails, call center interactions or other data sources. During the text mining process, data is transformed from unstructured text into a structured format (into a table/matrix) in order to identify meaningful patterns and new insights.

Sentiment analysis is a widely used text mining application that can track customer sentiment about a company, a product, or a service. Sentiment analysis mines the text with the objective of understanding the opinion expressed by it. The analysis classifies the terms (words) in the text as positive, negative, or neutral in order to track customer opinions (sentiments). Typically, the sentiments are quantified with a positive or

negative value, and neutral words are ignored. The sentiment score is calculated as the difference between the total number of positive and negative words in the text. An overall positive value for the sentiment score indicates that more positive words occurred in the text. A negative sentiment score reflects more negative than positive words used in the text description.

In this study, we have used Lexicon-based sentiment analysis. This type of sentiment analysis uses a pre-prepared sentiment lexicon to assign a sentiment value (positive or negative sentiment) to the terms in the text data. Each word that matches a corresponding word in the lexicon is assigned a sentiment value. The words not included in the lexicon are considered to be neutral words.

There are three commonly used general-purpose lexicons. These are bing (Hu & Liu, 2004), nrc (Mohammad & Turney, 2013) and affin (Nielson, 2011). These lexicons are available in the syuzhet and tidytext packages in R. The bing lexicon contains 6,786 words and categorizes words in a binary fashion into positive and negative categories. The nrc lexicon contains 5,468 words and categorizes words in a binary fashion into positive and negative sentiments. This lexicon also identifies 8 different emotions that the words represent, including anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. One word may be associated with more than one emotion. The affin lexicon contains 2,476 words and assigns words a score that runs between -5 and 5, with positive scores indicating positive sentiment and negative scores indicating negative sentiment.

In this analysis, we used the bing lexicon, with the largest number of words, to calculate the positive and negative sentiment scores for each individual instructor. The basic idea in sentiment analysis is to find the polarity of the text in order to classify it as positive, negative, or neutral. With large amounts of text, this computer aided process enables us to quantify the overall feeling and thus helps in human decision making. This task, detecting positive or negative sentiments from internal or external data sources, allows one to track changes in customer attitudes over time. It is commonly used to provide information about perceptions of brands, products, and services. Thus, the sentiment analysis task involves reading the text data, creating a corpus (the bag of words), cleaning the corpus, and calculating the sentiment scores using one of the lexicons.

An illustration from raw data to clean data to sentiment can be seen in the following example:

Consider a teacher course evaluation consisting of the following three reviews:

*[1] Course design was great, very easy to understand material and proper deadlines.*

*[2] Some of the instructions were confusing and homework assignments were difficult. Angry when students are late to the class. But, overall, it was an engaging, pleasant, and enjoyable experience.*

*[3] Whenever asked, the instructor provided constructive guidance to make the class engaging and enjoyable. Great professor and great organization.*

Is the sentiment expressed in these evaluations positive or negative overall? The reader is encouraged to evaluate them before moving to the bing analysis below.

Cleaning the data involves converting the text to lower case, removing common stop words (such as “and”, “is”, “that”, “but”, “since”, etc.), removing other words that do not reflect sentiments, removing punctuations, numbers, and white spaces. After cleaning the data, the text data in the example problem will be transferred to the following for sentiment analysis.

> *course design great easy understand material proper deadlines instructions confusing homework assignments difficult angry students late class overall engaging pleasant enjoyable experience whenever asked instructor provided constructive guidance make class engaging enjoyable great professor great organization*

Text mining identifies the words with positive and negative sentiments as listed in Table 1. The bing lexicon is used in this classification.

**TABLE 1**
**EXAMPLE OF POSITIVE AND NEGATIVE SENTIMENT WORDS**

| word | sentiment |
|---|---|
| great | positive |
| easy | positive |
| proper | positive |
| confusing | negative |
| difficult | negative |
| angry | negative |
| engaging | positive |
| pleasant | positive |
| enjoyable | positive |
| constructive | positive |
| guidance | positive |
| engaging | positive |
| guidance | positive |
| great | positive |
| great | positive |

From Table 1, we calculate the sentiment score, which is the difference between the number of positive and negative words in the text.

| Positive | Negative | Sentiment Score |
|---|---|---|
| 12 | 3 | 9 |

The final sentiment score of 9 is a very strong positive sentiment for this instructor (instructor 1). This calculation is very simple and straightforward and indicates whether the overall sentiment is positive or negative. When you compare two documents (e.g., reviews for two instructors), this sentiment score may favor longer reviews because they tend to have more counts of positive/negative words. For example, another instructor (instructor 2) with more reviews may have the following positive and negative sentiments.

| Positive | Negative | Sentiment Score |
|---|---|---|
| 24 | 12 | 12 |

The second instructor has a sentiment score of 12, which is higher than the sentiment score of the first instructor. To address the issue of comparing sentiments with a different number of reviews, additional data manipulationis necessary. We normalize the sentiment scores to eliminate the effect of longer vs shorter reviews. One method of normalizing is to use the ratio of positive words to the total number of sentiment words in text. This is calculated as follows:

$$\text{Percent Positive Score} = \frac{\text{number of positive words} * 100}{\text{number of positive words} + \text{number of negative words}} \quad (1)$$

The percent positive scores for the first and second instructors are (9*100)/15 and (12*100)/36, respectively. These ratios indicate that 80% of the sentiment words used are positive for the first instructor and that only 50% of the sentiment words used are positive for the second instructor.

In this paper, the percent method is used to compare the sentiment scores of the reviews by the school and RMP. That is, we will calculate the percent of positive sentiment in the School reviews and in the RMP reviews, then compare those numbers to avoid bias from review length.

Note that a text mining tool is not needed to determine the sentiments for a small example like this. But for text data involving a large number of reviews, it can be difficult to quantify or analyze the overall sentiment without the aid of a computer-generated analysis.

For the last sentiment comparison, we used the nrc lexicon to categorize specific emotions captured by the text. This lexicon aggregates the words used in the text comments into categories described as anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. Some of these emotions are negative: anger, disgust, fear, sadness; some are positive: anticipation, joy, trust. Surprises can be either positive or negative. We will test to see whether or not the distribution of words across these emotions are the same in the UNIV and RMP evaluations.

## DATA, HYPOTHESES AND METHODOLOGY FOR ANALYSIS

### Data

For this study, we collected data from reviews conducted by the University and from the Rate My Professor site. From within the University, this represented 2,284 reviews from students over 114 sections. From Rate My Professor, there were 500 reviews representing 37 classes. All data used reflected one department within the school of business. It included both core and major classes taken by both business and non-business students. The university data was obtained from the within-school evaluations administered by the university at the end of each term and included data from Fall 2019 through Summer 2021. All evaluations were done online. Data used from the university set included the instructor's name, the class name and section, the number of students enrolled in the class, the number who responded to the survey, an overall rating of the instructor (from 1 to 5), and textual comments from the students.

Data from the Rate My Professor site was acquired by downloading the reviews for each instructor from the RMP site. Data downloaded from the RMP site included the instructor's name, the class name, an overall numeric rating of the instructor, and textual comments from the students. Data at the RMP site is accessible by first specifying the school's name, then the name of the instructor. A drop-down list allows you to see the list of classes for which comments are available. After selecting a class name, you have access to a link with each student's overall quality rating (from 1-lowest to 5-highest) and individual comments for the course. The individual comments have some summary choices for the student to make that describe the class (is attendance mandatory, is the class for credit, etc.) followed by open space for the student to make any comments they wish. A set of words is also given from which the student can select descriptor tags for the instructor (caring, respected, awful, tough grader, etc.).

The sentiment analysis described in Section 2 is used to calculate sentiment scores for each individual instructor. A positive sentiment value is calculated by summing the instances of positive words occurring in the text, using the Bing lexicon.

The negative sentiment value is calculated in a similar way, using the negative words. The overall percent of positive sentiment, as described in Section 2 is the number of positive words divided by the sum of positive and negative words, times 100. Thus, for each instructor, two separate sentiment scores are developed that represent the percent of positive sentiment in each group of text comments; one using the text comments in the teacher course evaluations using the school results, and another using the reviews provided at Rate My Professor.

### Hypotheses and Methodology

In order to evaluate the results from the data obtained both from the school and the website, we consider the following hypotheses:

***Hypothesis 1****: There is no difference in the positive sentiment percent scores of the evaluations by UNIV and RMP.*

For each instructor, the university supplied text data was cleaned, and the bing lexicon was used to calculate a positive and a negative sentiment score. Further, the percent positive scores are calculated for each instructor. In the RMP site, comments from students that had been downloaded were also cleaned and processed using the bing lexicon. Positive and negative scores were calculated for each instructor over all the classes taught. Next, the percent positive scores are calculated. These percent positive scores for each instructor were then used in a paired t-test to evaluate this hypothesis.

To give additional insight into the data, an overall departmental score was calculated for each text set and compared. We also looked at the correlation between the positive percents for the instructors in these data sets.

***Hypothesis 2****: There is no difference in the emotions in the text from the two sources: UNIV evaluations and RMP.*

To further investigate the sentiment displayed by the texts, the nrc lexicon was used to calculate the number of times each of the eight emotions was referenced for the overall dataset. These values were changed from raw data to percents. This transformation was necessary in order to compare the different-sized data sets. A paired t-test was then used to compare the percents of the eight emotions in the University and RMP data.

## RESULTS AND DISCUSSION

### Results of Sentiment and Emotion Analysis

The percentage of positive sentiment scores for each instructor are given in Table 2. The number of reviews for each instructor is not the same since internal evaluations have a much greater response rate than those on RMP. Therefore, the magnitudes of the sentiment scores are not the same. The instructors with a larger number of reviews will have larger number of positive and negative sentiments. Thus, as described above, we have standardized them by using the percent positive scores as calculated in equation (1).

We see that the percentage of positive reviews by RMP are lower for all instructors except two. A paired t-test on these scores yields a t-Statistic of -3.16 with a two-tailed critical value of 2.23 and a P value of 0.01. This shows that the difference between percent positive scores is significant at 0.05 significance level. We can reject our null hypothesis that the amount of positive sentiment is equal in both UNIV and RMP text. Inspecting the average values shown in Table 2, we see that 83.18% of the internal bing lexicon words were positive compared to only 71.73% in the RMP reviews. This supports the belief that the reviews at RMP are less positive (more negative) than the reviews by the school.

**TABLE 2**
**PERCENT POSITIVE SCORES PER INSTRUCTOR USING BING LEXICON**

| Instructor | Percent Positive LUC | Percent Positive RMP |
|---|---|---|
| Instructor1 | 90.82% | 88.41% |
| Instructor2 | 83.86% | 78.79% |
| Instructor3 | 84.37% | 76.47% |
| Instructor4 | 81.71% | 69.57% |
| Instructor5 | 91.51% | 70.00% |
| Instructor6 | 85.13% | 67.69% |
| Instructor7 | 88.36% | 95.59% |
| Instructor8 | 86.24% | 70.97% |

| Instructor | Percent Positive LUC | Percent Positive RMP |
|---|---|---|
| Instructor9 | 71.05% | 50.00% |
| Instructor10 | 73.68% | 40.00% |
| Instructor11 | 78.21% | 81.58% |
| Average | 83.18% | 71.73% |

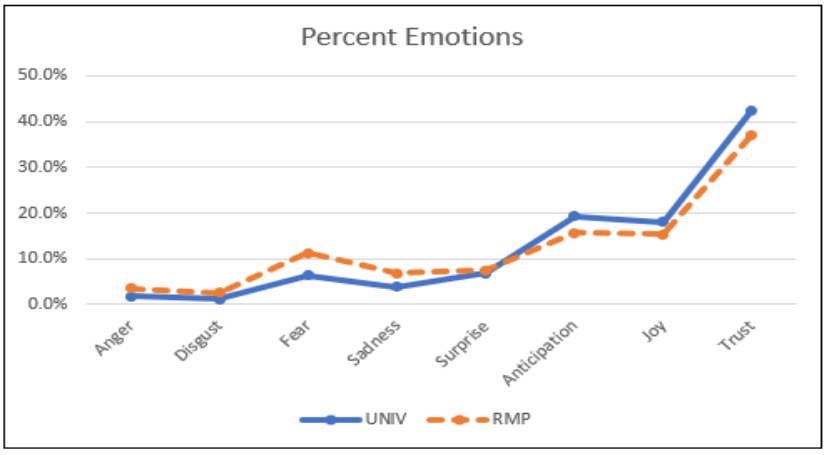
When we apply the nrc lexicon to the emotion expressed in these texts, we also see a difference. Table 3 displays the percentage of words in the student texts that the nrc lexicon classifies as describing each of these eight emotions. Figure 1 displays these graphically. Inspecting this Table, we see that the RMP percents are about twice as high for the emotions of anger, disgust, fear, and sadness, all negative emotions. In contrast, the UNIV texts are higher for the emotions of anticipation, joy, and trust, all positive emotions. For the remaining emotion, surprise, the values are much closer, being only slightly higher in the RMP texts.

**TABLE 3**
**PERCENT OF WORDS REPRESENTING EACH EMOTION IN THE NRC LEXICON**

| Emotion | Percent in UNIV text | Percent in RMP text |
|---|---|---|
| Anger | 1.8% | 3.5% |
| Disgust | 1.2% | 2.6% |
| Fear | 6.3% | 11.3% |
| Sadness | 3.8% | 6.9% |
| Surprise | 6.9% | 7.5% |
| Anticipation | 19.3% | 15.7% |
| Joy | 18.1% | 15.3% |
| Trust | 42.5% | 37.1% |

When we compare the percent of words falling into each of these emotions using a paired t-test, we get a t-statistic of -3.42E-16 with a critical value of 2.36, leading us to conclude that the percentages are not equal in the paired University and RMP texts. Inspection of these emotions shows that it is the negative emotions, Anger, Disgust, Fear, Sadness, that are higher in the RMP texts, and the positive emotions, Anticipation, Joy, Trust, that are lower in the RMP texts. This reinforces our decision that RMP descriptions are more negative and less positive.

**FIGURE 1**
**PERCENT OF TEXTS REPRESENTING EACH NRC EMOTION**



## SUMMARY AND RECOMMENDATION

In this paper, we used Sentiment analysis to analyze and compare text comments about professors from within one department using the texts from two different sources: Rate-My-Professor and the university end-of-term evaluations. While previous studies have analyzed the numeric scores in RMP, this paper focuses on the text accompanying the evaluations. The application of the Sentiment analysis methodology was based on two different standard lexicons: bing and nrc.

After cleaning the data to remove neutral words, the bing lexicon was applied to get a count of the positive and negative words used to describe each professor. The percentage of positive words from each source were compared using a paired t-test. There was a significant difference in the percent of positive sentiment expressed, with the university evaluations being more positive. The university evaluations represent a much larger sample of opinions than one finds on the RMP site. Given this larger sample in the university evaluations, we find that students express more positive sentiment overall in this larger group.

We next applied the nrc lexicon to identify the amount of specific nrc-identified emotions contained in the texts. The nrc lexicon identifies eight individual emotions. We found that the negative emotions occurred in higher percentages in the RMP evaluations than in those administered by the university, and that the percent of positive emotions occurred in lower amounts in RMP. This finding also supports the idea that the RMP texts have a more negative emphasis.

Thus, we found that the sentiments expressed in the RMP texts are less positive than those of the texts obtained from the end-of-term university assessments; the percent of negative emotion in RMP texts is greater than the percent of negative emotion in university text assessment while the percent of positive emotions expressed in the RMP texts was smaller than the percent of positive emotions expressed in the university assessments. Our conclusion from the analysis of this small case study is that the RMP textual comments, while perhaps more entertaining, are significantly less positive in tone than those of the larger, more representative comments from the classes as a whole.

While Sentiment analysis is a much newer tool than standard numeric analysis, we find that it offers expanded insight into the data available from course evaluations. It enables the user to quantify and compare the positive and negative attitudes within the textual comments. It also enables the user to identify which emotions are most evoked by professors or departments.

Some suggestions for future research would be to gain access to data from all departments within a school and repeat this type of analysis. As lexicons develop, it might also be possible to identify more specific sentiments or emotions.

## REFERENCES


Altice USA Reports Third Quarter 2021 Results. (2021, November 24). *Financial news details*. Retrieved from https://investors.alticeusa.com/investors/overview/financial-news-details/2021/Altice-USA-Reports-Third-Quarter-2021-Results/default.aspx

Altice USA. (2021, December 14). In *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Altice_USA

Balkin, D.B., Trevino, L.J., & Straub, S. (2021). The effect of gender inequalities in the classroom and beyond in U.S. business schools. *Journal of Management Education*. Retrieved from https://doi.org/10.1177/10525629211045604.

Boehmer, D.V., & Wood, W.C. (2017). Student vs faculty perspectives on quality instruction: Gender bias, 'hotness', 'easiness' in evaluating teaching. *Journal of Education for Business*, *92*(4), 173–178.

Cheddar (TV channel). (2021, October 20). In *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Cheddar_(TV_channel)downloaded from site on Dec 13, 2021

Donovan, J., Mader, C.E., & Shinsky, J. (2006). Constructive student feedback: Online vs. traditional course evaluations. *Journal of Interactive Online Learning*, 5(3), 283–296.

Felton, J., Micthell, J., & Stinton, M. (2004). Web-based student evaluations of professors: The relations between perceived quality, easiness, and sexiness. *Assessment and Evaluation in Higher Education*, *29*(1), 91–108.
Guder, F., & Malliaris, M. (2010). Online and paper course evaluations. *American Journal of Business Education*, *3*(2), 131–138. https://doi.org/10.19030/ajbe.v3i2.392
Hu, M., & Liu, B. (2004). *Mining Opinion Features in Customer Reviews*. Retrieved from https://www.aaai.org/Papers/AAAI/2004/AAAI04-119.pdf
Katrompas, A., & Metsis, V. (2021). Rate my professors: A study of bias and inaccuracies in anonymous self-reporting. *2021 2nd International Conference on Computing and Data Science (CDS)*, pp.536–542. DOI: 10.1109/CDS52072.2021.00098
Kim, C., & Hodge, N. (2000). Professor attitude: Its effect on teaching evaluations. *Journal of Management Education*, *24*(4), 458–473.
Lengauer, G., Esser, F., & Berganza, R. (2011). Negativity in political news: A review of concepts, operationalizations and key findings. *Journalism*, *13*(2), 179–202. https://doi.org/10.1177/146488491142780
Mohammad, S.M., & Turney, P.D. (2013). *Crowdsourcing a Word-Emotion Association Lexicon*. https://doi.org/10.48550/arXiv.1308.6297
Morrison, K. (2013). Online and paper evaluations of courses: A literature review and case study. *Educational Research and Evaluation*, *19*(7), 585–604. https://doi.org/10.1080/13803611.2013.834608
Nielson, F.A. (2011). *Sentiment Analysis (Lexicons)*. Retrieved from https://rpubs.com/chelseyhill/676279
Otto, J., Sanford, D.A., & Ross, D.N. (2008). Does ratemyprofessor.com really rate my professor? *Assessment and Evaluation in Higher Education*, *33*(4), 355–368.
RateMyProfessors.com. (2021, October 26). In *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/RateMyProfessors.com
Rosen, A.S. (2018). Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: A large-scale study of RateMyProfessors.com data. *Assessment and Evaluation in Higher Education*, *43*(1), 31–44.
Rozin, P., & Royzman, E.B. (2001). Negativity Bias, Negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. doi: 10.1207/S15327957PSPR0504_2
Trussler, M., & Soroka, S. (2014). Consumer demand for cynical and negative news frames. *International Journal of Press/Politics*, *19*(3), 360–379. https://doi.org/10.1177/1940161214524832
Van der Meer, T., Hameleers, M., & Kroon. A. (2020). Crafting our own biased media diets: The effects of confirmation, source, and negativity bias on selective attendance to online news. *Mass Communication and Society*, *23*(6), 937–967. https://doi.org/10.1080/15205436.2020.1782432