Cutting Through the Comment Chaos

A Supervised Machine Learning Approach to Identifying Relevant YouTube Comments

Möller, A.M.; Vermeer, S.A.M.; Baumgartner, S.E.

DOI
10.1177/08944393231173895

Publication date
2024

Document Version
Final published version

Published in
Social Science Computer Review

Article

Social Science Computer Review
2024, Vol. 42(1) 162–185

DOI: 10.1177/08944393231173895
journals.sagepub.com/home/ssc
Sage

Cutting Through the Comment Chaos: A Supervised Machine Learning Approach to Identifying Relevant YouTube Comments

A. Marthe Möller[1], Susan A. M. Vermeer[1], and Susanne E. Baumgartner[1]

Abstract

Social scientists often study comments on YouTube to learn about people's attitudes towards and experiences of online videos. However, not all YouTube comments are relevant in the sense that they reflect individuals' thoughts about, or experiences of the content of a video or its artist/maker. Therefore, the present paper employs Supervised Machine Learning to automatically assess comments written in response to music videos in terms of their relevance. For those comments that are relevant, we also assess *why* they are relevant. Our results indicate that most YouTube comments are relevant (approx. 78%). Among those, most are relevant because they include a positive evaluation of the video, describe a viewer's personal experience related to the video, or express a sense of community among the video viewers. We conclude that Supervised Machine Learning is a suitable method to find those YouTube comments that are relevant to scholars studying viewers' reactions to online videos, and we present suggestions for scholars wanting to apply the same technique in their own projects.

Keywords

user comments, YouTube, relevance, Supervised Machine Learning, music videos

Introduction

People's daily use of social media has made interpersonal communication via such platforms a frequently studied topic among social scientists. A specific focus within this research is the study of online user comments written on social media and on YouTube in particular. It has been shown,

[1]Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, the Netherlands

Corresponding Author:
A. Marthe Möller, Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Nieuwe Achtergracht 166, Amsterdam 1018 WV, Netherlands.
Email: a.m.moller@uva.nl

for example, that people's attitudes towards and experiences of videos depend on the comments that accompany these videos (e.g., Hsueh et al., 2015; Möller et al., 2021; Shi et al., 2014; Waddell & Sundar, 2017; Walther et al., 2010; Ziegele et al., 2018). Other studies focused on how viewers use comments to converse about the content that they are watching (e.g., Dubovi & Tabak, 2020; Poché et al., 2017; Song et al., 2015).

One central challenge for researchers in this field is the sheer abundance of available social media comments. Research has shown that many of the available social media comments are not relevant to researchers or social media users as many comments might be spam created by bots, or just off-topic comments written by users (Poché et al., 2017). These irrelevant comments might require unnecessary processing and storage capacities, and can seriously bias study findings (Poché et al., 2017; Serbanoiu & Rebedea, 2013). It is thus indispensable for researchers studying social media comments to distinguish between relevant and irrelevant user comments. With "relevance" we refer to comments that reflect individuals' experiences of, or opinions or thoughts about (the content of) a video, or the artist/maker of the video. Moreover, we consider a comment as relevant if the comment relates to the social experience of watching a video on a social media platform (e.g., referring to other users, asking for information).

Understanding the extent to which YouTube comments reflect viewers' reactions to or experiences of videos is a necessary step if we want to advance our theoretical knowledge about online interpersonal communication on social media platforms. For example, various content analyses of YouTube comments detected the sentiment of comments to learn more about viewers' opinions and attitudes (e.g., Siersdorfer et al., 2010; Song et al., 2015; Thelwall, Sud, et al., 2012). However, if we assume that large amounts of comments are irrelevant and not related to the respective video, sentiment analysis might falsely categorize these comments as meaningful and might draw a biased picture of the public opinion on the video. Hence, the first aim of this study is to explore the occurrence of relevant comments on YouTube. Moreover, we set out to develop specific categories of relevant comments.

As the number of comments present on YouTube is seemingly infinite, it is impossible to manually analyze even a small fraction. Therefore, this study employs a Supervised Machine Learning (SML) approach. In short, SML refers to a technique whereby a computer automatically categorizes texts. For example, Vermeer et al. (2019) used it to analyze online reviews and to automatically determine if a review required a company to write a response or not. SML has made its entrance as a popular research method in social sciences relatively recently (Baden et al., 2022; Breuer et al., 2020; Van Atteveldt & Peng, 2018). In an article identifying social scientists' needs when it comes to the usage of computational techniques such as SML, Baden et al. (2022) have called for more resources and guidance in applying this relatively novel technique in their work. Replying to this call, the second aim of this study is to formulate a set of guidelines based on a detailed description of the steps taken in using SML to analyze YouTube comments that scholars can use for future research projects.

This study takes an exploratory approach by focusing on one popular genre of YouTube videos, namely, music videos. Music videos are the most popular form of videos on YouTube and trigger high user engagement (Liikkanen & Salovaara, 2015). In addition, by using SML, this study uses a method that is relatively novel within communication science in particular and for which the procedures are not as standardized as other more established methods in this field. Hence, this study explores how SML can best be used to analyze social media comments. By setting up guidelines for using SML to analyze YouTube comments, future scholars can use this method to investigate the relevance of YouTube comments written in response to videos other than music videos, such as news videos, product reviews, or vlogs, thereby moving beyond the exploratory nature of the current work.

The Importance of Studying Relevant User Comments on YouTube

User comments on social media provide a rich source of information about the opinions, sentiments, evaluations, and experiences of social media users (e.g., Thelwall, 2018). Particularly YouTube is characterized by its large number of public comments, and active user engagement in the comment section (Khan, 2017). YouTube comments have been previously studied to understand attitudes and opinions of users (Alhujaili & Yafooz, 2021), to examine user discourse about specific topics (e.g., Schneider, 2016), or to understand the influence of these comments on users (e.g., Möller et al., 2021; Walther et al., 2010). When conducting studies on user comments, it is indispensable to first identify and eliminate irrelevant comments as these might seriously bias study results.

From a social science perspective, the extent of irrelevant comments has not yet been studied. However, a few studies from computer science and related fields examined the extent of spam among user comments. For example, Dubovi and Tabak (2020) examined knowledge co-construction of science video comments. Of the 1530 comments in their manual content analysis, approximately 25% of comments were discarded as irrelevant (e.g., ads, spam). Similarly, Poché et al. (2017) automatically classified user comments of coding tutorials in respect to whether these comments are useful for the creator of the tutorials. In a first manual coding step, they found that only about 30% of the comments might provide useful information to the video creators, all other comments were coded as "miscellaneous" and "uninformative." However, in their classification, Poché et al. (2017) also coded evaluative comments (e.g., "This video is trash") as useless comments which might be considered useful in other contexts.

To further understand the degree of relevance of a comment, Serbanoiu and Rebedea (2013) proposed a relevance-based ranking of YouTube music video comments. In a first step, they removed all irrelevant comments using a neural network classifier. In this approach, textual features of each comment were used for the classification of relevancy. For example, it was assumed that spam comments have a higher number of capital letters, digits, and punctuation marks. By only using these textual features, they classified around 30% of comments as irrelevant. However, it is likely that such a solely text-based approach might erroneously classify relevant user comments as spam. For example, using the classification above, a comment such as "LOVE IT 4EVER!!!!!" might be classified as spam, although the comment does express a user's genuine expression about the music video which makes it relevant for social scientists.

In sum, whereas scholars have previously worked on relevance detection of YouTube comments, this has led to varying results because different definitions of relevance were used in these studies (see Online Appendix A for an overview of various definitions of relevance used in the literature). These definitions were either specific to the context of the videos that were studied (e.g., comments are relevant if they are useful for the creators of tutorial videos; Poché et al., 2017) or based on the textual features of comments which limits our understanding of why comments are relevant (i.e., Serbanoiu & Rebedea, 2013). To further advance research on the usage and effects of user comments, a thorough understanding of the relevance of comments is required that can be applied to different types of videos and that can capture the different reasons *why* a comment may be relevant for social scientists to study.

Towards a Thorough Understanding of the Relevance of Comments

Although the studies discussed above provide useful information on the detection of *irrelevant* comments on YouTube, they do not further examine *why* comments can be *relevant.* We thus lack an understanding of the types of topics that users discuss in relevant comments. In the current study, we therefore set out to develop an initial categorization of relevant comment categories.

To develop such a categorization, it is valuable to examine the motivations of users to post these comments. The Uses and Gratifications theory (Katz et al., 1973; Rubin, 2009) can be a useful framework for this. Specifically, it has been shown that motivations for active engagement, such as writing a comment, differ from those of passively using social media (Khan, 2017; Shao, 2009). For posting comments on YouTube, Khan (2017) identified three main motivations: information sharing, information seeking, and social interaction with other users. We assume that these three categories of motivations should also be reflected in the content of user comments.

Whereas the categories for information sharing and seeking are rather straightforward, the motivation to socially interact with others on YouTube can take various forms. A qualitative study by Schneider (2016) identified nostalgia—sentimental feelings towards one's own past experiences—as a central motive for sharing comments about music videos on YouTube. Similarly, Areni et al. (2022) established that sharing own experiences in YouTube comments of old TV commercials is a way for users to cherish one's own past experiences and thereby establishing connections with other viewers with similar past experiences. This can be seen as a form of social interaction as it might trigger other users to share their own experiences or relate to similar situations. Viewers of music videos on YouTube also frequently try to create a sense of community among viewers by directly reaching out to other viewers to share their experiences, or by referring to the moment in time someone is listening to a specific song (e.g., "Anyone listening to this song in 2022?").

Next to sharing or requesting information, describing viewers' personal experience related to the video, or creating a sense of community among the viewers of a video, YouTube comments also frequently have a strong evaluative character (Schultes et al., 2013). Although YouTube provides the option of assigning "likes" to videos, many viewers of entertainment videos additionally voice their positive or negative evaluation of a music video in the comment section. In sum, we argue that relevant comments likely belong to one of the following categories: positive video evaluation, negative video evaluation, providing information, requesting information, description of a viewers' personal experience related to the video, or creating a sense of community among the viewers of a video. We pose the following two research questions:

RQ1: To what extent are user comments written in response to online music videos posted on YouTube relevant?

RQ2: To what extent do relevant user comments belong to the sub-categories of positive evaluation, negative evaluation, providing information, requesting information, describing viewers' personal experience related to the video, or creating a sense of community among the viewers of a video?

Previous Research Based on Automated Content Analyses of User Comments

As discussed above, the increasing volume of user comments on YouTube has made it considerably more difficult for scholars to identify relevant comments. Ongoing advances in computational methods and natural language processing can help to tackle this challenge. This emerging discipline is often referred to as Computational Social Science (Lazer et al., 2009) or Computational Communication Science (Shah et al., 2015). Computational tools and methods offer new opportunities to identify and analyze the vast size of user comments on YouTube. Computational methods prove particularly fruitful when the dataset of interest is so large that it is not feasible to code each unit manually. There are different methodological approaches varying from counting and dictionaries to (un)supervised machine learning (for an overview see, for

example, Boumans & Trilling, 2016; Van Atteveldt et al., 2021). Computational methods can help, for instance, to confirm what scholars might already have suspected based on qualitative or small-scale quantitative research.

Various scholars are already applying computational methods to study user comments on YouTube. For instance, Ksiazek et al. (2015) selected 515 news videos and collected 28,341 user comments to study hostility and civility on YouTube. To assess civility and hostility in user comments, they used Linguistic Inquiry and Word Count (LIWC) to develop custom word dictionaries. For example, abuse*, annoy*, frustrat*, punish*, worthless*, are included in the hostility dictionary. More recently, Möller et al. (2019) have examined the amount and valence of online entertainment videos' social information and compared this to the social information of online political videos. They used the SentiStrength algorithm (Thelwall, Buckley, et al., 2012) to assess the valence of each comment. In a different study, Ji et al. (2022) collected 107 YouTube videos and their corresponding 56,912 comments about gene-edited babies. They used sentiment polarity, sentiment divergence, and the number of topics (using IBM Watson Natural Language Understanding), among other things, to study the factors that influence video popularity and diverse opinions.

Although dictionary-based approaches (or sentiment analysis tools that rely on dictionary-based approaches to determine the sentiment of posts based on lists of words with negative or positive valence) can be applied to detect the presence of certain words, machine learning techniques can handle more complex meaning compared to dictionary-based approaches. Vermeer et al. (2019) tested and compared the efficacy of various sentiment analysis, dictionary-based approaches, and machine learning techniques to detect relevant electronic word-of-mouth (eWOM). They used a dataset of approximately 60,000 Facebook user comments and 11,000 user tweets about 16 different brands to study eWOM in social media. Their results show that machine learning techniques—deciding whether eWOM is relevant for the brand to respond to, and distinguishing seven different types of eWOM (e.g., question, complaint)—can achieve considerably higher accuracy compared to any kind of sentiment analysis. More recently, Van Atteveldt et al. (2021) compared various methods of sentiment analysis. They relied on a dataset including economic and financial news from a total of 10 newspapers and five websites published between 1 February and 7 July 2015. The results indicate that machine learning, especially deep learning, substantially outperforms dictionary-based methods but falls short of human performance.

Unlike pre-trained dictionary-based approaches or sentiment tools that look for the valence of a text, a SML algorithm learns from human coders' classifications and allows scholars to solve the classification problem for an unlimited amount of user comments. While requiring manual labor initially, as training a SML algorithm requires an existing dataset of YouTube comments and their classification (e.g., relevant, or not), this approach can be highly useful for coding latent variables (e.g., social influences processes) in a large dataset. Moreover, as SML does not start with pre-existing assumptions, this approach is generally able to handle complex meaning. As such, using SML does not only increase efficiency, but also transparency and reproducibility (Boumans & Trilling, 2016). The classifiers can be used by other scholars to study YouTube comments over and over again. Various scholarly fields are already applying machine learning techniques to study user comments. For instance, in the field of journalism studies, Stoll et al. (2020) test and compare different SML techniques to predict impoliteness and incivility in online discussions on German news media outlets on Facebook. More recently, Jakob et al. (2022) combined LIWC dictionaries with machine learning to study destructive incivility in online user comments on news websites, Facebook, and Twitter.

Although the usage of computational methods, and machine learning in particular, is increasing within the field of social sciences in general, it is not often used by scholars taking a media

psychological perspective to study YouTube comments. This is striking as entertainment scholars studying the effects that YouTube comments may have on video viewers base their work on the assumption that the comments that video viewers see are actually relevant (e.g., Möller et al., 2021; Walther et al., 2010). Relying on SML, we demonstrate how such models can help filter relevant user comments on YouTube before proceeding with classifications that allow us to predict the underlying meaning of comments. In doing so, we pose the following research question:

RQ3: To what extent can SML help to detect relevant YouTube comments?

Method

Sample Selection and Data Collection

To answer our research questions, user comments posted in response to music videos on www.youtube.com were collected and analyzed. To determine the sample for this study, we first referred to YouTube's automatically updated list of the approximately 500 most viewed videos on the platform (YouTube, 2022). From this list, we selected only those videos that are music videos. Second, as we only wanted to analyze comments written in English, songs that are sung in English or songs without lyrics were selected. Third, to avoid including duplicate songs in our sample, we only selected music videos that were uploaded by the artist(s) performing in them (either directly or via the YouTube channel VEVO which is dedicated to uploading music videos on YouTube). Finally, as it is not possible to write a comment in response to YouTube videos aimed at children, we excluded all videos of nursery rhymes or children's songs. This resulted in a list of 309 videos.

To download the accompanying comments, we used a Python script developed by Van de Velde (2017) that employs the Google Application Programming Interface (API) to search for and collect data associated with YouTube videos. Using the unique ID-number that YouTube assigns to each video, we collected the data of the videos in three steps. First, we searched for the following meta-data belonging to each video: (1) the video title, (2) the day on which the video was uploaded on YouTube, (3) the name of the YouTube channel that uploaded the video, and (4) the video description as provided by the uploader. Second, we requested the API to provide us with the 600 most recent comments written in response to each video as well as the date on which each comment was uploaded. For the current project, we only focused on top-level comments, and we did not collect replies written in response to comments. After entering the 309 ID-numbers belonging to the selected videos, the API returned the data of 305 videos while the data of four videos were not returned by the API.

In the third and final step of the data collection, we collected the lyrics of the songs. The lyrics were collected as some comments on YouTube consist of the lyrics sung in the video that they accompany. Hence, adding the lyrics to the information used to categorize comments can clarify the link between a comment and its video that may otherwise remain unclear. The lyrics were gathered by scraping the Genius website (www.genius.com). Using Regular Expressions, we compiled a list of URLs leading to the webpage within Genius corresponding to each individual song based on the video title as provided by the Google API. We then stored the lyrics of each song performed in the 305 YouTube videos. Through this procedure, 182,367 comments were collected as well as the meta-data, and the lyrics of the video that these comments were written in response to. These data were collected between 22 April and 29 April 2022.

Applying SML requires a sample of comments that is annotated by human coders which can be used as the training material for the computer to understand how to automatically assign comments to categories. We estimated that with the time and budget available for the current project, it would not be possible to manually annotate all 182,367 comments that we collected.

Therefore, we selected a random sample of 40% of the downloaded comments (i.e., to 72,947 comments) to be coded manually.

Codebook Development

To create an annotated sample, we developed a codebook for the categorization of comments. The preliminary codebook asked coders to indicate whether a comment (1) is written in English, and (2) is relevant. Finally, if a comment was deemed relevant, coders were asked to indicate (3) to what category or categories it belonged (i.e., positive evaluation, negative evaluation, providing information, requesting information, personal experience, and/or community). If a comment was deemed relevant but did not belong to any of the six categories, coders were asked to assign it to a seventh category labeled "other." In four sessions, one of the authors together with two research assistants coded comments using the preliminary codebook. By coding comments and discussing them together, the codebook was adjusted and finalized: general instructions were refined, definitions of the various categories were clarified, and examples for each category were added (the final codebook is available on the Open Science Framework, see: https://osf.io/x8vgt/.

Data Annotation

Three research assistants went through extensive coder training consisting of seven training sessions and two practice rounds in which coders individually coded comments. After completion of the training, a more accurate estimation of how many comments each coder could annotate could be made. Based on this estimation, 30% of the comments that were designated to be part of the training dataset were selected randomly and each research assistant received their own set of comments to code. This led to a sample of 21,152 comments of which 2256 comments (approx. 10%) were coded by all three research assistants. The research assistants coded the data between 21 June and 25 August 2022 (see Table 1 for the categorization by the coders).

We assessed the intercoder reliability for all variables (see Table 2) in two ways. First, the percentage agreement score indicates the proportion of comments on which all coders agree (Hayes & Krippendorff, 2007). The lowest percentage agreement score for the annotated data is .88, with most scores over .90 indicating good reliability. We also report Krippendorff's Alpha or Kalpha, for which the lowest acceptable score is .667 (Krippendorff, 2004). As can be seen in Table 2, this threshold was met for all except three variables, namely, relevance (Kalpha = .64), providing information (Kalpha = .65), and other (Kalpha = .34). In case the coders did not agree on the score a comment should receive for a variable, the majority vote determined the final score.

Data Preprocessing

To prepare the data for the SML process, we did not modify the text of the comments (as is often done when using SML, see: Van Atteveldt et al., 2019), but instead fed it to the machines exactly how it was also presented to the coders because this offers the most direct comparison between the classification of our machines and that of the human coders. In addition, we added additional information to the text of the comments. We decided to include everything that the coders based their annotation on, namely, (1) the unique ID-number of the comment, (2) the date on which the comment was published, (3) the ID-number of the video that the comment was written in response to, (4) the title of the comment's video, (5) the date on which the comment's video was published, (6) the name of the channel that uploaded the comment's video, (7) the description of the comment's video, (8) the lyrics of the comment's video, and (9) the text of the comment.

Table 1. Distribution of Comments Across Categories in the Annotated Sample.[1]

	English	Relevant	Positive Evaluation	Negative Evaluation	Providing Information	Requesting Information	Personal Experience	Community	Other
# of comments	16,625	13,819	5,457	371	1,127	126	3,536	4,092	866
% of comments[2]	78.60	83.12	39.49	2.68	8.16	.91	25.59	29.61	6.27

[1]Relevant comments could be assigned to more than one category with the exception of relevant comments that did not fall under any of the six categories of relevant comments, in which case they were assigned to the "other" category only.

[2]For the English category, the given number indicates the percentage of all manually annotated comments. For the relevant category, the given number indicates the percentages of all manually annotated comments that were categorized as English. For the remaining categories, the given number indicates the percentage of all manually annotated comments that were categorized as English and relevant.

Table 2. Intercoder Reliability Score.

	English	Relevance	Positive Evaluation	Negative Evaluation	Providing Information	Requesting Information	Personal Experience	Community	Other
Percentage agreement	.97	.90	.89	.98	.93	.99	.88	.89	.92
Krippendorff's Alpha	.90	.64	.76	.72	.65	.71	.71	.75	.36

Finally, we compared two different types of vectorizers, a count vectorizer and a term frequency—inverse document frequency, or tf•idf vectorizer. Whereas a count vectorizer simply counts how often each word occurs, a tf•idf vectorizer weighs the number of term frequency by the number of documents in which the term occurs at least once (Van Atteveldt et al., 2022). We wanted to learn which vectorizer would lead to the best performance of the machine. Hence, we trained our machines twice, once using a count vectorizer and once using a tf•idf vectorizer.

Supervised Machine Learning

Using the Scikit Learn package in Python (https://scikit-learn.org/), we trained and tested several SML models. Frequently used models within social sciences that use SML are Multinomial Naïve Bayes (NB), Logistic Regression (LR), K-nearest neighbor (KNN), C4.5 decision tree, and Support Vector Machine (SVM) (e.g., Kaiser & Bodendorf, 2012; Meppelink et al., 2021; Van Zoonen & Van der Meer, 2016; Vermeer et al., 2019). We thus trained five machines, one for each of these models. This allowed us to compare machines with different underlying models and determine which machine performs best. Seeing that we also compared the performance of models receiving different vectors as input, we trained 10 machines in total (2: count vectorizer or tf•idf vecorizer × 5: Multinomial NB, LR, KNN, C4.5 decision tree, or SVM). Each machine can indicate the value of a comment for one category (i.e., one machine can indicate if a comment belongs to one specific category). Seeing that we have a total of nine variables or categories (i.e., one variable indicating if a comment is English, one variable indicating if a comment is relevant, and seven variables indicating why a comment is relevant), 90 models were trained in total. These models were trained using a randomly selected 80% of the manually annotated data (i.e., the training set). The remaining 20% of the manually annotated data was used as the test set. For the test dataset, the categorization of the SML models was compared to that of the coders so that the performance of the models could be assessed.

To evaluate the performance of SML models, we report various metrics, namely, accuracy, precision, recall, and the F_1-score. The accuracy score indicates in how many cases the classifier was correct in predicting a score. Precision indicates the number of positive category predictions made by the machine that indeed belong to that category. Recall quantifies the number of positive category predictions made by the machine out of all the cases that belonged to that category. Finally, the F_1-score is the harmonic mean of precision and recall. While we report each of these metrics for all our machines in the results section, we will predominantly use the F_1-score to evaluate our models.

Our choice to focus on the F_1-score is based on several factors. First, the accuracy score is not very informative in our case, as we are dealing with class imbalance, meaning that the distribution of comments across the different categories is not equal (Guo et al., 2008). For example, with 13,815 comments being labeled as relevant, irrelevant comments are relatively rare in our dataset (see Table 1). If a machine categorizes all comments as relevant, it will achieve a relatively high accuracy score simply because relevant comments often occur in the dataset, but this does not necessarily indicate that the machine is performing well. Second, when it comes to the precision and recall score of a machine, the answer to the question of which metric is most relevant depends on what the results of a machine will be used for. If it is important that a machine only makes correct predictions, precision seems to be the main indicator of a machine's performance. However, if a machine is being used to identify as many of the positive cases as possible, recall seems to be the best indicator of a machine's performance. Seeing that the development of the machine is, in itself, a goal of the present project, we decided to focus on the harmonic mean of precision and recall, or the F_1-score, when evaluating our models.

The values for a model's precision, recall, and F_1-score are applicable to each value of a category. For example, an F_1-score can be obtained for the comments that a machine classified as relevant and for those comments that the machine classified as not relevant. To assess the performance of a model, the average of both values is an informative metric. Specifically, we take the weighted average of the precision, recall and F_1-score. This number takes class imbalance into consideration by weighing the average of a measure by support (i.e., the number of true instances for each value).

Results

The first models that we built were trained to indicate whether a comment is written in English. The model based on a count vectorizer and on Logistic Regression outperformed the other models for all performance evaluators (see Table 3). The same can be seen for the machines trained to predict if comments can be assigned to the following categories: relevance, positive evaluation, personal experience, and community. Among the machines trained to predict if a comment contains a negative evaluation, the machine using a count vectorizer and a C4.5 decision tree outperformed the other models. For the machines trained to predict the variables indicating if a comment provided information, included a request for information, or belonged to the "other" category, multiple machines had the same and highest scores on the performance indicators (see Table 3). However, several of the best performing machines for these variables did not assign any comment to their respective category (see Table 3). Consequently, using these machines to classify new data would result in no comments being assigned to those categories. Therefore, we selected one machine among the machines that performed best (based on the F_1-score) that did assign comments to the category that they were trained to assess as the preferred model, namely, the machines using a count vectorizer and a C4.5 decision tree. To account for the risk of overfitting our models, we re-evaluated the best performing machines using a 5-fold cross validation. The results of this procedure (see Online Appendix B) corroborated our original findings.

We used the best performing machines to predict the scores on variables for those comments that were not annotated by the coders ($N = 161{,}215$ comments). We first used the machine based on a count vectorizer and LR and trained to indicate if a comment is English to distinguish between English and non-English comments. A total of 135,960 (approx. 84%) were automatically classified as English (see Table 4). Next, we used the best performing model (based on a count vectorizer and LR) that was trained to categorize comments as relevant or not relevant to analyze the sample of comments that were automatically classified as English. Of the English comments, 105,785 comments (approx. 78%) were automatically categorized as relevant (see Table 4). For each remaining variable, we then used those models that performed the best to analyze all English and relevant comments. The results of these automatic analyses are present in Table 4.

Comparing the results of the automated classifications (see Table 4) to those of the manual categorization of comments (see Table 1) shows that the similarity between the two results varies depending on the occurrence of comments within specific categories. Looking at the percentages of comments in the categories that occurred often (i.e., categories to which more than 10% of the manually annotated comments belonged), the automated classification yields results that are in line with those of the human coders (e.g., in the manually annotated sample, 39.49% of the comments could be categorized as positive evaluation, compared to 38.10% in the automatically classified sample). However, for categories that do not occur often in the manually annotated data (i.e., categories to which less than 10% of the manually annotated comments belonged), the results of the machines deviated from the results of the human coders. For example, while in the manually annotated sample, .91% of the English and relevant comments included a request for information, among the automatically classified comments, only .05% of the English and relevant comments

Table 3. Evaluation Metrics of the Classifiers.

Classifier		Accuracy	Precision	Recall	F_1-score
Trained on all annotated comments (n = 21,152)					
English					
Multinomial NB	Count vectorizer	.68	.75	.68	.70
	Tf•idf vectorizer	.78	.71	.78	.69
LR	Count vectorizer[1]	.88	.87	.88	.87
	Tf•idf vectorizer	.67	.76	.67	.70
KNN	Count vectorizer	.77	.71	.77	.72
	Tf•idf vectorizer	.79	.75	.79	.71
C4.5 decision tree	Count vectorizer	.85	.84	.85	.84
	Tf•idf vectorizer	.83	.82	.83	.82
SVM	Count vectorizer	.78	.70	.78	.70
	Tf•idf vectorizer	.78	.72	.78	.70
Trained on comments that were annotated as English (n = 16,625)					
Relevance					
Multinomial NB	Count vectorizer	.65	.74	.65	.68
	Tf•idf vectorizer[2]	.83	.68	.83	.75
LR	Count vectorizer	.81	.82	.81	.81
	Tf•idf vectorizer	.60	.76	.60	.65
KNN	Count vectorizer	.81	.76	.81	.77
	Tf•idf vectorizer	.81	.75	.81	.76
C4.5 decision tree	Count vectorizer	.81	.78	.81	.79
	Tf•idf vectorizer	.80	.78	.80	.79
SVM	Count vectorizer	.83	.86	.83	.75
	Tf•idf vectorizer	.83	.77	.83	.75
Positive evaluation					
Multinomial NB	Count vectorizer	.56	.61	.56	.58
	Tf•idf vectorizer[2]	.67	.45	.67	.54
LR	Count vectorizer	.79	.80	.79	.80
	Tf•idf vectorizer	.57	.62	.57	.58
KNN	Count vectorizer	.58	.61	.58	.59
	Tf•idf vectorizer	.55	.59	.55	.56
C4.5 decision tree	Count vectorizer	.69	.68	.69	.68
	Tf•idf vectorizer	.65	.67	.65	.66
SVM	Count vectorizer	.66	.57	.66	.56
	Tf•idf vectorizer	.67	.59	.67	.57
Negative evaluation					
Multinomial NB	Count vectorizer	0.89	0.96	0.89	0.92
	Tf•idf vectorizer[2]	0.98	0.96	0.98	0.97
LR	Count vectorizer	0.97	0.97	0.97	0.97
	Tf•idf vectorizer	0.67	0.96	0.67	0.79
KNN	Count vectorizer	0.98	0.96	0.98	0.97
	Tf•idf vectorizer[2]	0.98	0.96	0.98	0.97

(continued)

Table 3. (continued)

Classifier		Accuracy	Precision	Recall	F_1-score
C4.5 decision tree	Count vectorizer	0.98	0.97	0.98	0.97
	Tf•idf vectorizer	0.97	0.96	0.97	0.97
SVM	Count vectorizer[2]	0.98	0.96	0.98	0.97
	Tf•idf vectorizer[2]	0.98	0.96	0.98	0.97
Providing information					
Multinomial NB	Count vectorizer	.79	.89	.79	.83
	Tf•idf vectorizer[2]	.93	.87	.93	.90
LR	Count vectorizer[1]	.89	.89	.89	.89
	Tf•idf vectorizer	.65	.90	.65	.74
KNN	Count vectorizer	.93	.89	.93	.90
	Tf•idf vectorizer	.93	.89	.93	.90
C4.5 decision tree	Count vectorizer	.93	.90	.93	.91
	Tf•idf vectorizer	.92	.90	.92	.91
SVM	Count vectorizer	.93	.92	.93	.91
	Tf•idf vectorizer	.93	.92	.93	.91
Requesting information					
Multinomial NB	Count vectorizer	.97	.99	.97	.98
	Tf•idf vectorizer[2]	.99	.99	.99	.99
LR	Count vectorizer	.98	.99	.98	.98
	Tf•idf vectorizer	.73	.98	.73	.84
KNN	Count vectorizer[2]	.99	.99	.99	.99
	Tf•idf vectorizer[2]	.99	.99	.99	.99
C4.5 decision tree	Count vectorizer	.99	.99	.99	.99
	Tf•idf vectorizer	.99	.99	.99	.99
SVM	Count vectorizer[2]	.99	.99	.99	.99
	Tf•idf vectorizer[2]	.99	.99	.99	.99
Personal experience					
Multinomial NB	Count vectorizer	.59	.72	.59	.63
	Tf•idf vectorizer[2]	.80	.63	.80	.71
LR	Count vectorizer	.85	.85	.85	.85
	Tf•idf vectorizer	.61	.74	.61	.64
KNN	Count vectorizer	.80	.76	.80	.73
	Tf•idf vectorizer	.79	.72	.79	.72
C4.5 decision tree	Count vectorizer	.80	.78	.80	.79
	Tf•idf vectorizer	.80	.78	.80	.78
SVM	Count vectorizer	.80	.74	.80	.71
	Tf•idf vectorizer	.80	.74	.80	.71
Community					
Multinomial NB	Count vectorizer	.59	.68	.59	.62
	Tf•idf vectorizer	.76	.69	.76	.66
LR	Count vectorizer	.85	.85	.85	.85
	Tf•idf vectorizer	.59	.69	.59	.62

(continued)

Table 3. (continued)

Classifier		Accuracy	Precision	Recall	F_1-score
KNN	Count vectorizer	.76	.73	.76	.74
	Tf•idf vectorizer	.69	.67	.69	.68
C4.5 decision tree	Count vectorizer	.84	.83	.84	.83
	Tf•idf vectorizer	.79	.79	.79	.79
SVM	Count vectorizer	.76	.74	.76	.67
	Tf•idf vectorizer	.76	.71	.76	.68
Other					
Multinomial NB	Count vectorizer	.79	.90	.79	.84
	Tf•idf vectorizer[2]	.94	.89	.94	.91
LR	Count vectorizer	.87	.90	.87	.89
	Tf•idf vectorizer	.61	.90	.61	.71
KNN	Count vectorizer	.89	.89	.89	.89
	Tf•idf vectorizer	.91	.89	.91	.90
C4.5 decision tree	Count vectorizer	.93	.89	.93	.91
	Tf•idf vectorizer	.89	.90	.89	.89
SVM	Count vectorizer[2]	.94	.89	.94	.91
	Tf•idf vectorizer[2]	.94	.89	.94	.91

[1]Training this machine produced a convergence warning. Therefore, the machine was trained again with scaled data. Reported results are based on the machine trained on this scaled data.
[2]This machine predicted none of the comments in the test dataset to belong to the assessed category, making the precision and F_1-score ill-defined.
Table 3 includes the following abbreviations: Multinomial NB (Multinomial Naïve Bayes), LR (Logistic Regression), KNN (K-nearest neighbor), SVM (Support Vector Machine), and tf•idf vectorizer (term frequency – inverse document frequency vectorizer).

included a request for information. Thus, whereas the results of the automated classification resemble those of the manual annotation when it comes to categories to which many comments belong, this similarity decreases as fewer comments belong to a category.

To further advance our understanding of the SML models, we applied the Python library eli5 (https://eli5.readthedocs.io/en/latest/index.html) to indicate which features (i.e., elements of comments and their meta-data) made it more likely that a comment was assigned to a specific category. Table 5 indicates, for each category, which 10 features made it more likely that a comment was assigned to that category by the computer. For the positive evaluation and negative evaluation categories, the results seem rather straightforward. For example, containing words such as "masterpiece," "legendary," and "banger" made it more likely that the model assigned a comment to the positive evaluation category. In addition, containing words such as "ruined," "cringe," and "terrible" made it more likely that a comment was categorized as containing a negative evaluation. Two other categories for which the results seem straight forward are the personal experience category, to which comments containing words such as "nostalgia," "nostalgic," and "bored" were likely to be assigned, and the community category, to which comments containing words such as "billion," "2022," and "everyone" were likely to be assigned.

The results for the three remaining categories (i.e., providing information, requesting information, and other), however, warrant more elaboration. As can be seen in Table 5, both the providing information and the other categories consist of comments that not only contain specific

Table 4. Distribution of Comments Across Categories in the Automatically Categorized Sample (n = 161,215).[1]

	English	Relevant	Positive Evaluation	Negative Evaluation	Providing Information	Requesting Information	Personal Experience	Community	Other
# of comments	135,960	105,785	40,301	467	3288	53	26,084	29,451	670
% of comments[2]	84.33	77.81	38.10	0.44	3.11	0.05	24.66	27.84	0.63

[1]Relevant comments could be assigned to more than one category with the exception of relevant comments that did not fall under any of the six categories of relevant comments, in which case they were assigned to the "other" category only.

[2]For the English category, the given number indicates the percentage of all automatically categorized comments. For the relevant category, the given number indicates the percentages of all automatically categorized comments that were classified as English. For the other categories, the given number indicates the percentage of all automatically categorized comments that were classified as English and relevant.

Table 5. Ten Most Characteristic Features for Each Category of Comments and Their Weights.

Relevance	Positive Evaluation	Negative Evaluation	Providing Information	Requesting Information	Personal Experience	Community	Other
2022 (2.60)	masterpiece (5.20)	ruined (0.02)	amp (0.05)	0ksoma3qbu0[1] (0.02)	nostalgia (5.06)	billion (4.22)	2022 (0.01)
masterpiece (2.60)	legendary (3.35)	cringe (0.01)	br (0.05)	Comment ID[2] (0.01)	nostalgic (4.37)	2022 (4.03)	sus (0.005)
song (2.42)	banger (2.87)	terrible (0.01)	href (0.01)	Comment ID[2] (0.01)	bored (2.55)	2021 (3.49)	song (0.005)
2021 (2.40)	awesome (2.76)	for (0.01)	that (0.01)	Comment ID[2] (0.01)	memories (2.52)	1b (3.41)	11t20 (0.004)
billion (2.13)	favorite (2.72)	so (0.01)	ooh (0.01)	Comment ID[2] (0.01)	reminds (2.47)	everyone (2.96)	ugyt (0.003)
video (1.95)	nice (2.70)	ridiculous (0.01)	39 (0.01)	Comment ID[2] (0.01)	miss (2.45)	hi (2.81)	boi (0.003)
rihanna (1.86)	favorite (2.68)	rapping (0.01)	you (0.01)	Comment ID[2] (0.01)	childhood (2.44)	anyone (2.61)	js (0.003)
nostalgia (1.85)	fav (2.66)	of (0.01)	real (0.005)	Comment ID[2] (0.01)	brought (2.28)	views (2.60)	info (0.003)
href (1.80)	best (2.52)	please (0.01)	don (0.005)	Comment ID[2] (0.01)	crying (2.22)	who (2.35)	lovato (0.003)
views (1.74)	amazing (2.39)	intros (0.01)	me (0.004)	Comment ID[2] (0.01)	concert (2.20)	2b (2.32)	yas (0.003)

[1]The provided information represents a video ID-number created by YouTube.
[2]To ensure the privacy of the authors of specific comments, the original comment ID-number was removed.

words (e.g., "that" for providing information and "song" for other), but also of numbers and meaningless words (e.g., "sus" for other) as well as parts of the time stamp of comments (e.g., "11t01"). These are not features that the coders typically looked at when manually annotating the data (see the Open Science Framework for the codebook and instructions that coders used: https://osf.io/x8vgt/). In addition, for the requesting information category, Table 5 shows that specific comments were more likely to be assigned to this category than others, but no general features such as words were used to determine if a comment contained a request for information. Finally, for all three categories, the weights that the models assigned to particular features are relatively close to zero, meaning that even the top 10 defining features of each category are not well able to mark comments that belong to that category.

Discussion

The present paper investigated to what extent comments written in response to music videos on YouTube are relevant in the sense that they reflect viewers' experiences of, or opinions or thoughts about (the content of) a video, or its artist/maker. By gathering and analyzing 182,367 YouTube comments and their meta-data using SML, we found that between 78% and 83% of English comments is relevant and can be used by scholars to learn more about viewers' experiences of and reaction to music videos. This large number of relevant comments is striking compared to previous work where the proportion of irrelevant comments ranged from 25% to 70% of the collected samples (Dubovi & Tabak, 2020; Poché et al., 2017; Serbanoiu & Rebedea, 2013). The difference between our finding and that of previous studies is likely due to the difference in definition of relevance: whereas we considered a rather broad definition of relevance, previous studies assessed relevance by examining the textual properties of comments (e.g, Serbanoiu & Rebedea, 2013), or adopted a definition of relevance that is applicable to the specific type of videos included in those studies only (e.g., Poché et al., 2017). This leads us to conclude that for researchers studying comments and reviewing past studies on user comments, it is crucial to understand which definitions of relevance previous studies used as this definition impacts what comments are included in their analyses.

Another reason for the difference in proportion of irrelevant comments between our study and previous work could be the development of the YouTube algorithm over time—which automatically detects spam based on the text of a comment, or by the behavior from a particular commenter. The improvement in YouTube algorithms for detecting spam might lead to larger filtering of irrelevant comments. In line with this idea, the number of comments that were removed by YouTube in 2022 due to them not complying with YouTube's guidelines was higher compared to that number in 2018 (Ceci, 2023). However, additional research comparing the proportion of irrelevant comments that are available on YouTube over time is necessary to confirm this notion.

Our findings also revealed that most comments could be assigned to at least one of six content categories. The three most commonly occurring categories are comments that contain a positive evaluation, describe a personal experience of the commenter, or are written to create a sense of community. The tone of the comments for music videos was therefore positive and supporting. On the one hand, this finding contradicts previous work on (YouTube) comments in which scholars focused on the potential of YouTube and other social media platforms to facilitate the publication of comments containing hostility and insults (e.g., Döring & Mohseni, 2020; Murthy & Sharma, 2019). On the other hand, our finding is in line with previous research based on the Uses and Gratifications framework (Katz et al., 1973; Rubin, 2009) finding that posting comments on YouTube is often motivated by a need for social interaction (Khan, 2017). Specifically, our findings suggest that individuals who post comments on YouTube do so because they are looking

for positive social interactions with others by sharing compliments on other people's work, and by sharing personal experiences and connecting to the community of video viewers.

Based on the good performance of our models, it seems that SML is a good method for classifying comments into categories that occur relatively frequently. This conclusion is based on three results, namely, the high scores of the machines on the performance metrics, the comparison of the results of the manual annotation and of the automated classification of the data, and the features that the models used to categorize comments. Regarding the first finding, our results show that even the machines with the lowest scores on the performance metrics still performed relatively well. Regarding the second finding, we see that the results of the automated classification resembled those of the manual annotation but only for the automated classification of comments to categories to which many comments were assigned in the manual annotation. Hence, it seems that the classifiers are able to classify comments when it comes to frequently occurring categories, whereas they do not perform as well when it comes to categories that are rarer.

The above may be explained based on two limitations of the current work. First, the distribution of comments across categories was rather skewed. Hence, for categories to which fewer comments were assigned, the machines received fewer examples of comments that are relevant to those categories and thus, fewer examples to learn from. This notion seems to be largely in line with our third finding regarding the comment features that the models used to categorize comments. Namely, for categories to which many comments were assigned, the computer used distinctive words to indicate if a comment was likely to belong to a category (e.g., "masterpiece" as a word that makes it likely that a comment contains a positive evaluation). However, for categories to which few comments were assigned based on the manual annotation, the computer did not always use words with a clear meaning to indicate if a comment was likely to belong to a category (e.g., mentioning a specific comment ID makes it more likely that a comment contains a request for information). In these cases, the computer failed to identify meaningful features to categorize comments. Future studies could solve this by ensuring that their training data contains sufficiently large number of comments for each category, for example, by collecting and coding more data.

A second limitation of the present study is based on the low intercoder reliability of some variables. Although the intercoder reliability for most of our variables was satisfactory, this was not the case for three variables, namely, for the variables indicating if a comment was relevant, provided information, or if a comment fell in the "other" category. Two of these variables, namely, providing information and "other" are also among the variables for which the results of the automated classification deviate from those of the manual annotation (i.e., relevance, providing information, "other," and requesting information). Hence, for these two variables, the low intercoder reliability may have had consequences for how well the machines were able to learn to detect what we actually aimed to detect when annotating comments (Van Atteveldt et al., 2019). To address this limitation, future research could include additional rounds of coder training to ensure that the intercoder reliability of all variables is sufficient.

Interestingly, although the intercoder reliability of the variable indicating if a comment contained a request for information was sufficient, we did find that the results of the automated classification of comments to this category (i.e., .05% of the English and relevant comments belonged to the category) deviated from the manual categorization of comments to this category (i.e., .91% of the English and relevant comments belonged to the category, which is more than 18 times as much compared to the automated classification of comments). This indicates that while it may have been relatively straightforward for humans to understand when a comment contains a request for information even though such comments occurred seldomly, this task proved to be more difficult for the computer. Our investigation of what specific comment features increased the likelihood of comments being assigned to the request information category by the computer corroborates this notion. Here, we found that the 10 features making it more likely that a comment

was assigned to this category were unique comment ID-numbers. In other words, specific comments were more likely to be assigned to the requesting information category than others. This implies that the model was not able to find words that comments requesting information typically contain. Hence, whereas for humans it is evident when a comment contains a request for information, this is not the case for our model. This shows that although higher intercoder reliability scores for variables of the manually annotated data may make it more likely that a machine can learn to detect what researchers aim to detect, it offers no guarantee for this.

A final limitation of our study pertains to our sample not being representative for all YouTube comments in general. This is due to two reasons. First, we only collected comments written in response to music videos while on YouTube, many different videos are available, such as news videos, instruction videos, or unpacking videos. Second, we only collected the top-level comments for each video, meaning that replies to comments were not gathered or analyzed. Because of the above, generalizing our findings to comments written in response to other types of videos and to replies to comments warrants caution. Moreover, it means that the machines that we trained are only suitable to analyze top-level comments written in response to music videos on YouTube, but not to analyze (replies to) comments written in response to other types of videos (Baden et al., 2022; Burscher et al., 2015; Van Atteveldt & Peng, 2018). Hence, scholars who want to analyze the relevance of comments from a different population than the one used in the present study need to repeat the process of manually annotating comments and training machines based on these data.

Based on the present project, we formulate several recommendations to scholars who aim to use SML to analyze comments or other textual social media data. First, there are many common practices using SML to analyze (social) media data, but to increase the transparency of the literature, it is important that scholars explain why they choose to follow a specific practice. For example, while it is common to remove stop words before feeding text as training data to a machine, the choice to do so is often not explained (Baden et al., 2022) and based on our results, it does not always seem necessary. In addition, our findings indicate that the specific vectorizer used has implications for the performance of machines. Hence, if researchers choose to use one preselected vectorizer in SML instead of comparing multiple vectorizers, their choice for that particular vectorizer should be clearly motivated. Explaining why specific steps are (not) taken in SML makes projects clearer and more transparent, and it allows future scholars to make informed decisions when they use SML in their next projects.

Second, although scholars often predominantly focus on the evaluation of machines, we argue that another focus that is at least as important is the manual annotation of the data. This is because the quality of the classification made by a machine can only be as good as that of the annotations that they were trained with. Third, as discussed above, our machines seem to perform well when it comes to assigning comments to frequently occurring categories, but they perform less well when it comes to assigning comments to categories that are relatively rare. Hence, it seems that setting up categories that can be assumed to cover many cases rather than creating categories to which only a few cases apply will contribute to creating well performing machines. The trade-off here, however, is that setting up categories that cover many cases may also be theoretically less informative as they may be more general. To future scholars, we thus recommend considering the question of how broad the categories that they want to classify can be for them to still be theoretically meaningful in their decision of whether to use SML.

The present project applied SML to select those YouTube comments that are relevant for social scientists. This resulted into new insights into the relevance of comments on YouTube as well as guidelines for scholars who want to use SML in the future. As interpersonal communication is often used as a theoretical background in computational studies (Van Atteveldt & Peng, 2018), the present work can be used as an additional resource for scholars

working in this field. However, the findings of this project also give rise to new questions. First, as this project focused on comments written in response to music videos, it remains unclear how different contextual genres affect the performance of SML algorithms (Baden et al., 2022; Burscher et al., 2015; Van Atteveldt & Peng, 2018). Various types of videos may evoke different reactions among viewers (Möller et al., 2019) that may be harder or easier to classify. Only by validating machines using samples from different populations can this question be answered. Second, while we focused on detecting relevant comments in an entertainment context, the question of how relevant social media comments are is important to scholars working in other fields as well. For example, this method might help political communication scholars to analyze the spread of misinformation on YouTube. It might also be relevant for health communication scholars to understand the power of YouTube comments to hinder important health discussions. Hence, we call for more collaboration between scholars from different subdisciplines as the algorithms developed for one discipline can offer substantial contributions to other disciplines as well. Moreover, as Brady (2019) and Van Atteveldt and Peng (2018) point out, with the introduction of computational methods in the field of social science, the requirements for scholars to develop programming skills also rise. Our field can only make optimal use of the opportunities that computational methods offer through collaboration—we aim for the current work to be a contribution to that.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Digital Communication Methods Lab (University of Amsterdam).

ORCID iDs

A. Marthe Möller https://orcid.org/0000-0002-2106-1422
Susan Vermeer https://orcid.org/0000-0002-9829-8057
Susanne Baumgartner https://orcid.org/0000-0001-6031-8836

Supplemental Material

Supplemental material for this article is available online.

References

Alhujaili, R. F., & Yafooz, W. M. S. (2021). March 25). Sentiment analysis for YouTube videos with user comments: Review. 2021 international Conference on artificial Intelligence and smart systems (ICAIS). https://ieeexplore.ieee.org/document/9396049

Areni, C. S., Momeni, M., & Reynolds, N. (2022). Ontological insecurity, nostalgia, and social media: Viewing YouTube videos of old TV commercials re-establishes continuity of the self over time. *Psychology of Popular Media*, *11*(2), 227–236. https://doi.org/10.1037/ppm0000352

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2022). Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, *16*(1), 1–18. https://doi.org/10.1080/19312458.2021.2015574

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*(1), 8–23. https://doi.org/10.1080/21670811.2015.1096598

Brady, H. E. (2019). The challenge of big data and data science. *Annual Review of Political Science*, *22*(1), 297–323. https://doi.org/10.1146/annurev-polisci-090216-023229

Breuer, J., Wulf, T., & Mohseni, M. R. (2020). New formats, new methods: Computational approaches as a way forward for media entertainment research. *Media and Communication*, *8*(3), 147–152. https://doi.org/10.17645/mac.v8i3.3530

Burscher, B., Vliegenthart, R., & De Vreese, C. (2015). Using supervised machine learning to code policy issues: Can classifiers generalize across contexts? *The Annals of the American Academy of Political and Social Science*, *659*(1), 122–131. https://doi.org/10.1177/0002716215569441

Ceci, L. (2023), January 4). "Number of video comments removed from YouTube worldwide as of Q3 2022". Statista. https://www.statista.com/statistics/1132989/number-removed-youtube-video-comments-worldwide/

Döring, N., & Mohseni, M. R. (2020). Gendered hate speech in YouTube and YouNow comments: Results of two content analyses. *Studies in Communication and Media*, *9*(1), 62–88. https://doi.org/10.5771/2192-4007-2020-1-62

Dubovi, I., & Tabak, I. (2020). An empirical analysis of knowledge co-construction in YouTube comments. *Computers & Education*, *156*, 103939. https://doi.org/10.1016/j/compedu.2020.103939

Guo, X., Yilong, Y., Cailing, D., Yang, G., & Zhou, G. (2008). On the class imbalance problem. *Proceedings of the 2008 fourth International Conference on Natural Computing*. DOI:10.1109/ICNC.2008.871.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89. https://doi.org/10.1080/19312450709336664

Hsueh, M., Yogeeswaran, K., & Malinen, S. (2015). "Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors? *Human Communication Research*, *41*(4), 557–576. https://doi.org/10.1111/hcre.12059

Jakob, J., Dobbrick, T., Freudenthaler, R., Haffner, P., & Wessler, H. (2022). *Is constructive engagement online a lost cause? Toxic outrage in online user comments across democratic political systems and discussion arenas*. Communication Research. https://doi.org/10.1177/00936502211062773

Ji, J., Hu, H., & Wei, S. (2022). YouTube comments on gene-edited babies: What factors affect diverse opinions in comments? *Social Science Computer Review*. https://doi.org/10.1177/08944393211073164

Kaiser, C., & Bodendorf, F. (2012). Mining consumer dialog in online forums. *Internet Research*, *22*(3), 275–297. https://doi.org/10.1108/10662241211235653

Katz, E., Blumler, J. G., & Gurevitch, M. (1973). Uses and gratifications research. *Public Opinion Quarterly*, *37*(4), 509–523. https://doi.org/10.1086/268109

Khan, M. L. (2017). Social media engagement: What motivates user participation and consumption on YouTube? *Computers in Human Behavior*, *66*, 236–247. https://doi.org/10.1016/j.chb.2016.09.024

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Sage.

Ksiazek, T. B., Peer, L., & Zivic, A. (2015). Discussing the news: Civility and hostility in user comments. *Digital Journalism*, *3*(6), 850–870. https://doi.org/10.1080/21670811.2014.972079

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Social science. computational social science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742

Liikkanen, L. A., & Salovaara, A. (2015). Music on YouTube: User engagement with traditional, user-appropriated and derivative videos. *Computers in Human Behavior*, *50*, 108–124. https://doi.org/10.1016/j.chb.2015.01.067

Meppelink, C. S., Hendriks, H., Trilling, D., Van Weert, J. C. M., Shao, A., & Smit, E. S. (2021). Reliable or not? An automated classification of webpages about early childhood vaccination using supervised

machine learning. *Patient Education and Counseling*, *104*(6), 1460–1466. https://doi.org/10.1016/j.pec.2020.11.013

Möller, A. M., Baumgartner, S. E., Kühne, R., & Peter, J. (2021). Sharing the fun? How social information affects viewers' video enjoyment and video evaluations. *Human Communication Research*, *47*(1), 25–48. https://doi.org/10.1093/hcr/hqaa013

Möller, A. M., Kühne, R., Baumgartner, S. E., & Peter, J. (2019). Exploring user responses to entertainment and political videos: An automated content analysis of YouTube. *Social Science Computer Review*, *37*(4), 510–528. https://doi.org/10.1177/0894439318779336

Murthy, D., & Sharma, S. (2019). Visualizing YouTube's comment space: Online hostility as a networked phenomena. *New Media & Society*, *21*(1), 191–213. https://doi.org/10.1177/1461444818792393

Poché, E., Jha, N., Williams, G., Staten, J., Vesper, M., & Mahmoud, A. (2017). Analyzing user comments on YouTube coding tutorial videos. *2017 IEEE/ACM 25th international Conference on program comprehension (ICPC)*, 196–206. https://doi.org/10.1109/ICPC.2017.26

Rubin, A. M. (2009). Uses-and-gratifications perspective on media effects. In *Media effects: Advances in theory and research* (3rd ed.). Routledge.

Schneider, C. J. (2016). Music videos on YouTube: Exploring participatory culture on social media. In *Symbolic Interactionist Takes on Music*(*47*, pp. 97–117). Emerald Group Publishing Limited. https://doi.org/10.1108/S0163-239620160000047016

Schultes, P., Dorner, V., & Lehner, F. (2013). Leave a comment! An in-depth analysis of user comments on YouTube. *Wirtschaftsinformatik Proceedings*, *42*, 15.

Serbanoiu, A., & Rebedea, T. (2013). Relevance-based ranking of video comments on YouTube. *2013 19th International Conference on Control Systems and Computer Science*, 225–231. https://doi.org/10.1109/CSCS.2013.87

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science. *The Annals of the American Academy of Political and Social Science*, *659*(1), 6–13. https://doi.org/10.1177/0002716215572084

Shao, G. (2009). Understanding the appeal of user-generated media: A uses and gratification perspective. *Internet Research*, *19*(1), 7–25. https://doi.org/10.1108/10662240910927795

Shi, R., Messaris, P., & Cappella, J. N. (2014). Effects of online comments on smokers' perception of antismoking Public Service Announcements. *Journal of computer-mediated communication: JCMC*, *19*(4), 975–990. https://doi.org/10.1111/jcc4.12057

Siersdorfer, S., Chelaru, S., Nejdl, W., & San Pedro, J. (2010). How useful are your comments? Analyzing and predicting YouTube comments and comment ratings. *Proceedings of the 19th international conference on world wide web* (p. 891). WWW '10. https://doi.org/10.1145/1772690.1772781

Song, M., Jeong, Y. K., & Kim, H. J. (2015). Identifying the topology of the K-pop video community on YouTube: A combined co-comment analysis approach. *Journal of the Association for Information Science and Technology*, *66*(12), 2580–2595. https://doi.org/10.1002/asi.23346

Stoll, A., Ziegele, M., & Quiring, O. (2020). Detecting impoliteness and incivility in online discussions: Classification approaches for German user comments. *Computational Communication Research*, *2*(1), 109–134. https://doi.org/10.5117/CCR2020.1.005.KATH

Thelwall, M. (2018). Social media analytics for YouTube comments: Potential and limitations. *International Journal of Social Research Methodology*, *21*(3), 303–316. https://doi.org/10.1080/13645579.2017.1381821

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, *63*(1), 163–173. https://doi.org/10.1002/asi.21662

Thelwall, M., Sud, P., & Vis, F. (2012). Commenting on YouTube videos: From Guatemalan rock to el big bang. *Journal of the American Society for Information Science and Technology*, *63*(3), 616–629. https://doi.org/10.1002/asi.21679

Van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2–3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

Van Atteveldt, W., Trilling, D., & Arcila Calderon, C. (2022). *Computational analysis of communication*. Wiley-Blackwell.

Van Atteveldt, W., Van der Velden, M. A. C. G., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, *15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

Van Atteveldt, W., Welbers, K., & van der Velden, M. (2019). Studying political decision making with automatic text analysis. In W. van Atteveldt, K. Welbers, & M. Van der Velden (Eds.), *Oxford research encyclopedia of politics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228637.013.957

Van de Velde, R. N. (2017). *YouTube scraper*. Retrieved from https://github.com/bobvdvelde/YouTube_scraper/tree/master

Van Zoonen, W., & Van der Meer, T.G. L. A. (2016). Social media research: The application of supervised machine learning in organizational communication research. *Computers in Human Behavior*, *63*, 132–141. https://doi.org/10.1016/j.chb.2016.05.028

Vermeer, S. A. M., Araujo, T., Bernritter, S. F., & van Noort, G. (2019). Seeing the wood for the trees: How machine learning can help firms in identifying relevant electronic word-of-mouth in social media. *International Journal of Research in Marketing*, *36*(3), 492–508. https://doi.org/10.1016/j.ijresmar.2019.01.010

Waddell, T. F., & Sundar, S. S. (2017). #thisshowsucks! The overpowering influence of negative social media comments on television viewers. *Journal of Broadcasting & Electronic Media*, *61*(2), 393–409. https://doi.org/10.1080/08838151.2017.1309414

Walther, J. B., DeAndrea, D., Kim, J., & Anthony, J. C. (2010). The influence of online comments on perceptions of antimarijuana Public Service Announcements on YouTube. *Human Communication Research*, *36*(4), 469–492. https://doi.org/10.1111/j.1468-2958.2010.01384.x

YouTube (2022). Most viewed videos of all time (over 450M views) https://www.youtube.com/playlist?list=PLirAqAtl_h2r5g8xGajEwdXd3x1sZh8hC

Ziegele, M., Koehler, C., & Weber, M. (2018). Socially destructive? Effects of negative and hateful user comments on readers' donation behavior toward refugees and homeless persons. *Journal of Broadcasting & Electronic Media*, *62*(4), 636–653. https://doi.org/10.1080/08838151.2018.1532430

Author Biographies

Dr. A. Marthe Möller is an Assistant Professor of Entertainment Communication at the Amsterdam School of Communication Research (ASCoR). In her work, she analyzes user-generated information to study social media users' responses to and online conversations about media messages. She does so through experiments and by employing computational methods for automated content analysis.

Dr. Susan Vermeer is an Assistant Professor at the Amsterdam School of Communication Research (ASCoR) of Political Communication and Journalism. In 2021, she defended her dissertation, in which she investigated news consumption in the digital society. Her research interests center around the impact of the digital media environment, such as social media and algorithms, on political behaviour and attitudes. She combines traditional research methods with innovative and computational methods to study such processes.

Dr. Susanne Baumgartner is an Associate Professor at the Amsterdam School of Communication Research (ASCoR), at the University of Amsterdam. Her research focuses on the role of digital media in adolescent development. Specifically, she is interested in media multitasking and how this affects the cognitive development of youth. Moreover, she studies the effects of digital media on stress and sleep. To study the impact of digital media, she employs innovative methodological approaches, such as experience sampling, smartphone tracking data, and eye-tracking.