International Journal of Research in Marketing 40 (2023) 54–74

Contents lists available at ScienceDirect

International Journal of Research in Marketing

journal homepage: www.elsevier.com/locate/ijresmar

Full length article

Complementing human effort in online reviews: A deep learning approach to automatic content generation and review synthesis

Keith Carlson [a], Praveen K. Kopalle [b,*], Allen Riddell [c], Daniel Rockmore [d,e], Prasad Vana [b]

[a] Department of Computer Science, Dartmouth College, Hanover, NH 03755, United States
[b] 100 Tuck Hall, Tuck School of Business at Dartmouth, Dartmouth College, Hanover, NH 03755, United States
[c] School of Informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, IN 47405, United States
[d] Departments of Computer Science and Department of Mathematics, The Santa Fe Institute, Santa Fe, NM 87501, United States
[e] Dartmouth College, Hanover, NH 03755, United States

ARTICLE INFO

Article history:
First received on February 5, 2021 and was under review for 6½ months
Available online 12 February 2022

Area Editor: Lan Luo
Accepting Editor: P.K. Kannan

Keywords:
Online reviews
Artificial intelligence
Machine learning
Wine reviews
Review synthesis
Automation
Deep learning

ABSTRACT

Online product reviews are ubiquitous and helpful sources of information available to consumers for making purchase decisions. Consumers rely on both the quantitative aspects of reviews such as valence and volume as well as textual descriptions to learn about product quality and fit. In this paper we show how new achievements in natural language processing can provide an important assist for different kinds of review-related writing tasks. Working in the interesting context of wine reviews, we demonstrate that machines are capable of performing the critical marketing task of writing expert reviews directly from a fairly small amount of product attribute data (metadata). We conduct a kind of "Turing Test" to evaluate human response to our machine-written reviews and show strong support for the assertion that machines can write reviews that are indistinguishable from those written by experts. Rather than replacing the human review writer, we envision a workflow wherein machines take the metadata as inputs and generate a human readable review as a first draft of the review and thereby assist an expert reviewer in writing their review. We next modify and apply our machine-writing technology to show how machines can be used to write a synthesis of a set of product reviews. For this last application we work in the context of beer reviews (for which there is a large set of available reviews for each of a large number of products) and produce machine-written review syntheses that do a good job – measured again through human evaluation – of capturing the ideas expressed in the reviews of any given beer. For each of these applications, we adapt the Transformer neural net architecture. The work herein is broadly applicable in marketing, particularly in the context of online reviews. We close with suggestions for additional applications of our model and approach as well as other directions for future research.

* Corresponding author.
E-mail addresses: keith.e.carlson.gr@dartmouth.edu (K. Carlson), kopalle@dartmouth.edu (P.K. Kopalle), prasad.vana@tuck.dartmouth.edu (P. Vana).

https://doi.org/10.1016/j.ijresmar.2022.02.004
0167-8116/

1. Introduction

Online product reviews are ubiquitous and helpful sources of information available to consumers for making purchase decisions (Wu, Che, Chan, & Lu, 2015; Zhao, Yang, Narayan, & Zhao, 2013; Zhu & Zhang, 2010). Consumers rely on both the quantitative aspects of reviews such as valence and volume (Chintagunta, Gopinath, & Venkataraman, 2010) as well as the text content of reviews (Chakraborty, Kim & Sudhir, 2021; Ludwig et al., 2013) to learn about product quality and fit. Firms are likewise concerned about the effect of reviews on sales (Chevalier & Mayzlin, 2006; Vana & Lambrecht 2021) and actively engage in review management (Chevalier, Dover, & Mayzlin, 2018). Much of the literature in this area has focused on reviews and opinions written by non-expert consumers (Kannan & Li, 2017; Moe & Schweidel, 2012; Moe & Trusov, 2011), experts (Camacho, De Jong, & Stremersch, 2014; Luo, Zhang, Gu, & Phang, 2017), and managers of businesses (Proserpio & Zervas, 2017). Notably, the reviews explored in all the cases comprised by this body of research are written by humans.

Herein we turn our attention to the task of the writing of product reviews by machines, that is, through the use of tools and techniques of artificial intelligence algorithms. The high-level schema is one in which the model is trained through the linkage of product review text to a set of features of the product (the "metadata") and as a result of this training, the model is able to write (generate) new reviews given new (and thus, unseen) metadata.

This endeavor is motivated by the possibility of addressing two particular review-related challenges. First, while consumers rely on the reviews written by experts for a variety of experience goods such as books (Berger, Sorensen, & Rasmussen, 2010), movies (Basuroy, Chatterjee, & Ravid, 2003; Reinstein & Snyder, 2005), wines (Friberg & Grönqvist, 2012) and new technology products (Luo et al., 2017), these domain-specific expert reviewers increasingly face an unmanageably high volume of products needing written reviews.[1] This issue becomes all the more challenging given the complexities involved in describing experience products such as wines (Moon & Kamakura, 2017). Machine-generated reviews could help reduce the workload of the expert. We have in mind a review workflow wherein the machine would write a first draft of a review, which may or may not need to be edited further for production-level quality.

The second challenge relates to the consumer as well as retailer problem of "review overload" typical in most online shopping environments today where products often accumulate hundreds or even thousands of reviews. While consumers find reviews useful in their search for products, they face (and may experience) significant search costs to sift through so much text to find the relevant information. On the other side of the sales coin, retailers find it challenging to optimize screen real estate for product pages so that they can succinctly display review information. A convenient solution to this problem could take the form of the online retailer providing a "review synthesis"[2] on the product page that captures the ideas present in the reviews for that particular product. Our approach to the machine-writing of expert reviews adapts well to this challenge.

We work in two specific empirical contexts – wine and beer. Both are experience goods and each gives us different kinds of data. Our data for the context of production of expert reviews includes approximately 125,000 Wine Enthusiast reviews[3] published from 1996 to 2016. For the challenge of the production of review synthesis, our data consists of 142,860 reviews made up of ten distinct reviews each of 14,286 distinct beers.[4] In both cases, the data include the text of the review, a numerical rating, and also a number of attributes (metadata) pertaining to the wine or beer such as alcohol content, taste etc.

We make use of a Transformer-based deep learning neural net architecture for both empirical contexts. This architecture is built on two main components: an *encoder* and a *decoder* (cf. (Vaswani et al., 2018) while herein, some detail is given in Section 3). As per the standard practice in machine learning, the data are split into training and test sets. The former is used to tune the Transformer-based model, a process that produces a set of parameters such that ultimately, given metadata from the training input, will generally generate text that is similar to the linked human review. Test set data is not seen by the model before and is used to validate the model – showing that given new metadata, the model can generate (write) a review *de novo*.

We validate the writing in a number of ways by varying the attributes of a product (such as the wine rating) to generate review text that corresponds to the new set of product attributes. These variations are imaginary (in the sense that they attach metadata to a wine – e.g., a rating -that was not assigned by the reviewer), but the comparison of the reviews of the modulated attribution set to those of the original are instructive and show the range of the application while also revealing the ways in which the different attributes affect the language production.

We test the degree to which our machine-generated "expert" wine reviews are human-like via an MTurk study. Respondents see randomly selected reviews from a mix of human-generated and machine-written reviews. The survey asks respondents to rate the extent to which they think each review was written by a machine or a human being by allocating a total of one hundred points, and also their purchase intention in response to each wine review. Our results show that respondents

[1] For example, 17 wine review experts in our data wrote about 125,000 reviews, averaging over 7,300 reviews each. Another example includes Kirkus book reviews, where 200 expert reviewers write about 10,000 book reviews a year for an average of about one book review per week per expert (https://www.kirkusreviews.com/indie-reviews/).

[2] We try to stay away from the term summary as this can have a range of meanings – from a statistical description of certain kinds of word usage, to particular kinds of textual representations that decide to highlight different kinds of qualities in a large number of reviews. This is discussed in more detail in Section 6.

[3] Reviews obtained from www.winemag.com.

[4] Reviews obtained from https://www.ratebeer.com/.

were unable to tell, with any statistical significance, if a review was written by a human being or a machine. Similarly, there is no statistical difference between the true sources (i.e., machine vs human) in the participants' responses to the purchase intention question.

We next focus on the validation of the review synthesis task using another MTurk study. Here, we ask participants to rate on a scale of 1 to 7 the degree to which they feel that the machine-generated review synthesis of beer reviews captures the ideas and concepts mentioned in the corresponding 10 human-written reviews for that beer.[5] The average participant evaluation statistically significantly exceeds the neutral response of 4, supporting our claim that our Transformer-based synthesis algorithm is competent at the task of generating a synthesis of product reviews.

Even when restricted to the context of machine-generated reviews for wines, our results have managerial implications beyond the efficiencies gained by assisting expert reviewers by providing a first draft of review text. One example could be for restaurants and bars to describe the wines that they offer on their menus. Since most restaurants cannot afford to have an in-house sommelier, they may lack the expertise to describe the wines present on their menus. Our algorithm could aid the managers of such restaurants and bars. Another use case could be in generating text for brochures of wineries in describing the wines they offer. The extension to other product spaces with analogous metadata and expert review text is straightforward. Our results have wide applications in marketing, particularly in the context of online reviews.

To give just one other possibility, this technology can also be used in other contexts where product data are available and a review of the product is required. A different non-expert review use case for our Transformer-based model relates to a challenge faced by many retailers on platforms such as Etsy. Small to medium size sellers on these platforms, of which there may be thousands, can find themselves confronted with the task of describing hundreds of products that they sell. Armed with slightly different training data than the current study, one where the output is a description of the product rather than a review, an extension of the current model could quickly automate this task at scale. Another application is to provide a synthesis of reviews where consumers does not have the time or energy to sift through a large number of reviews.

While the use of artificial intelligence tools such as Transformer-based models to create machine-generated reviews opens the doors for several productive opportunities, it also raises some important ethical issues, especially with respect to the construction of a review without an actual purchase or an experience (Anderson & Simester, 2014). In particular, there is a distinct possibility that a consumer reading the review could assume that the experiential aspects of the review, such as the description of the taste and smell, are the result of the expert reviewer's own personal experience with the product while the review could have been machine-generated. In fact, anecdotal evidence from our wine MTurk study hints at these issues. We asked respondents to comment on the benefits of machine-generated reviews and several respondents mentioned that machines cannot have the subjective experience of a product such as a wine or a beer or coffee the way humans do. We envision the machine-generated review to be a starting point for the expert reviewer rather than a finished, publishable review. In the event there is no human intervention in the review, the platform may indicate that the review was generated by a machine or at a minimum, say nothing about the source of the review or the synthesis.[6]

We uncovered a related ethical issue in our beer MTurk study. Our results indicate that participants are significantly more likely to think that a review synthesis better captures the ideas present in the reviews if they were either told that the synthesis was written by a beer enthusiast rather than a machine or the neutral condition where we do not mention the source. This result raises further important ethical concerns since it appears that retailers could make better (i.e., actionable) use of machine-generated reviews by (falsely) attributing them to humans. We look forward to studies that will investigate in detail these kinds of ethical complexities.

2. Natural Language Processing (NLP) and Artificial Intelligence (AI) in Marketing

Marketing researchers and practitioners have over the past decade identified the potential of user-generated content and leveraged contemporaneous developments in natural language processing (NLP) techniques. Early work in marketing-related NLP focused on mining various sources of text data such as online reviews (Decker & Trusov, 2010), comments (Singh, Hillmer, & Wang, 2011), blogs (Onishi & Manchanda, 2012), and social media (Ghose, Ipeirotis, & Li, 2012). Subsequently, this stream of research has used text data to either complement existing marketing models such as conjoint analysis (Lee & Bradlow, 2011), perceptual maps (Gabel, Guhl, & Klapper, 2019; Moon & Kamakura, 2017), survival analysis (Zhang & Luo, 2021) and market structure analysis (Netzer, Feldman, Goldenberg, & Fresko, 2012) or in developing new models that utilize text data to substitute traditional marketing models in various domains, including the identification of customer needs (Timoshenko & Hauser, 2019), and the analysis of brand perception (Liu, Dzyabura, & Mizik, 2020; Nam, Joshi, & Kannan, 2017).

Methodologically, NLP-based work in marketing has used techniques such as sentiment analysis (Hennig-Thurau, Wiertz, & Feldhaus, 2015), topic modeling (Huber, Kamakura, & Mela, 2014; Nam et al., 2017), and the use of word count software such as LIWC (Cavanaugh, Bettman, & Luce, 2015; Villarroel Ordenes, Ludwig, De Ruyter, Grewal, & Wetzels, 2017). With advance in computing power and NLP techniques, more sophisticated machine learning and AI methods are now being

[5] Participants were told that 1 stands for "does not capture the ideas at all" while 7 stands for "captures the ideas very well" with 4 being the neutral response.

[6] There is some analogy to the use of Photoshop for modifying images used in advertising, making models appear in a particular visual style. It is now the case in France that such images need to come with a warning label (see https://www.bbc.com/news/world-europe-41443027).

applied. This includes the use of support vector machines (Kübler, Colicev, & Pauwels, 2020), random forests (Netzer, Lemaire, & Herzenstein, 2019), XGBoost (Zhang & Luo, 2021), and deep neural networks (Hu, Dang, & Chintagunta, 2019; Liu, Singh, & Srinivasan, 2016). More recently, transfer learning (Liu et al., 2020), reinforcement learning (Schwartz, Bradlow, & Fader, 2017) and convolutional neural networks (Timoshenko & Hauser, 2019) have been used in marketing research. These models allow researchers great versatility in combining text data with other sources of unstructured data such as images (Li & Xie, 2020) and video (Li, Shi, & Wang, 2019).

The focus thus far in marketing has not been on models that generate text as output. This may be because generative models have primarily been used in translation tasks where the input is text in one language and the output is the translation of that text in a different language. Our contribution is to utilize this technology and have as input metadata of a product and generate as output human-readable text that describes a product or provides a synthesis of reviews.

Machine-written reviews naturally extend the many ways in which AI and machine learning are already affecting the world of marketing and sales. In particular, AI-driven marketing is evolving along the two dimensions of human-machine interaction and automated analysis (Kopalle, Gangwar, Kaplan, Ramachandran, & Reinartz, 2021). Technological advances in these areas have enabled the easy processing of large-scale unstructured data and have demonstrated strong predictive performance in several marketing tasks given their flexible model structures (Ma & Sun, 2020). Several traditionally human-handled marketing tasks such as the targeting of ads (Forbes 2020), chat (Luo, Tong, Fang, & Qu, 2019), lead generation (Syam & Sharma, 2018), market research (Haenlein, Kaplan, Tan, & Zhang, 2019), and services (Huang & Rust, 2018) have now been successfully transferred to machines. In addition, research has also explored interesting ways in which humans and AI can co-create products (Zhang, Sun, Luo, & Golden, 2021). We hope that our novel approach to review generation will enable marketing scholars and practitioners to study and employ AI technologies in an effective manner for firms and consumers.

3. A Transformer Framework For Machine-Writing Of Expert Wine Reviews

As a first task we take up the challenge of machine-writing human-level expert wine reviews. More explicitly, we take on the challenge of using a standard set of wine features as the input to a review-writing Transformer-based algorithm or technology. The features are the metadata that accompany each entry in a curated collection of reviews and suggests a clear schema for tackling the problem. In what follows we explain the data (3.1), outline the implementation (this includes a description of the Transformer architecture) (3.2), and discuss the results (3.3) and model requirements (3.4). A related experiment in which we detail a Turing Test-style evaluation of the output can be found in Section 4.

3.1. Wine review data

Wine reviews are an excellent use case for the machine-writing of expert reviews for two broad reasons. First, wines are complex products whose reviews generally include and link objective characteristics (e.g., price, alcohol content, vineyard) with experiential characteristics (e.g., taste and aroma) that the typical consumer values. Second, as a data source for a machine learning problem, the linkage of the wine review text with its metadata is well-suited to the deep learning approach.

We work with data obtained from *Wine Enthusiast* reviews[7] published between 1996 and 2016. The data are collected directly from www.winemag.com. As a datum, a wine review is a pairing of wine review text with metadata (descriptors). The metadata fields are *name of wine*, *points* (rating given by reviewer), *review author*, *price*, *designation*, *variety* (Riesling, Merlot, etc.), *appellation* (geographic location where grapes were grown), *winery*, *alcohol content*, *category* (red, white, rosé, dessert, fortified, etc.), and *date of review*. Table 1 shows a few examples.

We discard from the full set the handful of entries where there is no text representing the actual review. After this simple culling, our data has 201,431 unique wines, approximately 125,000 of which are identified as having been written by one of 17 authors, with the remaining having no identified author. Fig. 1 shows the distribution of wine category by year while Fig. 2 presents the distribution of ratings across the corpus of reviews. The figures show that the majority of wines in the data are reds and the ratings occupy a relatively narrow range between 80 and 100. Due to the proprietary nature of this data (owned by the website), we are unable to make the full data publicly available, but those interested in working with it may collect the data directly from winemag.com.[8]

Next, we perform some standard and necessary processing preparations. We first build a vocabulary from the reviews. This comprises approximately 20,000 tokens[9] built on both the wine attributes and the review text. Words which appear infrequently are not included in the vocabulary directly, but sub-words are included to represent them. For example, the word "government" is rare in our data and is represented by three tokens, "go", "vern", and "ment", when it appears. This tokenizing is a standard preparatory stage in NLP work. Using a fixed vocabulary helps to constrain the size of the model while still enabling it to learn representations of rare words. We then randomly split the reviews into three sets: a training set of 90% of our data and

[7] www.winemag.com.

[8] The reviews can be "scraped" using standard website scraping/data gathering techniques.

[9] "Tokens" is the natural language/text processing term for words or character strings that are found with some frequency in the text corpus.

Table 1
Example of the Formatted Paired Data.

Wine Information	Wine Review
<Ancient Peaks 2010 Cabernet Sauvignon (Paso Robles)> <84> <N/A> < $17> <N/A> <Cabernet Sauvignon> <Paso Robles, Central Coast, California, US> <Ancient Peaks> <14%> <Red> <12/31/2012>	This is a sound Cabernet. It's very dry and a little thin in blackberry fruit, which accentuates the acidity and tannins. Drink up.
<Feiler-Artinger 2012 Beerenauslese Traminer (Burgenland)> <92> <Roger Voss> <N/A> <Beerenauslese> <Traminer, Other White> <Burgenland, Austria> <Feiler-Artinger> <12%> <Dessert> <11/1/2013>	Very young wine, dominated by its textured spiciness and ripe tropical fruits. It is rich, full of pepper, cinnamon, and a dry core that makes it as much rich as sweet.

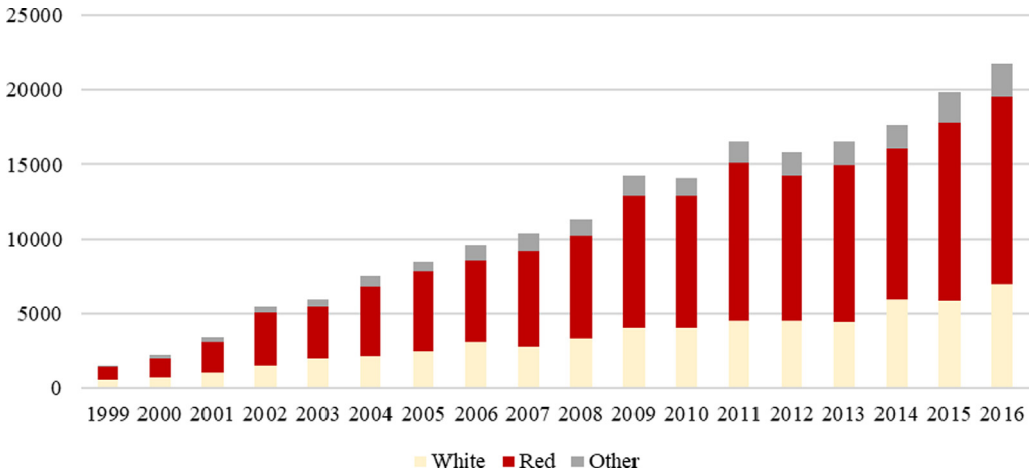

Fig. 1. Category of Wines in the Data by Year.

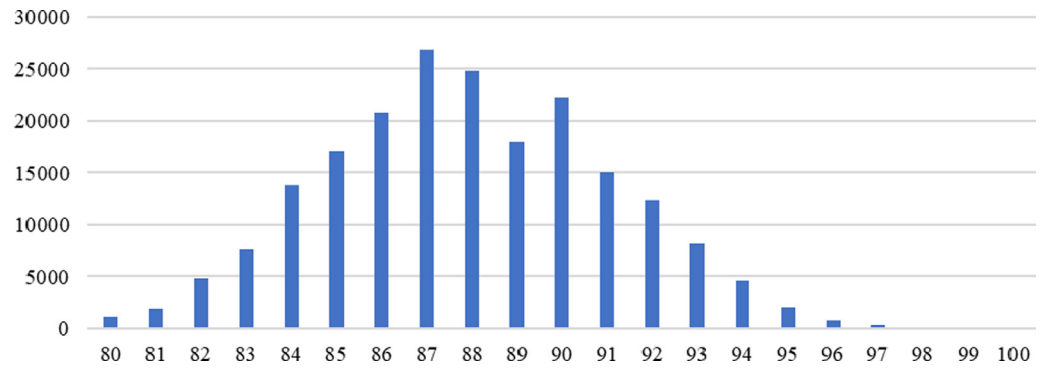

Fig. 2. Distribution of Rating of Wines in the Data.

test and development sets each of 5%. This is a common train/test/development apportionment in machine learning and NLP research.

3.2. The Transformer deep learning model

Computer models to create natural language text have a long history. Relevant examples from the computer science literature include the automatic generation of stories (McIntyre & Lapata, 2009), summarization of text (Conroy & O'leary, 2001; Nallapati, Zhou, Gulcehre, & Xiang, 2016), style transfer (Carlson, Riddell, & Rockmore, 2018; Xu, Ritter, Dolan, Grishman, & Cherry, 2012), and machine translation (Cho, Bart Van, Caglar, Dzmitry, Fethi, Holger, & Yoshua, 2014; Koehn et al., 2007). The last of these is key to our work and progress on machine translation models and has been the engine for many of the extraordinary recent advances in NLP. Early efforts focused on rule-based methods (Kay, 1984), which were later replaced by statistical methods (Brown et al., 1990; Koehn et al., 2007). Neural networks have been widely applied to

these types of tasks, achieving state-of-the-art results on a variety of language generation tasks including translation (Bahdanau, Cho, & Bengio, 2014).

Our work makes use of the *Transformer* (Vaswani et al., 2017), a *feed-forward neural network* architecture originally designed and used for machine translation. Our implementation relies on the publicly available *tensor2tensor* software library (Vaswani et al. 2018) package. We now give a brief overview of the model. Note that all of our terminology and references are standard for the neural network learning paradigm. A full description of this important machine learning paradigm is well-beyond the scope of this paper. We encourage the interested reader to see Vaswani et al. (2017, 2018) and the references therein for more detail.[10]

As in many neural translation models, at the heart of the Transformer model rests an *encoder* and a *decoder*. The encoder takes in a sequence of tokens, in our case the words describing the attributes of a product and creates vector representations of the tokens. These vector representations reflect the parameters of underlying neural networks related to the text structure as embodied in the inputs. The decoder then uses those vector representations to create a sequence of probability distributions over output tokens, which are the words in the review vocabulary. During training, the model is given data about a product as input and generates guesses about the words in the linked review. The model adjusts internal parameters so that its predictions of words in the review get closer (by some metric) to the words in the linked review. After enough examples and adjustments (all done automatically within the training stage), the parameters settle onto values such that the predicted output improves enough to produce plausible reviews, both for products the model has never seen before (i.e., outside the training set), or even for products that don't exist (in the sense that there does not actually exist a product with the particular attributes fed to the network).

Fig. 3 is an illustration of the Transformer architecture. We see a standard sketch of the network, along with an actual example input and machine-generated output from our work. As shown, the model takes in the sequence of input tokens, $x_1, \ldots, x_n$, describing the attributes of a wine (the metadata). The first component of the model is the encoder, which produces vector representations, $z_1, \ldots, z_n$, of the tokens, turning each token into a real-valued vector in a common high-dimensional space (the z_i are akin to word embeddings (Mikolov, Yih, & Zweig, 2013)). The embedding is such that the relative positions of each of the words in this high-dimensional space enable a measure of the likelihood of their co-occurrence in the wine reviews. Ultimately, the coordinates are a reflection of parameter weights in the underlying neural network.

Fig. 3 shows that the encoder part of the model is in fact a cascade of several layers of encoders, where each layer takes the embedding of the previous layer (the first layer takes a word embedding combined with an embedding of the position of the token) in a succession of updates of each token's embedding. This process uses a "self-attention mechanism" (generally, "attention" refers to the ability to base calculations on specific embeddings created earlier or in the case of self-attention, simultaneously) meaning that the embedding is successively updated based on the embeddings of the other input tokens from the same level of the encoder. This allows for successive and simultaneous integration of parameter information that effectively incorporates "neighborhood" information from the training data.

Similar to the encoder, the decoder is also a cascaded structure, built out of several decoder layers which operate like the layers of the encoder, albeit (roughly) backwards. That is, each decoder layer also adds an additional attention module which attends to the embeddings produced by the final layer of the encoder. Each decoder layer produces one output token at a time and the attention mechanism of each embedding is only allowed to see tokens that are produced earlier in the output than the current token. For example, the fourth token of the decoder can only attend to the first three output token embeddings in each layer. In this manner, the embedding of each decoder position can rely on the embeddings of all of the embeddings of the input as well as the embeddings of the output from earlier time steps.

Once embeddings from the final layer of the decoder are produced, a *logit model*[11] is used to turn these embeddings into conditional probability distributions over the words in the vocabulary. The model has the encoder's z_i vectors that are conditioned on the input X vector and since it has the previous tokens of the output, the Transformer architecture can condition on the output of previous timesteps, i.e., those $y < t$, thereby producing an overall conditional probability on output word strings. That is,

$$p(Y|X) = \prod_{t=1}^{T} p(y_t|X, y_{<t}). \tag{1}$$

During training, the model is given the input and generates probability distributions over the words in the review as output. By comparing the actual human-written review with the model-generated output at each iteration, the model adjusts internal parameters in all layers of the encoder and decoder so that its predictions get closer to the truth. These internal parameters are grouped into θ in Equation (2) below and the overall log-conditional probability is modified to minimize a loss function:

$$\ell(\theta) = -\sum_{t=1}^{T} \log p(y_t|X, y_{<t}; \theta). \tag{2}$$

[10] More details about the Transformer can also be found at http://jalammar.github.io/illustrated-transformer/.

[11] In the computer science literature this is often referred to as a "softmax" routine.

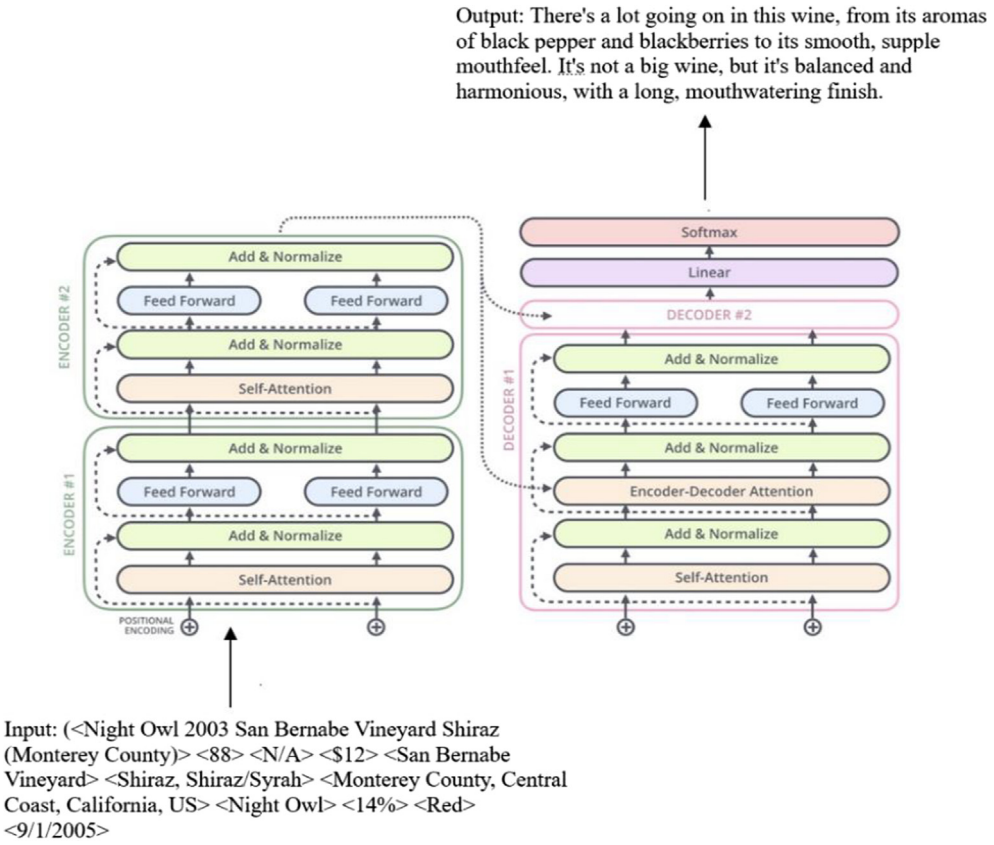

Fig. 3. Representation of the Transformer network with example input and output. (Source. http://jalammar.github.io/illustrated-transformer/ (Jay Alammar's Visualizing machine learning one concept at a time). Used with permission.)

During inference, these probability distributions are used to identify the tokens most likely to be incorporated into a new machine review.

We set most of the parameters of the Transformer network as described in Vaswani et al. (2017). These choices include using six layers each for the encoder (where each layer is itself a connection of sublayers of a so-called "multi-head self-attention layer" and a full connected feed-forward neural network) and decoder (which includes a second self-attention layer), embeddings with 512 dimensions, and dropout (node removal) probability of 0.1 between layers. Thus, ultimately the inputs generate 512-dimensional representations (embeddings) of the tokens.

At each iteration, loss is noted for the development data, which is similar to the test data as both represent a random 5% sample of the collected data. The difference is in their use, with development data being used to evaluate the model during training and test data being used for evaluation of the fully trained model. Training is stopped when the loss on development data has not improved over the current best for two consecutive evaluations. During our training algorithm, this occurred after 65,000 iterations (steps). Fig. 4 shows the loss throughout training for both the training and development data.

The model is fit using our training data with a batch size of 1024 (meaning each training step processes input with a total of 1024 tokens) and evaluated using the development set after every 2,500 iterations. The trained model is then given wine metadata from the test set and the model automatically generates the review text for these wines in the test data. As mentioned above, we train the model using 90% of our data. This splitting of the data means that the model never sees the wines in the test set during the training phase.

As mentioned earlier, a deep dive into the workings of the model are beyond the scope of the paper. That said, we need to address two specific model parameters – or rather "hyperparameters" as they are called – that govern large-scale algorithmic choices: the *alpha parameter*, which controls the length penalty applied to output, and the *beam size*, which affects how many candidate reviews the model considers at each step as it generates the output word-by-word. Alpha was introduced in Wu, Schuster, Chen, Le, Norouzi, Macherey, and Klingner (2016) and they explain the motivation, stating: "Without some form of length-normalization regular beam search will favor shorter results over longer ones on average since a negative log-probability is added at each step, yielding lower (more negative) scores for longer sentences. We first tried to simply divide by the length to normalize. We then improved on that original heuristic by dividing by length$^{\alpha}$ . . ." In their experiments they report that an alpha between 0.6 and 0.7 usually gave the best results.

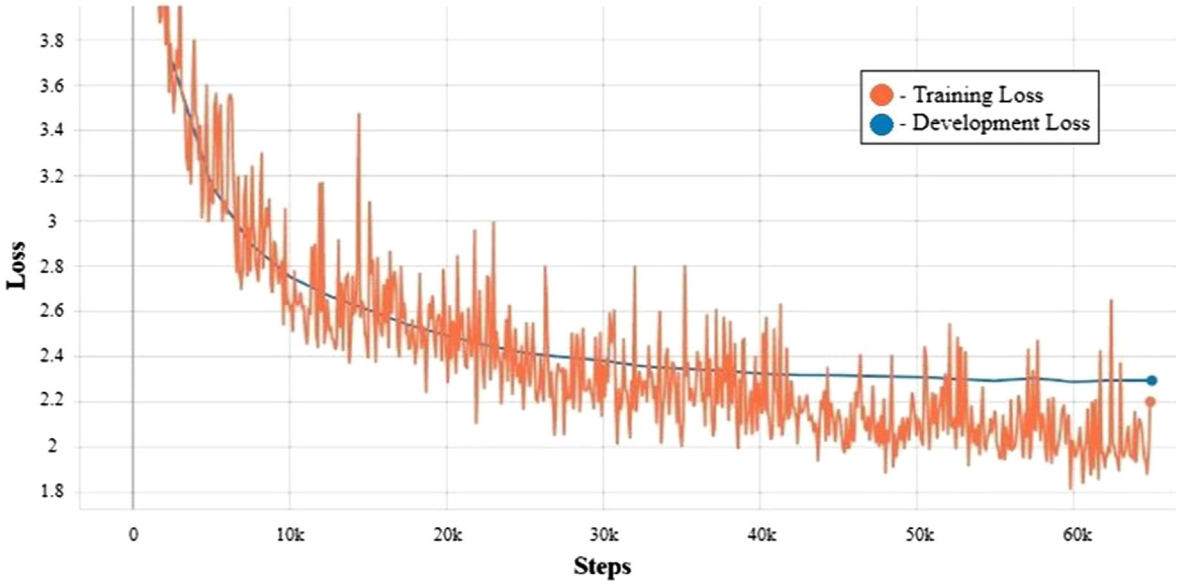

Fig. 4. Transformer Loss During Training.

Beam size controls the number of candidates considered during each step as the model generates an output word by word. With a beam size of 5, for example, the model begins by finding the 5 most likely words to start the review. For the second word, it considers all possible words paired with any of the current 5 candidates and of these possibilities chooses the 5 (now two-word phrases) scored most likely by the model. These 5 candidates after the second step may all use the same first word or may be spread over the 5 candidates for the first word. Generation of the third word is built on these 5 two-word phrases and so on.

Table 2 shows reviews generated by the model for one example wine (a Merlot) in the test data for various settings of the two parameters. As can be seen, the model is responsive to these parameters and generates slightly different reviews for (most) different settings of the parameters (it will sometimes generate identical reviews if the parameter values are similar). Table 3 shows reviews generated for another example (a Rosé) for various parameter settings. We find no appreciable differences among the reviews for the range of settings of these parameters and for the reviews presented in the rest of the paper, we set alpha to 0.7 and the beam size to 20. These correspond to (deep learning) industry standards and seem to work well for this task. Smaller values for alpha would cause the model to produce shorter reviews, and at the extreme, an alpha of 0.0 would mean there is no length normalization at all. Higher values for alpha create longer reviews which can suffer from an obvious decline in quality as the model is incentivized to produce more words even if it doesn't have a good guess of what those words should be. The goal here is to find a balance between these extremes that produce output that looks the most human. Fortunately, alpha can be changed at decoding time using the same trained model so some experimentation with alpha is not very computationally costly as the training of the model doesn't need to be repeated.

3.3. Results

Having tuned the model, we create reviews for every wine in the test set. These are new inputs (metadata), never seen by the model in training. We can compare the machine-generated review with the human review. We also perform a class of experiments altering characteristics of the input. In particular, we experiment with both changing the author and the rating of the review. Since both fields are independent of the wine itself, this allows us to generate multiple reviews for the same wine – effectively review-writing "Gedanken experiments" where one review is the wine as experienced by a different expert and second as experienced differently by the same expert.

In Table 4, we show an original review from the test set, the machine-generated review using the original metadata and then a machine-generated review where the metadata has been changed by increasing the rating. Note reasonable agreement in sentiment and descriptors between the human- and machine-written reviews with the original metadata and the much more effusive review with increased rating. In Table 5, we do the same, except the ratings are decreased in the second case. Again, we see nice agreement between the sentiment and descriptors of the human-written and machine-written reviews of the same metadata but a much more muted review with the decreased rating. Additional examples are given in Tables 6, 7, 8, and 9. Overall, we note that our Transformer architecture model appears to be reflecting many of the human observations about the wines with good accuracy. As we discuss later in Section 4, a simple comparison of word usage in the human- and machine-written reviews shows nice agreement in the statistics.

Table 2
Reviews Generated with Different Decoder Settings for a Merlot.

Beam Size	Alpha	Generated Review
1	0.7	A nice, easy-drinking Merlot, with pleasant cherry, blackberry and herb flavors. It's dry and balanced, with a good grip of tannins
20	0	There's lots to like in this dry, full-bodied wine. It has flavors of cherries and herbs, and is very dry
20	0.6	There's lots to like in this dry, full-bodied wine. It has flavors of cherries and herbs, and is very dry
20	0.7	There's lots to like in this dry, full-bodied wine. It has pleasant flavors of cherries and blackberries, and is very dry
20	0.9	There's lots to like in this dry, full-bodied wine. It has pleasant flavors of cherries and blackberries, and is very dry
20	1	There's lots to like in this dry, full-bodied wine. It has pleasant flavors of cherries, blackberries and herbs, and is very dry
50	0	There are some good flavors of blackberries and cherries in this dry, full-bodied wine. It's a little rough around the edges
50	0.6	There are some good flavors of blackberries and cherries in this dry, full-bodied wine. It's a little rough around the edges, though
50	1	There are some good flavors of blackberries and cherries in this dry, full-bodied wine. It's a little rough around the edges, but it's a nice sipper

Table 3
Reviews Generated with Different Decoder Settings for a Rosé.

Beam Size	Alpha	Generated Review
1	0.7	This is a very dry, acidic wine, with a scour of tannins. It has subtle flavors of cherries, herbs and spices, and is very clean
20	0	Pretty dark for a rosé, and full-bodied, with cherry, raspberry, vanilla and spice flavors. It's dry, with good acidity
20	0.6	Pretty dark for a rosé, and full-bodied, with cherry, raspberry, vanilla and spice flavors. It's dry, with good acidity
20	0.7	Pretty dark for a rosé, and full-bodied, with cherry, raspberry, vanilla and spice flavors. It's dry, with good acidity
20	0.9	Pretty dark for a rosé, and full-bodied, with cherry, raspberry, vanilla and spice flavors. It's dry and balanced, with a clean, fruity finish
20	1	Pretty dark for a rosé, and full-bodied, with cherry, raspberry, vanilla and spice flavors. It's dry, with good acidity and a clean, spicy finish
50	0	A little heavy and full-bodied for a rosé, but there's no denying the ripe cherry, raspberry and chocolate flavors
50	0.6	A little heavy and full-bodied for a rosé, but there's no denying the ripe cherry, raspberry and chocolate flavors
50	1	Pretty dark for a rosé, and full-bodied for a rosé. It's dry and full-bodied, with cherry-berry flavors and a spicy finish

Table 4
Source Review, Machine Review, Machine Review with Increased Rating.

Rating	Source	Review
84	Human	This blush wine has Viognier blended in with Syrah. It's a pretty salmon color, and has flavors of raspberries, peaches and vanilla. Fully dry, it will be nice with grilled salmon
84	Transformer	Pretty dark for a rosé, and full-bodied, with cherry, raspberry, vanilla and spice flavors. It's dry, with good acidity
97	Transformer	This is one of the best California rosés I've ever had. It's so dry, so rich in fruit, it's almost like a dessert wine, except it's totally dry. The flavors are enormously complex, ranging from cherries and raspberries to white chocolate, and the finish goes on and on

Table 5
Source Review, Machine Review, Machine Review with Decreased Rating.

Rating	Source	Review
94	Human	The estate vineyard is in the Vaca Mountains, east of the city of Napa. You can taste the cool climate in the brisk acidity that frames this 100% Cabernet Sauvignon. It's very ripe in blackberries, with rich layers of toast from new French oak. For all the sweet fruit, the finish is dry, even rasping. Such are the tannins that the wine feels tough, even at this age. Give it another five years, at least
94	Transformer	This is one of the best Cabernets of the vintage. It's so delicious, you can hardly keep from drinking it now, but it should develop over the next 6–8 years
83	Transformer	Soft and simple, with sweet-and-sour cherry and raspberry flavors. Drink now

3.4. Model requirements

Generally speaking, the bigger the size of the training data for a deep learning architecture, the better it is at the prediction task. Deep learning architectures tend to need thousands of examples (at least) to train on. One study that has experimented with data size is Murray, Kinnison, Nguyen, Scheirer, and Chiang (2019), in which methods are developed for automatically optimizing the model size and parameters of a Transformer for translation tasks. They test with training data of varying sizes, ranging from 15,000 to 235,000 pairs of sentences. They generally find better results with larger datasets, although the numbers are not directly comparable across datasets as the language pairs are different for the different datasets. That said, for our wine review generation task, we recommend using data sets comprising on the order of at least 10,000

Table 6
Source Review, Machine Review, Machine Review with Increased Rating.

Rating	Source	Review
87	Human	Stone fruit, fragrant white rose, natural rubber and tangerine blossom mark the nose of this pretty effort. The wine's texture is thick and oily, and there's a touch of sweet spice on the close
87	Transformer	This opens with a fragrant bouquet of white flower, honey and exotic fruit. The mouthfeel is smooth and rich, with a touch of bitter almond on the close
97	Transformer	This is a gorgeous expression of Gewürztraminer from one of the best producers in the world. It opens with a beautiful bouquet of white flower, honey and exotic fruit. The mouthfeel is smooth, silky and long-lasting

Table 7
Source Review, Machine Review, Machine Review with Increased Rating.

Rating	Source	Review
90	Human	Lime zest and grassy herb notes infuse freshness into this dry, light-footed Riesling. It's particularly sprightly in feel, a crush of tart white grapefruit and pineapple flavors, but it penetrates long on the finish, accented by a kiss of citrus blossoms. Drink now–2017
90	Transformer	While the nose is a bit closed, the palate of this off-dry Riesling is chock full of juicy white grapefruit and tangerine flavors. It's not a deeply concentrated wine, but it's balanced neatly by a strike of lemon-lime acidity that lingers on the finish
97	Transformer	While demure on the nose, the palate of this off-dry Riesling is chock full of luscious white grapefruit and tangerine flavors. It's feather light on the palate, yet deeply concentrated, with a long, mineral-tinged finish. Drink now through 2020

Table 8
Source Review, Machine Review, Machine Review with Decreased Rating.

Rating	Source	Review
90	Human	While many dry wines from the Mosel are overly tart, this offering is a very pleasant exception. It is stony, minerally and a bit austere, offering up the barest hints of peach kernel and spice, but it's also characterful. It finishes with good length, showing firmness alongside dried-spice notes
90	Transformer	This is a medium-bodied, off-dry Riesling, with aromas and flavors of green apple, lime and wet stones. It's crisp and refreshing, with a long, mouthwatering finish
83	Transformer	This medium-bodied, off-dry Riesling is a bit of a disappointment from this normally reliable producer. There's a hint of nail polish on the nose, followed by simple apple and citrus flavors

Table 9
Source Review, Machine Review, Machine Review with Decreased Rating.

Rating	Source	Review
90	Human	Golden and extracted. Butter-sautéed bananas, intense tangerine and peach and pineapple custard and vanilla are among the flavors. Good acidity, but so sweet and delicious
90	Transformer	If you're looking for a dry, crisp white wine, try this one. It's filled with citrus, peach, wildflower and vanilla flavors, with a rich, creamy texture and a long, spicy finish
83	Transformer	There's not much going on in this wine. It's bone dry and acidic, with lemon and lime flavors

reviews to produce reasonable output. Note that the exact amount of training data needed for a particular Transformer model run would depend on the variation in number of metadata variables available, amount of variation within a variable, differences in degree of correlation between the variables in metadata and reviews, range of common vocabulary in reviews, consistency of format, and the quality of reviews.

To test the boundary conditions of our model, we conduct two small experiments. The objective is to check how robust the Transformer architecture is to having fewer variables in the metadata or less data. In one experiment, we use only 4 metadata fields rather than 11 as in our full specification. We chose the fields that seem most important for creating reviews: name of wine, rating of wine, author of review, and variety of wine. In a second experiment, we reduce the number of reviews in our dataset by about half, from approximately 200,000 to 100,000, but use all 11 wine metadata fields. In Fig. 5, we see the measures of the loss during training, both for the data used for training (orange) and for the development data (blue). In both cases, the loss on development when training stops is slightly greater than with our original model run (Fig. 4). This indicates to some extent that these reduced models were worse at predicting the reviews in the held-out development data. However, even with that, the output reviews are "readable" and, at face value, of acceptable quality. Examples are presented in Tables 10 and 11.

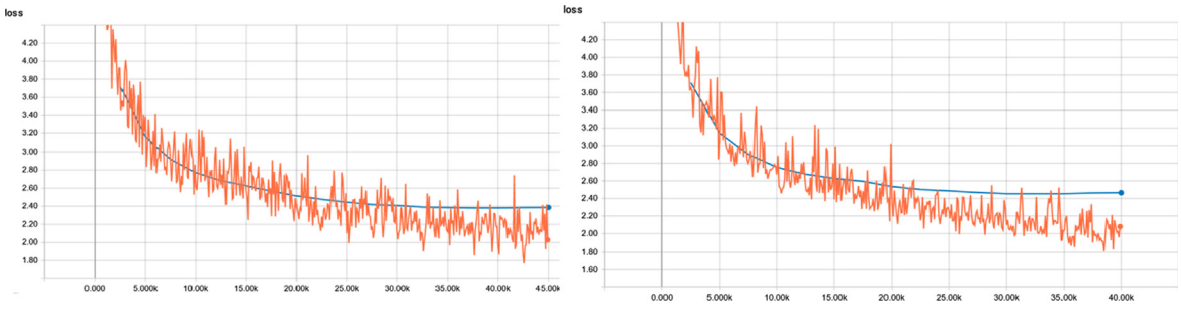

Fig. 5. (L) Training loss using only 4 of 11 metadata fields for the wine reviews. (R) Training loss using only 100,00 wine reviews.

Table 10
Example Machine Reviews Using Reduced Metadata (4 of 11 fields).

Review No.	Review
1	A bit of a letdown after the previous two vintages, but still a very good wine. It's dry and full-bodied, with a rich array of blackberry, plum, coffee and spice flavors that finish long and spicy. Drink now
2	Here's a good, everyday sort of Syrah. It's dry, full-bodied and tannic, with blackberry and coffee flavors that finish in a swirl of dusty spice
3	A blend of 60% Cabernet Sauvignon, 30% Merlot and 10% Shiraz, this medium-weight wine is balanced and easy to drink with soft tannins and a medium-length finish
4	While the nose is a bit closed, the palate of this dry Riesling is chock full of fresh apple and pear flavors. It's zesty and spry, finishing on a zesty mineral note. Drink now through 2019
5	A blend of Pinot Noir and Pinot Noir, this is a rich, full-bodied rosé, packed with red fruits and a touch of toast. It is ripe and full in the mouth, with a soft aftertaste. Sweet and oaky, with buttered toast and caramel flavors swamping the underlying pineapple and pear fruit
6	A solid, chunky wine, its tannins well integrated into the ripe red fruits and acidity. The wine has weight and richness, while the tannins are already well integrated. Age for 3–4 years

Table 11
Example Machine Reviews Using Reduced Training Data (100,000 out of approximately 200,000 reviews).

Review No.	Review
1	A blend of 60% Pinot Noir and 40% Chardonnay from the Montagne de Reims, this is a ripe, full-bodied wine. It has a touch of toastiness as well as crisp apple and citrus flavors. The wine is ready to drink
2	There's a nice roundness to the nose and mouth of this wine, with aromas and flavors of red apple, orange rind and a touch of sweet spice. Round and full in the mouth with a medium-length finish
3	This is one of Chile's best wines to date. It's one of the best we've tasted on the market. The nose is pure and fruity, while the palate is loaded with cherry, raspberry and plum flavors. The finish is long and smooth, with just enough acidity to keep it moving in the right direction. Drink now through 2010. Imported by Ecovalley Quality Wine Group and National Refrescos Import Company, LLC
4	Here's a soft and approachable Cab that's ready to drink now. Medium weight with soft tannins and a medium-length finish, it's easy to drink now
5	A 100% varietal wine from one of the Livermore Valley's top-of-the-line Petite Sirah vineyards, this is a dark, brooding wine, full-bodied in tannin and oak, with a long-lasting finish
6	A blend of 50% Grenache, 25% Syrah and 25% Mourvèdre, this is a full-bodied, creamy-textured wine that's ready to drink now

4. Experimental Evaluation with Human Respondents

The previous section provides some examples of machine-generated reviews and shows how varying the metadata affects the review. This begs a question central to this paper: "Does a machine-written review seem human-written?". That is, given a corpus of human-written reviews plus product characteristics, can a machine learn to write a wine review of "human quality"? Here, we assess whether the machine is able to generate reviews similar to that of human beings with an experiment involving an independent sample of five hundred and one Amazon Prime Mechanical Turk (MTurk) participants. We also test how likely consumers are to purchase the wine based on the expert-generated review versus the machine-generated review. Our survey puts respondents in the shoes of online shoppers and tests what decision they would make when a wine review is presented to them.

4.1. A wine review "Turing Test"

In this experiment, survey respondents see fourteen randomly selected reviews from a set of six hundred reviews, three hundred of which were human-written (by the expert wine reviewers) while the remaining three hundred were generated by our Transformer-based model (i.e., are machine-written). Please see the Web Appendix for the set of three hundred human-written reviews (Web Appendix WA.1) and the corresponding set of three hundred machine-written reviews generated by the Transformer architecture (Web Appendix WA.2) for a common set of three hundred wines. Again, the set of machine-written reviews and the set of human-written reviews are based on the same set of three hundred wines, all part of the test set. Thus, none of these wines were used to train our Transformer network.

Respondents were told that some of the reviews were written by humans who are wine reviewers and some were generated by a machine. We asked them to read each review carefully and indicate – through the answers to two questions – to what extent they thought the review was written by a machine or a human being: (i) indicating if they believed the review was written by a human being or a machine or unable to tell, and (ii) asking for an allocation of 100 points to the two types: review written by a human being and review generated by a machine.

The survey was completed by five hundred and one respondents from ten states in the U.S. Each respondent earned a reward of $1.50. We did not delete any observations either due to non-compliance or for any other reasons. Appendix A.1 provides the survey used in this study and Appendix A.2 presents the data descriptives. Overall, about 39.3% reported being female and 60.3% male. 74.3% report having Bachelor's degree or higher, and 50.7% report a household income $50,000 or higher.

While 10.96 percent of respondents said they were unable to tell if a review was written by a human being or a machine, a chi-square test ($\chi^2 = 0.24$ with $df = 2$) revealed no significant difference ($p = .886$) for this question between the two true sources of reviews: human being or machine. This supports the claim that from a consumer's perspective there were no perceived differences regarding the true review source. There was also no significant difference in respondents' mean allocation of 100 points to each of the two types of sources: written by a human being versus generated by a machine ($p = .99$). A comparison of the word usage in the two MTurk review sets also shows that there is little difference in the vocabularies and word frequencies (see Appendix B for a "bag of words" analysis).

As a next step, we estimated a random effects regression model that controls for the following: (1) review-specific effects, (2) respondent-specific effects. The dependent variable in the regression is the number of points (out of 100) each participant allocated for each review in terms of how likely the review was generated by a machine. The key independent variable is the true review source, which is a dummy variable, 1 for human-generated and 0 for machine-generated. Respondent specific effects were controlled by 500 dummy variables to represent the 501 respondents. Review specific effects were controlled via a random-effects specification. In addition, the standard errors were clustered at the review level.

Column (I) of Table 12 presents the results. As seen in Table 12, after controlling for review specific effects, respondent specific effects, and clustering the standard errors at the review level, the coefficient for the review source is insignificant ($p = .57$). This provides further support that the true source of the review had no significant impact on the respondents' perceptions of the source of the review being human-generated or machine-generated.

4.2. Purchase intention

For each review, we asked the participants to indicate the likelihood they would purchase the wine either as a gift or for themselves. Again, there was no significant difference in responses from the two types of review source in terms of their purchase likelihood ($p = .18$). We next test via a random-effects regression model, if machine-generated reviews are as informative and engaging as human-written ones by exploring if they induce the same amount of purchase intention among the respondents as human-written ones. Our dependent variable here is the response on a scale of 1 to 7 to the question of respondents' likelihood of purchasing the wine either as a gift or for themselves based on the review they saw. As in the regression of Column (I) of Table 12, the key independent variable is the true review source, which is 1 for human-generated and 0 for machine-generated review. Similarly, we use five hundred dummy variables to control for respondent-specific fixed effects and use review-specific random effects to control for review-specific effects.

Column (II) of Table 12 reports the results of this regression. As can be seen, the coefficient for review source is statistically insignificant ($p = .16$). This result adds further support to the claim that the true source of the review had no significant impact on the respondents' purchase intention of the wine.

While overall purchase intent does not depend on the mode of expression of the review, it is possible, as explored by Luo et al. (2019), that there could be specific patterns within the wine attributes where human-generated reviews drive greater purchase intention than machine-generated reviews. To check for any such systematic differences among the attributes of wines, we have conducted the following analysis. We computed the average rating for the purchase intention question for each wine across all participants when they saw the human-written and machine-generated review. We then selected those wines for which the average purchase intention is higher when participants saw the human-written review rather than the machine-generated review. Of the 300 wines in our data, this is the case for 148 wines. The machine-generated review has the higher average purchase intention for 140 wines and the average purchase intention scores are tied for 12 wines. We next compared the attributes of the 148 wines that have higher average purchase intention with human-written reviews with the 140 wines where the machine-generated reviews have higher average purchase intention. The hypothesis tests

Table 12
Panel Regression with Review-Level Random Effects and Individual-Level Fixed Effects.

	(I. Turing Test)		(II. Purchase Intention)	
	Mean	SE	Mean	SE
Dependent variable	How much (out of 100 points) Mturk respondents thought it was a machine- generated review		How likely (on a scale of 1 to 7) Mturk respondents were to buy the wine either as a gift or for themselves	
Review source is human (dummy variable)	0.409	0.726	−0.039	0.027
Individual specific effect	FE		FE	
Review specific effect	RE		RE	
Number of individuals	501		501	
Number of reviews	600		600	
N	7,014		7,014	
R^2	0.167		0.4217	

for the means of price, rating for the wine (score out of 100), and alcohol by volume (ABV) content are presented in Table 13. As can be seen, while there is marginal significance in the price differential ($p = 0.08$), there is no significant difference in the rating or alcohol content across the two types of wines. This analysis provides further evidence that there was no particular pattern among the wines for which participants had more purchase intent when they saw a human-written review compared to a machine-generated one.

5. Review Synthesis Use-Case Study

Our first use-case addresses the machine writing of an expert review and the degree to which a machine-written review is similar to one that is written by a human and the effects thereof. In this section, we conduct a second use-case study: the capability of the Transformer architecture to be used to generate from a corpus of multiple reviews of individual products a "synthesis" that faithfully integrates the content of the reviews for any single product. We then test the quality of a review synthesis with seven hundred and fifty MTurk participants.

5.1. Review synthesis creation

In the wine review dataset each wine (and its metadata) is associated with a single (expert) review and the proposed task was to generate an expert review. For the review synthesis task, we are in a different setting: the dataset is assumed to be a collection of products, each of which multiple reviews. Oddly enough, another class of alcoholic beverages is a good and available data source: a collection of online user-written beer reviews (https://snap.stanford.edu/data/web-RateBeer.html), consisting of a total of approximately 2.9 million reviews written for 110,419 unique beers.

The beer reviews were written by users rather than professionals and so the reviews display greater variation (in length, language, style, etc.) than the wine reviews. Thus, for training, we culled the set to impose some consistency. We first selected all reviews of length between 20 and 35 words.[12] From this set we then discarded any reviews for a beer which had fewer than ten reviews. This left us with 14,286 unique beers. From each of these we selected ten random reviews out of those available for a total of 142,860 reviews in our dataset.

As in the wine dataset the text of each review was paired with metadata: beer name, beer alcohol by volume (ABV), beer style, and five reviewer ratings. Beers were rated on appearance (out of 5), aroma (out of 10), palate (out of 5), taste (out of 10), and also given an overall rating (up to 20). The Transformer model was initialized with the same parameters as in the case of the wine review study. In this case the training was done on 90% of the reviews in our dataset and the remaining 10% is used as development data to determine when to stop training the model. Training stops when the loss stops improving on this development data. This sets the model parameters.

We then used the trained model to provide a review synthesis for each of 100 randomly selected beers, each with 10 individual reviews, for a grand total of 1000 reviews for 100 beers in the dataset. The overall rating of each beer used for encoding was the average score for that beer in the ten reviews in our data. We then used some automatic filtering to identify the machine-written synthetic reviews to test with our MTurk participants. This automatic filtering code went through the 100 machine-written review syntheses and removed those which:

[12] We use only reviews between 20–35 words as a way to add consistency to the data since these reviews were written by online users (rather than paid experts) and hence display far less consistency in their length, style, and coherence. We thus eliminate possibilities such as reviews being either single words (for example "Draft" and "good" both appear as reviews in the raw data) or a set of rambling stories.

Table 13
Comparison of Wine Attributes for Wines Where the Average Purchase Intent is Higher for Human-Written Reviews.

Attribute	Wines where the average purchase intent is higher for human-written review			Wines where the average purchase intent is higher for machine-generated review			Hypothesis test for difference of means	
	Mean	SD	n	Mean	SD	n	t-stat	p-value
Price ($)	30.269	37.131	148	37.131	41.423	140	−1.715	0.088
Score (out of 100)	87.811	2.963	148	87.793	3.255	140	0.049	0.961
Alcohol by volume (%)	0.135	0.012	148	0.137	0.01	140	−1.116	0.266

1. Were an exact match of a review in the data set,
2. Repeated the same word with no other words in between, and
3. Repeated the same two-word phrase anywhere within the review.

We then calculate a "matchScore" for all reviews. This attempts to measure the extent to which the machine-written synthesis captured the most important words from the 10 human reviews in our data. We do this as follows: Treating all 10 human reviews for a beer as a single document, and the set of all of these documents in our data as the corpus, we calculate term frequency-inverse document frequency (TF-IDF)[13] a measure of how important particular words are to a document relative to the whole corpus. We then sum, for each word in our generated review, the TF-IDF score of the word in the 10 human reviews for the same beer. It is this sum of TF-IDF scores that we call "matchScore", which ranges from 0 to 1.20. We choose all syntheses whose products have a matchScore $\geq$ 0.10. Our matchScore and filtering criteria result in sixty-nine review syntheses (out of 100) for evaluation.

5.2. Experimental evaluation of review synthesis with human respondents

In this experimental evaluation with human respondents, we have two goals. One is to test the degree to which machine-written syntheses appear to capture the essence of the set of reviews written by human beings who are beer enthusiasts. The other is to test whether it makes a difference to the reader if they know the synthesis is written by a machine rather than a human versus providing no such information to the reader of the synthesis. Our study puts respondents in the shoes of online shoppers in a situation where they see ten reviews for a beer and a corresponding synthesis and asks them the extent to which the review synthesis captures the ideas presented in the ten reviews for the beer. Our study involves a three-cell-(review synthesis conducted by a beer enthusiast/machine/control) between-subjects design. Appendix C provides the survey used in this study.

The study was completed by seven hundred and fifty respondents from the U.S. Each respondent earned a reward of $1.50. Overall, 46.00% reported being female and 53.47% male, with 70.27% having bachelor's degree or higher, and 56.67% with household income $50,000 or higher. Each study respondent saw ten reviews (written by ten different beer enthusiasts) for each of three randomly selected beers from a set of sixty nine beers for which our Transformer-based network had generated a review synthesis for each of the sixty-nine beers. Please see Web Appendix for the set of six hundred ninety reviews (ten reviews for each of the 69 beers) written by beer enthusiasts (Web Appendix WA.3) and a corresponding set of sixty-nine review syntheses "written" by our Transformer architecture (Web Appendix WA.4) for the respective sixty-nine beers. Note that the set of human-written reviews and the set of machine-written review syntheses are based on the same set of sixty-nine beers.

A third of the respondents were told that the review synthesis was done by a machine; another third were told that the review synthesis was done by a beer enthusiast; and for the remaining two hundred fifty respondents in the control condition no information was provided about who wrote the review synthesis. We asked the participants to read the corresponding ten reviews carefully for each of the three beers assigned to them. The reviews were followed by the review synthesis and the (same) participants were asked to indicate the extent to which they thought the review synthesis captured the ideas presented in the ten reviews using a 1–7 Likert scale where 1 indicates "does not capture the ideas at all" and 7 for "captures the ideas extremely well". Thus, 4 stands for the neutral response.

Overall, about 75% of the participants provided a rating of 4 or higher, thus indicating that the review syntheses did indeed capture the ideas in the reviews well. The mean score of the response is 4.67 with a standard error of 0.03. We contend that the review synthesis task is performed well if the average rating by the participants statistically significantly exceeds the neutral response of 4. We conduct a hypothesis test for this mean score and find this to be the case ($p < 0.001$). In fact, we found that the average score is significantly higher than 4.5 ($p < .01$). This indicates that our Transformer model is successful at taking the reviews of a product as input and generating a synthesis review for that product.

Note that while the study in Section 4 attests to the fact that (our) machine-written "expert" reviews are indistinguishable from human-written expert reviews and also induce the same purchase intention as human-written ones, it still leaves

[13] For more details, please see https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction, https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html.

open the question of the machine-written review's credibility. That is, the indistinguishability of the author of the review does not imply that the machine-written review of a given product is actually relevant to that product. The results we show above that the machine is capable of generating a review synthesis that captures well the ideas present in several human-written reviews speaks positively to the credibility of machine-generated reviews.

Further, we explore the results from our experimental manipulation. A one-way ANOVA showed a significant impact ($p < .01$) of our experimental manipulation (review synthesis written by a beer enthusiast/machine/control) on how well the respondents thought the synthesis captured the ideas presented in the reviews (average score for machine, beer enthusiast, and control conditions were 4.60, 4.83, and 4.58 respectively with corresponding standard deviations 1.60, 1.48, and 1.59). Interestingly, there was no significant difference in the mean responses between the control condition versus the review synthesis done by a machine ($p = 0.83$). Equally interesting, there was a significant difference between the control/machine condition versus the beer enthusiast condition ($p < .001$). This is consistent with prior research (Luo et al., 2019) where consumers showed a negative reaction when it was revealed to them that the chatbot was a machine and not an actual human being. This provides evidence that from a consumer's perspective, while the reviews were either automatically generated by a machine or providing no information about the source, there were no differences with respect to the perceived quality of the review synthesis. The significant difference in respondents' evaluation of the two types of sources of review synthesis, i.e., written by a human being versus generated by a machine, has important ethical implications for firms such as Amazon. We discuss this issue more in the next section.

6. Discussion

In this paper, we address the overarching question: "To what extent can a machine learn to write a review that is as engaging, informative, and appropriate as a human-written review?" We consider this question in two different contexts: (1) The expert review and (2) The synthesis of multiple reviews for a single product. In the former case, we compare human- versus machine-written reviews in the spirit of a Turing Test and show that the machine generally succeeds in writing human-like reviews. In the latter, we use human evaluation to show that the review syntheses effectively capture the ideas in the various reviews. Both contexts make use of a Transformer-based architecture.

For exploring expert reviews, we work with a data set of wine reviews, while for the synthesis challenge we work with a data set of beer reviews. In these settings we demonstrate that our deep learning approach can – with suitable and sufficient training data – (1) learn to produce expert reviews that are similar to those written by the experts in the training set and (2) learn to produce an informative synthesis of multiple reviews. Our work here is easily generalized or adapted to similar data contexts that could exist for other products. We test the readability and utility (in terms of purchase intention) through analyses of MTurk surveys. We find no significant difference in respondents' identification of whether a review was written by a machine or human being. We thus show that machines can indeed learn to write "human-quality" reviews, with similar purchase intentions, so that at the very least they give reviewers a significant assist (first drafts) in writing such text. This issue is relevant for consumers and retailers alike. We see our work as a first step in the use of machines for writing reviews.

We suggest three related possible applications of our model and approach and provide directions for future research. First, the machine-generated output could either be used as is or as a near-to-complete draft for a retailer to share with consumers. For example, in the context of wine reviews, it can be expensive to produce high quality wine reviews. The number of wine reviews produced by top publications is significant. *Wine Spectator* and *Wine Enthusiast* in total create nearly 40,000 reviews per year in recent years (see Fig. 6). These reviews are written by professionals and *Wine Enthusiast* suggested a top wine writer could draw an annual salary of up to $150,000.[14] These magazines likely also employ copyeditors and proofreaders, positions with a mean annual salary of approximately $42,000.[15]

Researchers have noted that positive reviews in publications increase sales (Friberg & Grönqvist, 2012), and suggested that detailed information about wines is valued by consumers making a purchasing choice (Chaney, 2000; Danner, Johnson, Ristic, Meiselman, & Bastian, 2017). Additionally, reading information about a wine before drinking it appears to actually increase the satisfaction of the imbiber (Danner et al., 2017; Siegrist & Cousin, 2009). We envision that with a new wine, a wine expert can first plug in the wine attributes into the model and generate a draft of the wine review, then the wine expert can taste the wine and modify the machine-written draft if needed. A similar application lies in the movie industry where experts are needed to write reviews for new movies (Liu, 2006).

Apart from the needs of wine publications, wineries may also be interested in generating descriptions of wines. Wineries usually include some information on the labels, and research suggests that an "elaborate taste description" is one of the features most valued by consumers (Mueller, Lockshin, Saltman, & Blanford, 2010). As previously discussed, the presence of these descriptions likely increase consumers' likelihood of purchasing a wine and their enjoyment (Danner et al., 2017; Siegrist & Cousin, 2009). This means the wineries have incentive to generate descriptions for their labels, and those who do also have costs associated with this task. Once again, the first draft provided by the machine could be a useful assist.

A second major use case is the synthesis of the sentiments expressed in large numbers of reviews about a given topic. We have used the Transformer framework in the context of beer reviews to take as input, the various attributes of the product

[14] https://www.winemag.com/2001/05/01/careers-in-wine/.
[15] https://www.bls.gov/oes/current/oes439081.htm.

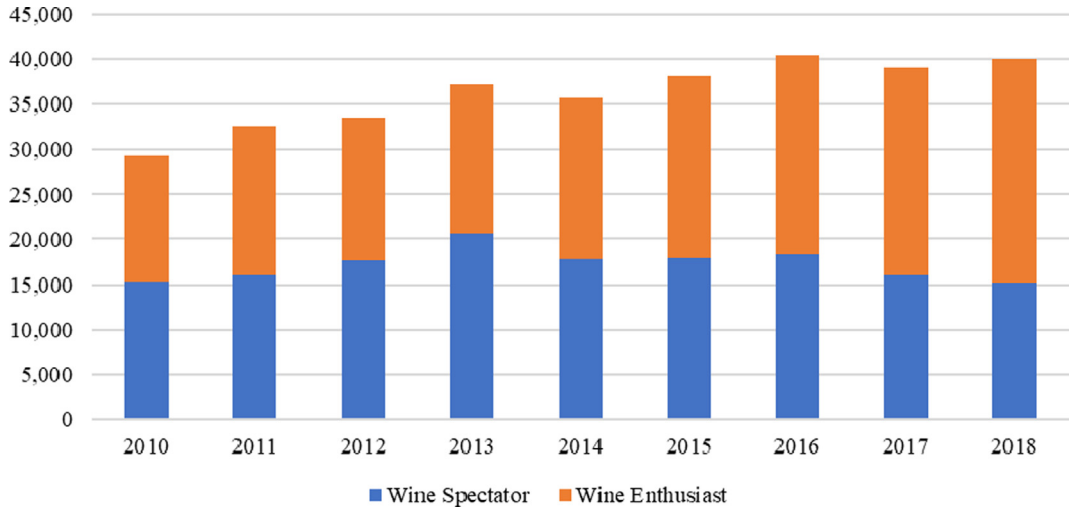

Fig. 6. The number of reviews published by Wine Spectator and Wine Enthusiast online during each year. (Data collected from www.winemag.com and www.winespectator.com.)

along with its reviews, and generate a review "synthesis" able to capture the main ideas present in all the reviews. More generally, this approach can be used to address the issue of having many reviews at retailers such as Amazon where products can have thousands of reviews making it a challenge to present this review information succinctly to customers. Retailers could then display this review synthesis on the product page to optimize screen real estate. Such a review synthesis could improve customer experience by significantly reducing search costs incurred in acquiring information about a product and making the right purchase decision.

We intend our study to be a proof of concept for the idea of using Transformer models to generate a review synthesis. Our study has limitations that future research can address. In the review synthesis use case, we utilize measures of similarity between machine-generated review syntheses and their corresponding human-written reviews only for filtering between the various machine-generated reviews. Future studies can explore incorporating this similarity into the objectives of the model. Additionally, future research could examine how our Transformer model can be modified so that it could run through, say, thousands of reviews on websites like Amazon.com for a product, and create a gestalt review for the product on customized dimensions. For example, there are over 55,000 reviews for a microfiber bed sheet set on Amazon[16] where thousands of reviews may be separated along the following word mentions: "fitted sheet", "super soft", "wrinkle free", "thread count", "good quality", "great value", etc. A modification of the Transformer model may be able to (machine-) read through the reviews under each category of word mentions and create a summary review for each dimension. This would be beneficial to consumers because no consumer will be able to devote their resources (time and effort) to go through thousands of reviews under each dimension and come up with an overall gestalt review that summarizes the content across all these reviews. We leave this issue for future research. Future studies can also explore in detail the effect of using such an algorithm on business performance.

Third, going beyond the direct application of our use cases, a broader potential application of our technology exists in contexts where text description of a product is needed and only the metadata of that product is available. This technology can be used by small independent sellers on platforms such as Etsy to automatically generate text descriptions for the hundreds or thousands of products that they may offer on their store. In other words, there are many products on platforms such as Amazon.com or Etsy.com which have very few or no reviews and many consumers would be hesitant to purchase products without any reviews (https://sellercentral.amazon.com/forums/t/no-reviews/410368). For products with either few or no reviews, future research could examine how our Transformer model may embed learning from reviews of similar products and provide a machine-generated "review" or a "product description" for such products. With this approach one can envision the usefulness of providing relevant information to consumers to aid their decision-making. At the same time, as seen in Anderson and Simester (2014) we now know that sometimes reviews are posted without an actual purchase, and this can be used as a form of deception. Machine-writing is – like all technologies – Janus-faced in its potential uses.

To that end, further research will continue to unpack the debate with regard to the benefits, perils, and ethics of machine-writing. In our study, we make a small empirical contribution to this kind of work as we test the effect of the source that generates the review synthesis. We find that the participants are significantly more likely to agree that the synthesis captures the ideas present in the reviews well if they were told that the synthesis was generated by a beer enthusiast rather than a machine or if told nothing at all (the neutral condition). This result raises important ethical concerns since it appears

[16] https://www.amazon.com/gp/product/B00NQDGCW8/?ots=1&slotNum=14&imprToken=d4bb8e18-436d-203b-f3b&tag=1000rev-20.

that retailers could get greater acceptance of machine-generated reviews by attributing them to humans. We look forward to future studies that will investigate in more detail the ethical complexities of machine-generated content meant for human consumption.

Finally, we believe that our model architecture could be useful in a context like Yelp!, say in the particular category of restaurants where there is a lot of metadata and (often) many reviews per restaurant. The critical aspect here is – even more so than our beer example – the heterogeneity of the review text. The outputs generated by the model will only be as good as the training data and a filtering of reviews akin to what we do in Section 5 would be necessary. Thus, while work remains to be done, empirically studying the role of automatic generation of content and online reviews in particular seems viable and worthy of the effort required to more fully understand it.

Acknowledgements

We thank the Editor, Area Editor, and two anonymous IJRM reviewers for their helpful comments on earlier drafts of this paper. We also thank Anja Lambrecht, conference participants at the 2020 Conference on Artificial Intelligence, Machine Learning, and Business Analytics, and 2020 Northeast marketing consortium attendees for their comments on previous versions of the paper. All authors contributed equally to this paper.

Appendix A

A.1 Wine MTurk survey

In this 10-minute survey you will be asked to reflect on your experience as a consumer of reviews and detailed product descriptions. Often when you are shopping for a product online, e.g., a wine, you will find a review or a detailed product description for that product. Here you will be presented with a set of descriptions for different wines. Some of the descriptions that you see are generated by human beings and some of them were written by a machine using an artificial intelligence algorithm. Please read each description of a wine carefully and after reading each of the descriptions, please answer the following question.

Q1) In my opinion (please check one box):

This review was written by a human being ____

This review was written by a machine ___

I am unable to tell whether this review was written by a human being or a machine ___

Q2) Based on your reading, please divide 100 points between the following two options. The number of points you give should reflect your guess in terms of whether the review was written either by a machine or a human being – so, you can give 0 points to the machine and 100 points to the human being if you are sure that the review was written by a human being; you can also give 100 points to the machine and 0 points to the human being if you are sure that the review was written by a machine; or you can allocate 50/50 between the two if you think it is equally likely that it could have been written by a human being or a machine; etc.

The review is written by a machine

The review is written by a human being

Q3) 7 point Likert from strongly disagree to strongly agree

Assume that the wine you just read is available at a reasonable price at your favorite store. "Based on the description of the wine that I just read, I am likely to buy the wine either for myself or as a gift."

Q4) In your opinion, do you see any benefits to consumers if machines are able to learn from people on how to write online product reviews?

Which state do you primarily reside in? _____(pull down menu)

Zipcode___

Gender__

Age___

Education ____

Income ____

How interesting did you find wine descriptions presented in this study?

In your opinion, how realistic do you think were the wine descriptions presented in this study?

A.2 Data descriptives

Tables A1–A4.

Table A1
Question 1. Review Source.

	In my opinion. . .			
	This review was written by a human being	This review was written by a machine	I am unable to tell whether this review was written by a human being or a machine	Total
Review source: human	1,946	1,164	390	3,500
Percent	55.60%	33.26%	11.14%	100.00%
Review source: machine	1,966	1,169	379	3,514
Percent	55.95%	33.27%	10.79%	100.00%

Table A2
Question 2. Review Source.

	Based on your reading, please divide 100 points between the following two options: (the review is written by a machine)			
	N	Mean	Median	SD
Review source: human	3,500	44.7	50	26.52
Review source: machine	3,514	44.71	50	27.22
	(the review is written by a human)			
	N	Mean	Median	SD
Review source: human	3,500	55.3	50	26.52
Review source: machine	3,514	55.29	50	27.22

Table A3
Question 3. Likely to Buy.

	Based on the description of the wine that I just read, I am likely to buy the wine either for myself or as a gift (1 to 7)			
	N	Mean	Median	SD
Review source: human	3,500	3.36	4	1.13
Review source: machine	3,514	3.33	4	1.13

Table A4
Other Questions.

	N	Mean	Median	SD
How interesting did you find the wine descriptions presented in this study?	501	5.49	5	1.19
How realistic do you think the wine descriptions were that you read?	501	5.68	6	1.12

Appendix B. A bag-of-words comparison of human- and machine-written reviews

Using the Python package NLTK (https://www.nltk.org/), we built a bag-of-words (BoW) model. The most frequently occurring words in a corpus describe the context and semantic style of that corpus. If two corpuses have similar semantic styles (i.e., in our case, the three hundred human-written and the three hundred machine-generated reviews), then there should be similarity among the most frequently occurring words across the two corpuses. As highlighted below in italics, 23 of the top 30 words are in common across the two corpuses, suggesting the semantic styles are similar across human-written and machine-generated reviews. Table B1

Appendix C. Beer MTurk survey

In this 10-minute survey, you will be asked to reflect on your experience as a consumer of reviews and detailed product descriptions. Upon completion and approval of your work, you will be paid $1.50. Please take your time in carefully reading all instructions. Please note that there are no right or wrong answers. On the next page, you will indicate whether you are willing to participate in the study after reading a "Consent Form."

Often when you are shopping for a product either in a store or online, you will find a review for that product online. Here you will first be presented with a set of 10 reviews for a beer written by ten different beer enthusiasts. Once you carefully

Table B1
The Most Frequently Occurring Words in Human-Written and Machine-Generated Reviews.

Human-written reviews		Machine-generated reviews	
Word	Frequency	Word	Frequency
wine	208	*wine*	233
flavors	150	*flavors*	142
fruit	117	*full*	90
finish	85	*fruit*	89
aromas	75	*aromas*	78
acidity	73	*finish*	78
palate	72	*dry*	76
cherry	63	*palate*	75
black	60	bodied	75
tannins	55	*drink*	73
ripe	54	*cherry*	67
sweet	49	*black*	66
red	49	*nose*	60
dry	47	*berry*	50
drink	47	touch	50
spice	46	*tannins*	49
berry	44	*red*	48
rich	44	*ripe*	48
oak	39	*blend*	47
vanilla	39	*soft*	44
nose	38	*rich*	42
full	36	*spice*	42
notes	36	*acidity*	42
soft	33	*plum*	39
blend	31	dark	39
fruits	30	light	35
fresh	30	white	35
good	30	*fruits*	34
plum	30	color	34
citrus	29	bit	32

read through the set of 10 reviews for a beer, we will then present to you a Review Synthesis to capture the ideas from the beer reviews. The Review Synthesis was put together by a beer enthusiast/machine/no such information provided. You will then be asked to what extent the Review Synthesis captures the ideas from the corresponding ten beer reviews, where 1 indicates Does Not Capture the Ideas at All and 7 indicates Captures the Ideas Very Well. We will begin with a set of 10 reviews for Beer A plus the review synthesis and then your task, followed by a similar sequence for Beers B and C.

In your opinion, to what extent do you think the "Review Synthesis" captures the ideas presented in the corresponding ten reviews above it?

(7 point Likert from "Does Not Capture the Ideas at All" to "Captures the Ideas Very Well")

Which state do you primarily reside in? _____(pull down menu)

Zipcode___

Gender__

Age___

Education ____

Income ____

How interesting did you find wine descriptions presented in this study? (7-point: Not at all interesting/Very Interesting) Mean = 5.35 (S.D. = 1.54)

In your opinion, how realistic do you think were the reviews and their descriptions presented in this study? (7 point: Not at all realistic/Very realistic) Mean = 5.91 (S.D. = 1.16)

Appendix D. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijresmar.2022.02.004.

References

Anderson, E. R., & Simester, D. I. (2014). Reviews without a purchase: Low ratings, loyal customers, and deception. *Journal of Marketing Research, 51*(June), 249–269.

Bahdanau, D., Cho, K. & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint*: 1409.0473.

Basuroy, S., Chatterjee, S., & Ravid, S. A. (2003). How critical are critical reviews? The box office effects of film critics, star power, and budgets. *Journal of Marketing, 4*, 103–117.
Berger, J., Sorensen, A. T., & Rasmussen, S. J. (2010). Positive effects of negative publicity: When negative reviews increase sales. *Marketing Science, 29*(5), 815–827.
Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J., ... Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics, 16*(2), 79–85.
Camacho, N., De Jong, M., & Stremersch, S. (2014). The effect of customer empowerment on adherence to expert advice. *International Journal of Research in Marketing, 31*(3), 293–308.
Carlson, K., Riddell, A., & Rockmore, D. (2018). Evaluating prose style transfer with the Bible. *Royal Society Open Science, 5*(10) 171920.
Cavanaugh, L. A., Bettman, J. R., & Luce, M. F. (2015). Feeling love and doing more for distant others: Specific positive emotions differentially affect prosocial consumption. *Journal of Marketing Research, 52*(5), 657–673.
Chakraborty, I., Kim, M., & Sudhir, K. (2021). EXPRESS: Attribute Sentiment Scoring with Online Text Reviews: Accounting for Language Structure and Missing Attributes. *Journal of Marketing Research*. 00222437211052500.
Chaney, I. (2000). A comparative analysis of wine reviews. *British Food Journal, 102*(7), 470–480.
Chevalier, J. A., Dover, Y., & Mayzlin, D. (2018). Channels of impact: User reviews when quality is dynamic and managers respond. *Marketing Science, 37*(5), 688–709.
Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research, 43*(3), 345–354.
Chintagunta, P. K., Gopinath, S., & Venkataraman, S. (2010). The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science, 29*(5), 944–957.
Cho, K., Bart Van, M., Caglar G., Dzmitry B., Fethi B., Holger S., & Yoshua B. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint*: 1406.1078.
Conroy, J. M., & O'leary, D. P. (2001). Text summarization via hidden Markov models. In *Proceedings of the 24th annual international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 406–407).
Danner, L., Johnson, T. E., Ristic, R., Meiselman, H. L., & Bastian, S. EP. (2017). "I like the sound of that!" Wine descriptions influence consumers' expectations, liking, emotions and willingness to pay for Australian white wines. *Food Research International, 99*, 263–274.
Decker, R., & Trusov, M. (2010). Estimating aggregate consumer preferences from online product reviews. *International Journal of Research in Marketing, 27*(4), 293–307.
Forbes (2020). 3 New Ways Artificial Intelligence Is Powering The Future Of Marketing, accessed at [https://www.forbes.com/sites/cathyhackl/2020/06/28/3-new-ways-artificial-intelligence-is-powering-the-future-of-marketing/?sh=f77577b1a96e] on February 5, 2020.
Friberg, R., & Grönqvist, E. (2012). Do expert reviews affect the demand for wine? *American Economic Journal: Applied Economics, 4*(1), 193–211.
Gabel, S., Guhl, D., & Klapper, D. (2019). P2V-MAP: Mapping market structures for large retail assortments. *Journal of Marketing Research, 56*(4), 557–580.
Ghose, A., Ipeirotis, P., & Li, B. (2012). Designing ranking systems for hotels on travel search engines by mining user-generated and crowd-sourced content. *Marketing Science, 31*(3), 493–520.
Haenlein, M., Kaplan, A., Tan, C. W., & Zhang, P. (2019). Artificial intelligence (AI) and management analytics. *Journal of Management Analytics, 6*(4), 341–343.
Hennig-Thurau, T., Wiertz, C., & Feldhaus, F. (2015). Does Twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies. *Journal of the Academy of Marketing Science, 43*(3), 375–394.
Hu, M., Dang, C., & Chintagunta, P. K. (2019). Search and learning at a daily deals website. *Marketing Science, 38*(4), 609–642.
Huang, M.-H., & Rust, R. (2018). Artificial Intelligence in Service. *Journal of Service Research, 21*(2), 155–172.
Huber, J., Kamakura, W., & Mela, C. F. (2014). A topical history of JMR. *Journal of Marketing Research, 51*(1), 84–91.
Kannan, P. K., & Li, H. A. (2017). Digital Marketing: A Framework, Review, and Research Agenda. *International Journal of Research in Marketing, 34*(1), 22–45.
Kay, M. (1984). Functional unification grammar: A formalism for machine translation. In *Proceedings of the 10th International Conference on Computational Linguistics* (pp. 75–78).
Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... Dyer, C. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions* (pp. 177–180).
Kopalle, P. K., Gangwar, M., Kaplan, A., Ramachandran, D., & Reinartz, W. (2021). Artificial Intelligence (AI) Technologies in Global Marketing: Current Trends and Future Research Opportunities. Forthcoming. *International Journal of Research in Marketing*..
Kübler, R. V., Colicev, A., & Pauwels, K. H. (2020). Social media's impact on the consumer mindset: When to use which sentiment extraction tool? *Journal of Interactive Marketing, 50*, 136–155.
Lee, T. Y., & Bradlow, E. T. (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research, 48*(5), 881–894.
Li, X., Shi, M., & Wang, X. S. (2019). Video mining: Measuring visual information using automatic methods. *International Journal of Research in Marketing, 36*(2), 216–231.
Li, Y., & Xie, Y. (2020). Is a picture worth a thousand words? An empirical study of image content and social media engagement. *Journal of Marketing Research, 57*(1), 1–19.
Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing, 70*(3), 74–89.
Liu, L., Dzyabura, D., & Mizik, N. (2020). Visual listening in: Extracting brand image portrayed on social media. *Marketing Science, 39*(4), 669–686.
Liu, X., Singh, P. V., & Srinivasan, K. (2016). A structured analysis of unstructured big data by leveraging cloud computing. *Marketing Science, 35*(3), 363–388.
Ludwig, S., De Ruyter, K., Friedman, M., Brüggen, E. C., Wetzels, M., & Pfann, G. (2013). More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing, 77*(1), 87–103.
Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science, 38*(6), 937–947.
Luo, X., Zhang, J. J., Gu, B., & Phang, C. (2017). Expert blogs and consumer perceptions of competing brands. *MIS Quarterly, 41*(2), 371–395.
Ma, L., & Sun, B. (2020). Machine learning and AI in marketing–Connecting computing power to human insights. *International Journal of Research in Marketing, 37*(3), 481–504.
McIntyre, N., & Lapata, M. (2009). Learning to tell tales: A data-driven approach to story generation. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (Vol. 1, No. 1, pp. 217–225). .
Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies* (pp. 746–751).
Moe, W. W., & Schweidel, D. A. (2012). Online product opinions: Incidence, evaluation, and evolution. *Marketing Science, 31*(3), 372–386.
Moe, W. W., & Trusov, M. (2011). The value of social dynamics in online product ratings forums. *Journal of Marketing Research, 48*(3), 444–456.
Moon, S., & Kamakura, W. A. (2017). A picture is worth a thousand words: Translating product reviews into a product positioning map. *International Journal of Research in Marketing, 34*(1), 265–285.
Mueller, S., Lockshin, L., Saltman, Y., & Blanford, J. (2010). Message on a bottle: The relative influence of wine back label information on wine choice. *Food Quality and Preference, 21*(1), 22–32.
Murray, K., Kinnison, J., Nguyen, T. Q., Scheirer, W., & Chiang, D. (2019). Auto-sizing the transformer network: Improving speed, efficiency, and performance for low-resource machine translation. *arXiv preprint*:1910.06717.
Nallapati, R., Zhou, B., Gulcehre, C. & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence runs and beyond. *arXiv preprint*: 1602.06023.
Nam, H., Joshi, Y. V., & Kannan, P. K. (2017). Harvesting brand information from social tags. *Journal of Marketing, 81*(4), 88–108.

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science, 31*(3), 521–543.
Netzer, O., Lemaire, A., & Herzenstein, M. (2019). When words sweat: Identifying signals for loan default in the text of loan applications. *Journal of Marketing Research, 56*(6), 960–980.
Onishi, H., & Manchanda, P. (2012). Marketing activity, blogging and sales. *International Journal of Research in Marketing, 29*(3), 221–234.
Proserpio, D., & Zervas, G. (2017). Online reputation management: Estimating the impact of management responses on consumer reviews. *Marketing Science, 36*(5), 645–665.
Reinstein, D. A., & Snyder, C. M. (2005). The influence of expert reviews on consumer demand for experience goods: A case study of movie critics. *The Journal of Industrial Economics, 53*(1), 27–51.
Schwartz, E. M., Bradlow, E. T., & Fader, P. S. (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science, 36*(4), 500–522.
Siegrist, M., & Cousin, M. (2009). Expectations influence sensory experience in a wine tasting. *Appetite, 52*(3), 762–765.
Singh, S. N., Hillmer, S., & Wang, Z. (2011). Efficient methods for sampling responses from large-scale qualitative data. *Marketing Science, 30*(3), 532–549.
Syam, N., & Sharma, A. (2018). Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice. *Industrial Marketing Management, 69*, 135–146.
Timoshenko, A., & Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science, 38*(1), 1–20.
Vana, P., & Lambrecht, A. (2021). The Effect of Individual Online Reviews on Purchase Likelihood. *Marketing Science, 40*(4), 708–730.
Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A. N., Gouws, S., . . ., Uszkoreit, J. (2018). Tensor2tensor for neural machine translation. *arXiv preprint*: 1803.07416.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems. s.l.:s.n* (pp. 5998–6008).
Villarroel Ordenes, F., Ludwig, S., De Ruyter, K., Grewal, D., & Wetzels, M. (2017). Unveiling what is written in the stars: Analyzing explicit, implicit, and discourse patterns of sentiment in social media. *Journal of Consumer Research, 43*(6), 875–894.
Wu, C., Che, H., Chan, T. Y., & Lu, X. (2015). The economic value of online reviews. *Marketing Science, 34*(5), 739–754.
Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., . . ., Klingner, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint*: 1609.08144v2 [cs.cl].
Xu, W., Ritter, A., Dolan, W. B., Grishman, R., & Cherry, C. (2012). Paraphrasing for style. *Proceedings of COLING, 2012*, 2899–2914.
Zhang, M., & Luo, L. (2021). Can Consumer-Posted Photos Serve as a Leading Indicator of Restaurant Survival? Evidence from Yelp. *SSRN*, working paper, accessed on January 15, 2022 at: https://dx.doi.org/10.2139/ssrn.3108288.
Zhang, M., Sun, T., Luo, L., & Golden, J. (2021). Consumer and AI Co-creation: When and Why Human Participation Improves AI Creation. *SSRN*, working paper, accessed on January 15 2022 at: https://dx.doi.org/10.2139/ssrn.3929070.
Zhao, Y., Yang, S., Narayan, V., & Zhao, Y. (2013). Modeling consumer learning from online product reviews. *Marketing Science, 32*(1), 153–169.
Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing, 74*(2), 133–148.