Creating Synthetic Experts with Generative Artificial Intelligence

Daniel M. Ringel
dmr@unc.edu

Working Paper
December 5, 2023

Abstract

Classification is paramount in today's data-rich environment as marketers increasingly depend on machine learning to distill intelligence from vast amounts of unstructured text such as news, reports, and social media. Modern classification models swiftly identify constructs of interest, such as sentiment or product categorizations to inform research and managerial decision-making. Training an effective classification model requires many correctly labeled examples. While simple constructs can be labeled via crowdsourcing, more abstract and multifaceted constructs necessitate expert labelers—a scarce resource. We study whether generative AI, specifically ChatGPT4, can replace domain experts for identifying a central marketing construct in microblogs: brands' marketing mix. We find that, unlike crowdsourced labels, those generated by ChatGPT4 are in high agreement with expert labels. We overcome ChatGPT4's proprietary nature, slow speed, high cost, and limited reproducibility by approximating it with an open-source model that is fine-tuned on ChatGPT4's labels. The created *Synthetic Expert* exhibits near-parity with ChatGPT4 in terms of expert agreement, is highly scalable, free from third-party constraints, and produces perfectly replicable results. When paired with sentiment analysis, it reveals different distributions of consumer sentiment across the marketing mix of 699 brands, with substantially varying strengths and weakness among competing brands. Deeper analysis uncovers marketing mix specific topics that consumers raise online. By introducing *Synthetic Twins*, AI-generated replicas of training texts that correspond in idea and meaning to their original counterparts, this research mitigates privacy, confidentiality, and intellectual property concerns for model training and data sharing.

Keywords: Classification, Generative Artificial Intelligence, Marketing Mix

Daniel M. Ringel is an Assistant Professor of Marketing at Kenan-Flagler Business School, University of North Carolina at Chapel Hill

1. Introduction

The emergence of machine learning technologies revolutionized the business landscape, spawning countless innovations (Ban and Rudin 2019; Wu et al. 2020; Burnap et al. 2023). In particular, the ability to distill actionable insights from masses of unstructured text—such as news articles, firm reports, and social media—represents a potent capability that offers competitive advantages (Avramov et al. 2023; Yoganarasimhan 2020; Zhang et al. 2023). Among the myriad tools and techniques at analysts' disposal, classification models have garnered significant attention due to their versatility and potency (Chakraborty et al. 2022; Frankel et al. 2022; Hartmann et al. 2019). This pivotal technique forms the focal point of this research.

Classification models analyze large volumes of data to identify patterns and categorize them into predefined classes, thereby helping firms harness the potential of their unstructured information assets. By identifying constructs of interest such as sentiment, arguments, type of rhetoric, or relevant product categories, classification models unlock hidden potentials that managers can leverage to drive business growth (Abbasi et al. 2019; Lawrence and Reed 2020; Şeref et al. 2023). The versatility of classification models extends their utility across sectors and functions, making them instrumental in decision-making processes, customer understanding, and risk mitigation. More than just a technical solution, classification is an analytical imperative—an essential catalyst in ensuring that firms stay agile, informed, and prepared to respond to dynamic market conditions and competitive pressure.

The efficacy of modern classification models relies heavily on their training, which in turn, hinges on the availability of a substantial amount of correctly labeled examples (Hartmann et al. 2023). Here, labeled examples denote datasets wherein each item is assigned a specific class, providing the model with illustrative outcomes learns to predict. For simple constructs such as sentiment, spam, or emotions, analysts can resort to crowdsourced labels (Snow et al. 2008). However, more complex constructs—those imbued with elevated levels of abstraction, ambiguity, and multifaceted dimensions—require expert labelers with domain knowledge and specialized skills (Hartmann et al. 2023). Unfortunately, experts are typically a scarce resource. This resource limitation introduces a substantial bottleneck in the labeling process, posing

a key problem in deploying modern classification models that can identify fundamental constructs such as the marketing mix, drivers of brand equity, or dimensions of service quality in unstructured data.

To overcome this limitation, we propose to use generative AI—specifically OpenAI's ChatGPT—as a proficient substitute for expert labelers. ChatGPT is a generative pretrained transformer model (GPT) that, given a user prompt, predicts words to formulate a likely answer (Radford et al. 2018). It was trained on a large collection of texts—books, websites, articles, Wikipedia entries—to embody vast amounts of knowledge. ChatGPT received considerable attention from both academia (Van Dis et al. 2023) and industry (Chui et al. 2022) who view it as a transformative technology with a lasting impact on society. Early research finds that GPT4, the underlying model of OpenAI's newest ChatGPT version at the time of writing, can solve novel and difficult tasks, such as mathematics, coding, vision, medicine, law, and psychology (Bubeck et al. 2023).

This research surrogates expert labelers with ChatGPT4 to identify the marketing mix variables that consumers' posts on Twitter pertain to. The marketing mix, or "four Ps of marketing" (product, price, place, and promotion), is an essential part of marketing theory, and one of the most powerful concepts ever developed for executives (Kotler et al. 2012; Shapiro 1985, Kotler et al. 2012). Understanding what part of a brand's marketing mix consumers talk about on social media is important because it guides managers' assessments of their marketing strategy, alerts them to potential risks and opportunities, and identifies which marketing mix levers require attention.

We contend that classifying Tweets regarding marketing mix variables is a complex undertaking. The lack of mutual exclusiveness among the 4 classes (product, price, place, and promotion) in their widespread definition creates ambiguity that labelers must deal with. Although Van Waterschoot and Van den Bulte (1992) improved the mutual exclusiveness of the 4 classes, they did so at the cost of a lower mnemotechnic appeal, complicating their faithful identification by crowdsourced amateurs.

We find that the labels generated by ChatGPT4 are in high agreement with those of four domain experts (i.e., scholars), whereas crowdsourced labels from Amazon mTurk workers exhibit substantially lower agreement. Despite these promising results, using ChatGPT4 as a surrogate for expert labelers is

subject to multiple limitations that marketers must consider. First, because ChatGPT4 is a proprietary model, firms that use it depend fully on its owner, OpenAI, who controls access, pricing, and capabilities. Second, passing information to a generative AI that is out of a firm's control can expose it to significant confidentiality and data privacy risks (Busch 2023, Daniels 2023). Third, we find that ChatGPT4's responses are not deterministic, which curtails researchers' abilities to replicate their findings, a major concern in the research community (Van Noorden and Perkel 2023). And fourth, generative AI models like ChatGPT4 are still costly and slow such that they do not scale well to the large volumes of data that are typical of production environments.

In response to these limitations, we develop an alternative solution that involves approximating ChatGPT4 with an open-source large language model (LLM). We fine-tune the LLM on ChatGPT4's labels for marketing mix (MMX) variables in thousands of Tweets, giving rise to what we call a "Synthetic Expert". Our synthetic expert not only demonstrates near-parity with ChatGPT4 in terms of expert label agreement, but also exhibits high scalability, full independence, guaranteed replicability, and freedom from third-party constraints.

To demonstrate the benefits of identifying more complex constructs of interest at scale, we disambiguate brand sentiment on Twitter into individual marketing mix variables using our synthetic expert. Our analysis reveals different distributions of consumer sentiment across the marketing mix of 699 brands. We further find that competing brands have different strengths and weaknesses in their marketing mix from consumers perspectives. Deeper analyses of Tweets that pertain to the weakest marketing mix variable (i.e., where sentiment is lowest) of a major fashion brand reveal four distinct topics that alienate consumers. We contend that marketing managers profit from such disambiguation because it helps them focus their efforts on those marketing mix variables that need the most attention.

By leveraging state-of-the-art generative AI as an effective surrogate for human expertise, we invite a paradigm shift in machine learning, which has traditionally relied on expert human annotation for complex classification tasks. The introduction of synthetic expertise opens a wealth of research opportunities, inspiring follow-up studies in diverse domains, such as healthcare for categorizing patient symptoms, law

for legal document classification, finance for extracting insights from complex economic reports, and education for personalized learning experiences. Notably, synthetic experts facilitate access to and sharing of expertise across the broader research community. And unlike current generative AI models, synthetic experts contribute to mitigating the replication crises by guaranteeing reproducibility. To firms, synthetic experts represent a scalable solution for the complex task of comprehending vast data repositories. Societally, introducing synthetic experts can significantly streamline public sector operations by, for instance, aiding in processing and analyzing public opinions on policy matters.

As such, our work contributes to the global discourse on the potential role of generative AI in managerial decision-making, research, and society. We shed light on new possibilities in this exciting frontier and provide hands-on solutions to research and practice: To expedite the adoption of synthetic expertise in complex classification tasks, we make all code required to train synthetic experts with generative AI freely available at www.synthetic-experts.ai and publish our fully trained MMX classifier on https://huggingface.co/dmr76/mmx_classifier_microblog_ENv02.

2. Related Literature

Our research connects the rich literature on text classification with the emerging literature on generative AI in the marketing context. Past research demonstrates the value of accurately classifying text to inform marketing decisions (Berger et al. 2020; Rust et al. 2021). At the forefront of the literature on text classification is sentiment analysis, which involves classifying the evaluative nature of a piece of text regarding its positivity, neutrality or negativity (Kiritchenko et al. 2014). Recent advances in sentiment analysis include paralinguistic classification (Luangrath et al. 2023), measures of certainty (Rocklage et al. 2023), and fine-tuning of LLMs to improve classifier performance (Hartmann et al. 2023).

Although undeniably of great value, sentiment analysis alone cannot fully access the rich information embedded in textual content (Archak et al. 2011). Indeed, as Chakraborty et al. (2022) show, attributing sentiment to constructs of interest—such as which element of a dining experience restaurant reviews refer to—creates deeper insights with immediate managerial implications. Identifying more intricate constructs

in a text is a complex undertaking. For instance, to build a brand reputation index from Twitter posts, Rust et al. (2021) had to develop and validate 11 language dictionaries that capture the dimensions of the value-brand-relationship framework by Rust et al. (2004). Similarly, Suslava (2021) had to construct an elaborate dictionary of corporate euphemisms and augment it with hand-crafted syntactic rules to study the impact of euphemisms on investor reactions.

While dictionary-based approaches for text classification are advantageous because of their transparency, recent research shows that supervised machine learning approaches can significantly improve classification performance (Frankel et al. 2022; Hartmann et al. 2023). Xia and Liu (2022) initial exploration into classifying brand-authored Tweets regarding brands' marketing mix highlights the nascent stage of research on more complex constructs of interest. Acknowledging the limitations of their small training sample, they call for more extensive datasets labeled by domain experts.

Because domain experts are often scarce and expensive (Hartmann et al. 2023), various approaches emerged to make the labeling process more efficient. In particular, active learning, which dates back to the work of Cohn et al. (1994), proved to be effective at streamlining the labeling and model training process. The main idea behind active learning is to select only those examples for labeling (e.g., texts such as Twitter posts) that would improve a classifier's performance most when labels are expensive to obtain, but unlabeled data are abundant (Chen et al. 2023). Like active learning, our proposed use of generative AI makes the labeling process more efficient. However, unlike active learning, which is about *what* to label, we seek to overcome a more fundamental problem: *how* to label more complex constructs in text.

Our approach of surrogating domain experts with OpenAI's ChatGPT4 also falls into the emerging literature on generative AI in the management sciences. Our work follows the call for research on how generative AI can contribute value to research and practice (Chui et al. 2022; Peres et al. 2023; Van Dis et al. 2023). Most relevant to our proposed approach is the study of Horton (2023), who demonstrates that various LLMs from OpenAI respond to economic scenarios in ways consistent with intuition and experience. More broadly, Guo et al. (2023) find that ChatGPT3.5 (the predecessor of ChatGPT4) provides

very similar answers as human experts across various domains. The findings of both studies support our idea to surrogate human experts with generative AI for complex labeling tasks.

In marketing, recent work demonstrates the use of generative AI as a possible substitute for human subjects in market research. Li et al. (2022) query ChatGPT4 for consumer perceptions of automotive brands to generate perceptual maps. Brand et al. (2023) explore the uses and benefits of GPT3 for researchers and practitioners who seek to understand consumer preferences. Both studies exploit model stochasticity as a proxy for heterogeneous consumer responses, which makes them fundamentally different from our work in two regards. First, we seek domain expertise, not consumers' perceptions or preferences. Second, we query a generative AI about well-documented and typically time-invariant constructs, not contemporary market information.

3. Empirical Investigation

We pursue three objectives in our empirical investigation. First, assess the viability of surrogating human experts with generative AI for identifying complex constructs in unstructured text. Second, approximate OpenAI's proprietary ChatGPT4 with a synthetic expert—a scalable, open-source classifier specialized in identifying a complex construct of interest in unstructured text. And third, demonstrate how synthetic experts can generate deeper insights for marketers—at scale, reproducibly, and free of third-party constraints.

The classification task of this study is to determine which of the four marketing mix variables (product, price, place, and promotion) Twitter posts pertain to. We chose the marketing mix because it is at the heart of firms' marketing decisions, and because the construct is sufficiently complex for our investigation (i.e., a multifaceted abstraction). Although central to marketing, we found no publicly available marketing mix classier online (in contrast, there are over 2,100 publicly available sentiment classifiers alone on www.huggingface.co). We chose Twitter posts because social media continues to be a valuable information source for marketing research (e.g., Liaukonytė et al. 2023; Mallipeddi et al. 2022; Schoenmueller et al.

2023), and because their brevity and use of informal language make them a sufficiently challenging baseline for advanced text analysis.

By example of the marketing mix, we seek answers to the following four questions: First (R1), can crowdsourced amateurs correctly identify a complex construct in text? Second (R2), can generative AI correctly identify a complex construct in text? Third (R3), does a task-specific approximation of ChatGPT4 perform equally well as ChatGPT4? And fourth (R4), is the disambiguation of consumer sentiment into brands' marketing mix necessary?

3.1. Data

We use Tweets collected through Twitter's API in this study. Our data comprise Tweets from 2019 to 2021 that mention major brands' Twitter handle, for example:

@nike in "*Best cushioning ever!!! 🤗🤗🤗 my zoom vomeros are the bomb🏃‍♂️💨!!! @nike #run #training*

We include 699 major brands in our investigation—compiled from the list of brands investigated by Dew et al. (2022) and top brands according to YouGov[1].

We built a pool of Tweets by randomly sampling 50 Tweets for each brand and removing duplicate Tweets. We preprocessed all Tweets by removing excessive spaces and breaks, translating HTML entities, and removing URLs. Finally, we drew two mutually exclusive samples from our pool: 1,000 Tweets as validation sample to investigate label agreement among experts, crowdsourced amateurs, and generative AI; 30,000 Tweets as training sample to train a synthetic expert. We stratified the validation sample by user engagement to ensure more meaningful content (i.e., half of the sampled Tweets had at least ten likes, quotes, retweets, replies, or combinations thereof).

3.2. Humans vs. Generative AI in Text Classification

To assess the viability of surrogating experts with generative AI in complex classification tasks, we measure how strongly human labels for the four marketing mix variables in the validation sample agree with those

[1] https://business.yougov.com/product/brandindex

of ChatGPT4. We measure label agreement using Krippendorff's α[2] to control for chance agreement (Hayes and Krippendorff 2007). Additionally, we use machine learning metrics for classification (i.e., precision, recall, and F1-score) to examine type 1 and 2 errors.

We first obtain expert labels for the validation sample. To obtain an objective ground truth and mitigate the risk of label noise (i.e., incorrect labels), four domain experts participated in six workshops to jointly examine, discuss, and label each Tweet of the validation sample. On average, the team of experts labeled 1.4 Tweets per minute. We use the 4,000 expert labels (1,000 Tweets × 4 labels each) as ground truth in our assessment.

Next, we crowdsourced amateur labels from workers on Amazon's Mechanical Turk (mTurk) platform. Our objective was to investigate whether our labeling task necessitated the use of scarce and costly experts. Workers were briefed on the definition of the four Ps of marketing and tasked to identify which of the four Ps of marketing, if any, a presented Tweet pertains to. To avoid worker fatigue, each worker was limited to labeling 50 Tweets. We used attention checks to ensure higher label quality. We recruited only mTurk Masters (workers with a record of superior performance) and had at least a 90% HIT approval rate (HITs are Human Intelligence Tasks and, in our case, correspond to labeling a single Tweet)[3]. We paid workers an equivalent of USD 12 per hour and obtained a University IRB exemption before fielding the task. Three different workers labeled each Tweet[4]. We used majority votes to obtain a final set of labels for each Tweet[5].

Finally, we used OpenAI's research API to query ChatGPT4 on the four Ps of marketing in the validation sample. We instructed ChatGPT4 with the following role-task-format (RTF) prompt:

[2] Krippendorff's α is the ratio between the observed weighted percent agreement and the chance weighted percent agreement of labels. It ranges from -1 to 1, with 1 representing unanimous agreement, 0 indicating random label assignments, and negative values suggesting systematic disagreement.

[3] We had a rejection rate of 13.3% due to attention check failures.

[4] In total, 53 mTurk workers participated. Mean labeling time per Tweet was 29 seconds. Total cost for the study was $410.75.

[5] Krippendorff's α for label agreement among amateur labels is .391, indicating some agreement.

You are a renowned marketing scholar and an expert on the 4 Ps of Marketing: Product, Place, Price, and Promotion. When given a numbered list of Tweets, you examine each Tweet individually. For each Tweet, determine which of the 4 Ps it is about, if any. Output all relevant Ps for each Tweet. Use only the terms Product, Place, Price, and Promotion. Do not provide notes or an explanation.

We provided ChatGPT4 with Tweets in batches of 25 and parsed out the labels from its response. We set the temperature hyperparameter of ChatGPT4 to zero to obtain consistent responses (temperature controls response stochasticity; as it approaches zero, the model becomes deterministic). Nonetheless, we found variation in ChartGPT4's labels for the same Tweets. Hence, we followed the same approach as with crowdsourced amateurs: we obtained three sets of labels for each Tweet from ChatGPT4 and took the majority vote[6].

Table 1. Bootstrapped Label Agreement among Human Experts, Amateurs, and Generative AI

	Mean Krippendorff's α			Mean Classification Metrics		
	Expert	Amateur	ChatGPT4	Precision	Recall	F1-score
Expert	1 *(.000)*			1 *(.000)*	1 *(.000)*	1 *(.000)*
Amateur	.512 *(.000)*	1 *(.000)*		.835 *(.000)*	.546 *(.000)*	.660 *(.000)*
ChatGPT4	**.786** *(.000)*	.470 *(.000)*	1 *(.000)*	**.893** *(.000)*	**.833** *(.000)*	**.862** *(.000)*

Notes: $N = 4{,}000$ labels. Analysis based on 1,000 bootstrapped Tweets. Standard errors are in parentheses. Human expert labels are the ground truth for classification metrics.

To evaluate label agreement among experts, crowdsourced amateurs, and ChatGPT4, we bootstrap the labeled validation sample 1,000 times, calculate four agreement metrics (Krippendorff's α, precision, recall and F1-score) for each bootstrap, and report their means with standard errors in Table 1. We find strong agreement among experts and ChatGPT4 across all four metrics (R2). In contrast, crowdsourced labels from amateurs exhibit substantially less agreement with expert labels (R1). Although not a perfect substitute, we conclude that generative AI is a viable alternative to scarce and costly domain experts for labeling text.

[6] Krippendorff's α for label agreement among three sets of ChatGPT4 labels is .700, indicating good, but not perfect agreement.

3.3. Creating a Synthetic Expert

To overcome ChatGPT4's proprietary nature, high cost, limited reproducibility, and slow processing speed, we propose to approximate it for our classification task with an open-source LLM that we fine-tune on ChatGPT4's labels. LLMs are general-purpose models that provide a universal understanding of language that can be used for a wide variety of tasks (Manning 2022). An abundance of pretrained LLMs, models already trained on various datasets, are available online[7]. These pretrained LLMs can easily be fine-tuned with task-specific training data using transfer learning (Howard and Ruder 2018).

We contend that a specific classification task does not necessitate a massive generative AI model like ChatGPT4. Instead, analysts typically require an expert model that (1) understands natural language, and (2) can identify specific constructs of interest in text. Fine-tuning a pretrained LLM with examples labeled by ChatGPT4 meets both requirements and provides analysts with a synthetic expert that they have full control over. In this study, we fine-tune a pretrained LLM called RoBERTa (Liu et al. 2019) with training data that we label using ChatGPT4. Previous research found RoBERTa to perform well on various datasets when fine-tuned for classification (Hartmann et al. 2023).

We proceed as follows: First, we use ChatGPT4 to label the training sample. We follow the same procedure as for the validation sample. Next, we download a pretrained RoBERTa model[8] and fine-tune it as a multi-label classifier using our labeled training sample. Fine-tuning involves adapting the pretrained model's output layer to suit our multi-label classification task. Specifically, we replace the original output layer (also referred to as classification head) with a new one that is tailored for multi-label classification, and randomly initialize the weights and biases of this new layer (i.e., classification head). We retrain the model on our training sample, leveraging the pretrained weights and biases of the model's other layers, allowing us to build upon the language representations learned by RoBERTa without learning everything from scratch.

[7] See https://huggingface.co/models for a large selection of pretrained LLMs.
[8] See https://github.com/facebookresearch/fairseq/tree/main/examples/roberta for model details.

Because the possibility exists that a Tweet pertains to more than one marketing mix variable (i.e., P), we face a multi-label classification task (a text can pertain to multiple classes), not a multi-class classification task (a text pertains to one of the multiple classes). Consider the following text:

> *@Sony's XM3's ain't as sweet as my bro's airpod pros but got a real steal* 🤑 *the other day* #deal #headphonez.

The writer compares in-ear headphones of two brands (product) and rationalizes purchasing the inferior headphones with a big discount (price). The distinction between multi-label and multi-class classification tasks is conceptually important because it determines the loss function required for fine-tuning the LLM. While multi-class classifiers are typically trained on cross-entropy loss, we use binary-cross-entropy loss in our multi-label classification task because each label corresponds to an independent, binary decision.

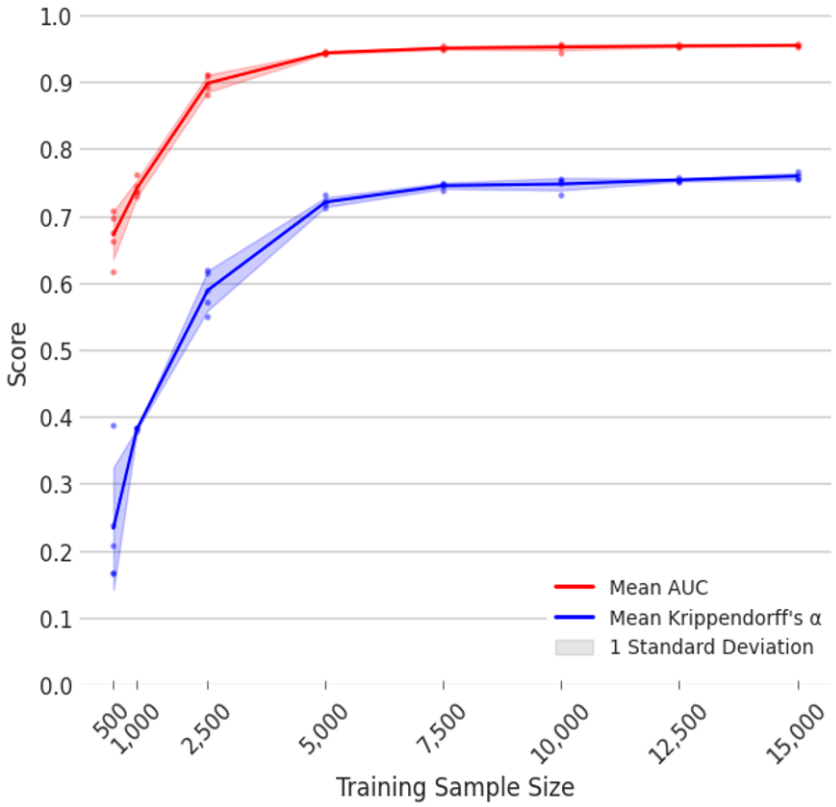
Finally, fine-tuning an LLM involves setting several hyperparameters. To improve model performance, we tune the learning rate, weight decay, number of training epochs, warm-up steps, and batch size using the hyperparameter optimization framework OPTUNA (Akiba et al. 2019). Because model performance can vary depending on the utilized training samples and how the model training is seeded (Dodge et al. 2020), we fine-tune RoBERTa five times with different seeds for different randomly drawn training samples of different sizes to ensure robust findings.

We evaluate each fine-tuned model (i.e., each synthetic expert) on our validation sample using AUC and Krippendorff's α to capture different aspects of the agreement between human experts and synthetic experts. AUC is a common performance metric in machine learning that evaluates a model's ability to distinguish between classes, regardless of where the probability threshold for assigning a class (here, marketing mix variable) to a sample (here, a Tweet) is set. Krippendorff's α controls for chance agreement but is fixed to a particular probability threshold (here, .5). We present our findings in Figure 1.

Overall, our synthetic experts perform very well in identifying which of the four marketing mix variables a Tweet pertains to. We observe a very high AUC of .898 with as few as 2,500 training samples. Our finding is in line with that of Hartmann et al. (2023), who found that relatively few training samples can already produce good results when fine-tuning LLMs. Krippendorff's α rapidly rises as the number of

training samples approaches 5,000. These rapid gains diminish as the sample size increases further. At 15,000 training samples, the mean of Krippendorff's α reaches .759, indicating very good agreement among synthetic and human experts. Notably, the performance of our synthetic expert is only 3.4% below ChatGPT4 (R3).

Figure 1. Agreement of Synthetic Expert Labels with Human Expert Labels on Validation Sample for different Training Samples and Model Initializations



Notes: N = 4,000 labels. Human expert labels are the ground truth for scores. Area under receiver operator characteristic curve (AUC) ranges from 0 to 1, with 1 indicating perfect class discrimination, and .5 indicating no discrimination between classes. Krippendorff's α ranges from -1 to 1, with 1 representing unanimous agreement, 0 indicating random label assignments, and negative values suggesting systematic disagreement. Dots indicate individual scores. The shaded area indicates one standard deviation from the mean score for each training sample size. We observe a slight variance in performance of the synthetic expert across the four marketing mix variables (F1-Score σ2 = .001; Krippendorff's alpha σ2 = .005), with the lowest relative performance for promotion, and the highest for price.

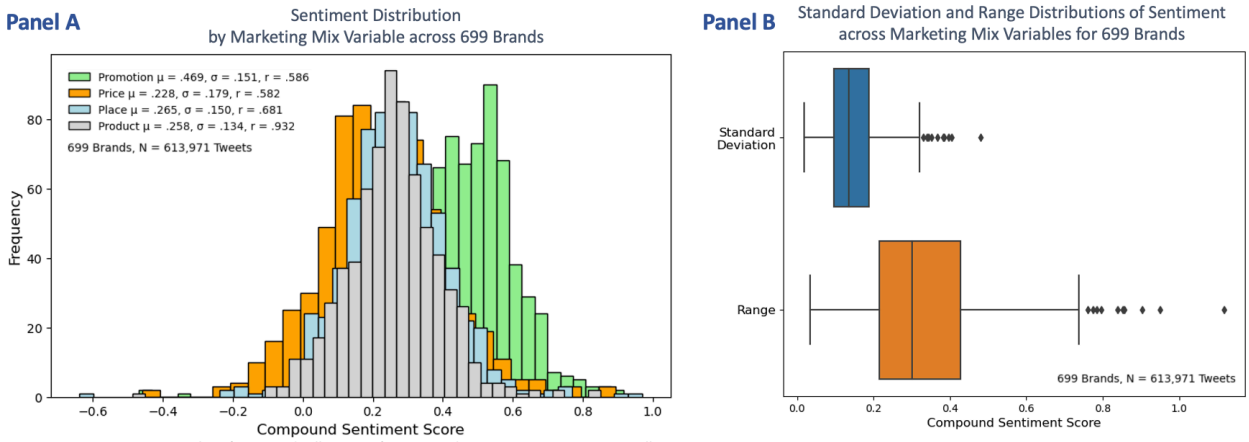
3.4. Consumer Sentiment in the Marketing Mix

Sentiment analysis is one of the most common classification tasks in marketing (Hartmann et al. 2023). It provides marketers with a lens on consumer perception, and contributes a meaningful variable to marketing

research (e.g., Tirunillai and Tellis 2012; Homburg et al. 2015; Park et al. 2021). Recent research shows that measuring sentiment for specific constructs of interest creates deeper insights than overall sentiment (Zhong and Schweidel 2020; Puranam et al. 2021; Ananthakrishnan et al. 2023). In what follows, we examine the necessity to disambiguate consumer sentiment on Twitter into brands' marketing mix (R4).

We sample over half a million Tweets of 699 brands and identify the marketing mix variables each of them pertains to using our MMX synthetic expert. We measure the sentiment of each Tweet using VADER (Hutto and Gilbert 2014), an established sentiment analysis tool for of Tweets. We find that the sentiment distributions among marketing mix variables deviate considerably (Figure 2, Panel A), and that the standard deviation and range of sentiment across the marketing mix variables of individual brands can be substantial (Figure 2, Panel B). Our finding calls for a more differentiated evaluation of consumer sentiment to discover strengths and weaknesses in brands' marketing mix.

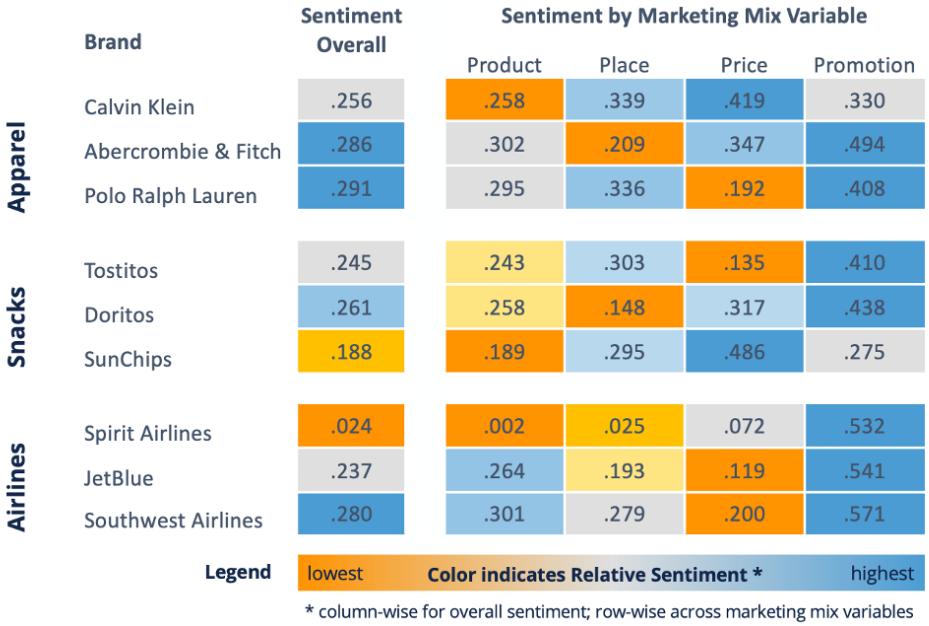
Figure 2. Differences in Sentiment Distributions by Marketing Mix Variable



***Notes**: VADER Sentiment Analysis (Hutto and Gilbert 2014), 699 Brands, N = 613,971 Tweets, Overall Sentiment: μ = 0.258, σ = 0.133, Range = $Sent_{mam}$ - $Sent_{min}$*

We zoom-in on consumer sentiment by marketing mix variable for nine major brands in three different categories in Figure 3. We find that brands have different strengths and weaknesses in terms of consumer sentiment across the marketing mix. Notably, brands within the same category such as Polo Ralph Lauran and Abercrombie & Fitch (A&F) may exhibit nearly the same overall sentiment, but very different sentiment distributions across their marketing mix: while Polo Ralph Lauren's lowest consumer sentiment is for price, A&F's lowest consumer sentiment is for place.

Figure 3. Sentiment by Marketing Mix Variable for Brands of different Categories



***Notes**: VADER Sentiment Analysis (Hutto and Gilbert 2014), 9 Brands, N = 9,000 randomly sampled Tweets from 2020; stratified by brand*

We conclude that marketers can be misled by overall sentiment (Figure 3) because it obfuscates potentially harmful weakness in their marketing mix. Our analysis further shows that substantial differences in sentiment across brands' marketing mix exists (Figure 2), which calls for its disambiguation (R4).

3.5. Synthetic Experts as Focal Lens for Richer Insights

In today's digital marketing era, understanding consumer conversations on social media is essential for effective brand management. The use of synthetic expertise to efficiently reveal specific aspects of these conversations, here by example of the marketing mix, promises a new level of strategic insight. Synthetic expertise shifts the focus from high-level analysis to a detailed exploration, revealing deeper insights that marketers require to adjust and fine-tune their strategies.

To illustrate the practical impact of synthetic expertise, we zoom-in on A&F Tweets that pertain to "place", identified as an area of low sentiment and thus of high importance. Specifically, we discover topics in consumer Tweets pertaining to "place" using unsupervised machine learning and natural language processing. The basic idea is to first cluster A&F "place" Tweets into semantically similar topics using

sentence embedding and density-based clustering. We then describe each topic through the most relevant words that occur in its Tweets and visualize our findings in bar charts. We follow a four-step process.

In step one, we use SBERT by Reimers and Gurevych (2019) to capture the semantic meaning of Tweets in high-dimensional numerical representations that are known as embeddings. SBERT is a fine-tuned version of the large language model BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2018). SBERT's fine-tuning process adjusts BERT's model parameters to reduce the distance between embeddings of semantically similar sentences and increase it for dissimilar ones. As a result, SBERT's embeddings are ideal for cluster analysis, as they group similar Tweets closer together in the embedding space based on their semantic content, facilitating the identification of distinct thematic clusters (i.e., topics).

Before we cluster the embedded Tweets to discover topics, we reduce their dimensionality using UMAP (Uniform Manifold Approximation and Projection) by McInnes and Healy (2018). UMAP is adept at preserving the relationships between data points in clusters, ensuring the data's structural integrity is maintained even in a lower-dimensional space. By reducing the dimensionality of our embeddings while preserving their topological structure, UMAP helps accentuate the natural groupings or clusters among the embedded Tweets. That is, UMAP brings semantically similar Tweets closer together while keeping dissimilar ones apart. In high-dimensional space, clusters might be obscured or less discernible due to the complexity of natural language. However, when reduced to a lower-dimensional space, the semantic similarities and differences between Tweets become more pronounced and easier to identify.

We then apply the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) clustering algorithm by McInnes et al. (2017) to the dimensionality-reduced embeddings to uncover topics among the embedded Tweets. Our choice of HDBSCAN is driven by its ability to effectively handle clusters of varying sizes, making it ideal for the diverse and uneven distribution of topics in social media content. Furthermore, as a non-exhaustive clustering method, HDBSCAN handles noise, distinguishing between significant clusters and outliers. This is especially beneficial given the often noisy and unstructured nature of social media data. Crucially, HDBSCAN eliminates the need to set the number of clusters a priori,

allowing for more organic and data-driven topic discovery, which is particularly useful given the unpredictable and dynamic nature of social media conversations.

In the final step, we identify the most relevant n-grams (i.e., contiguous sequence of n words from a given sample of text) for each topic using TF-IDF (Term Frequency - Inverse Document Frequency). TF-IDF is a numerical score that reflects how relevant an n-gram is to a topic within a collection of topics. It is calculated by multiplying the frequency of an n-gram in a specific topic by the inverse of its frequency across all topics, thereby highlighting n-grams that are most relevant to individual topics. We use BERTopic (Grootendorst 2022) to visually present the top 10 most relevant n-grams for each topic[9] in Figure 4. The topics are labeled based on their top-10 n-grams and randomly sampled Tweets of each topic.

Figure 4. "Place" Topics in Tweets that mention Abercrombie & Fitch

The insights gleaned from Figure 4 highlight the practical efficacy of synthetic expertise in marketing analysis. These findings not only corroborate with existing narratives about Abercrombie & Fitch (e.g., Klayman 2022), thereby validating their authenticity, but also showcase the depth of insight that synthetic expertise can enable. Our analysis demonstrates that synthetic expertise transcends the classification of simple constructs, enabling marketers to penetrate the layers of consumer discourse more effectively. It

[9] BERTopic also offers a convenient way to automatically daisy-chain the utilized models. Nonetheless, we recommend carrying out each step individually with the most current version of the respective models and then passing the results to BERTopic for visualization.

facilitates the swift identification of focal points in texts and conversations, areas that merit particular attention due to their relevance. In essence, synthetic expertise serves as a discerning lens, sharpening the focus of marketers on the most critical elements embedded in textual data, and thus, enhancing their decision-making.

4. Discussion and Conclusion

In an era characterized by an overwhelming abundance of textual information, understanding and classifying text is fundamental for firms and society. This research introduces synthetic experts—scalable, replicable, and fully independent classification models that approximate generative AI to identify complex constructs in vast amounts of data.

We empirically investigate the ability of a generative AI, ChatGPT4, to identify a complex construct in text that is challenging to analyze. We find that ChatGPT4's labels strongly agreed with those of human experts. In contrast, crowdsourced labels agreed substantially less. Our findings highlight the great potential that generative AI models like ChatGPT4 hold for complex classification tasks.

Despite the promising performance of ChatGPT4, it is crucial to consider its limitations. Being a proprietary model, firms and organizations relying on ChatGPT4 may become dependent on OpenAI, which controls access, pricing, and capabilities. This dependence also poses confidentiality and data privacy risks as sensitive information may be passed through a third-party system. Moreover, ChatGPT4's slow processing speed and high cost make it less feasible for large-scale applications. Finally, as this research shows, generative AI's answers are not necessarily deterministic, which represents a substantial reproducibility risk to the research community.

Synthetic experts mitigate the dependency on proprietary models, offering scalability, full control, reproducibility, and independence from third-party constraints. Notably, Krippendorff's α for our MMX synthetic expert was only 3.4% lower than that of ChatGPT4, making it viable substitute.

Nonetheless, our research is not without its own set of limitations. Our approach for creating synthetic experts was validated on Twitter posts pertaining to brands' marketing mix. Its performance in classifying

other constructs in other types of text remains to be tested. We contend, however, that our approach will also work for various other constructs—provided that they are established and broadly documented such that the utilized generative AI is "aware" of them (i.e., learned them from the data it was trained on). Finally, synthetic experts' performance may vary depending on the utilized text and construct of interest.

We believe that synthetic expertise lays the foundation for a wealth of research opportunities and practical applications. On the technical front, pairing synthetic experts with active learning strategies might further improve classification performance. Additionally, the vast array of pretrained language models, including those specialized in user-generated content, allows for fine-tuning models that most suitable for a given task. Researchers can also explore fine-tuning OpenAI's proprietary models on its platform with context-specific data and examples. Such fine-tuning may produce better labels, which in turn improves the performance of the entailing synthetic expert. The approach may also open a path to utilize generative AI for new constructs that are not yet broadly documented (i.e., general purpose models are not yet sufficiently "aware of them").

On the substantive front, synthetic experts facilitate the generation of theoretically founded variables from unstructured data—independent of the field of study. Researchers could use these variables as input to their models to answer research questions, test hypotheses, and explain mechanisms. Firms and organizations, on the other hand, find a practical tool in synthetic experts that enables them to harness their data assets. Marketing offers an abundance of complex constructs such as SWOT analysis, dimensions of service quality, customer experience, or branding (i.e., equity, identity, image). In public policy, analysts could use synthetic experts to identify agenda-setting and policy frames in documents, government communications, or news reports. And in management, firms might use them to identify leadership styles such as autocratic, democratic, or laissez-faire from corporate communications, internal memos, or employee reviews.

The field of generative AI is rapidly evolving, with new models being introduced by various organizations (e.g., Google, Meta, Anthropic). Some models may be more suitable for creating synthetic

experts than others. Multiple models could even be pooled to improve the training process. We leave this exciting frontier to further exploration by fellow scholars.

By sharing our code and model online at www.synthetic-experts.ai, we hope to catalyze research and applications in this exciting and rapidly evolving field. We also share what we call "*Synthetic Twins*" of Tweets. Synthetic twins correspond semantically in idea and meaning to original texts. However, wording, people, places, firms, brands, and products were changed by a generative AI. As such, synthetic twins mitigate, to some extent, potential privacy, confidentiality, and intellectual property concerns. Analysts can use our synthetic twins in their own models or run our code to create synthetic twins on their own.

In sum, our work bridges the gap between conventional text classification methods and generative AI. As an industry manager recently put it: *"You don't buy the whole candy store when all you need is a lollipop."* The analogy succinctly captures the essence of the proposed synthetic experts: tailored, efficient solutions that address specific needs without the overhead and constraints of larger, more general AI models. Importantly, the idea of creating synthetic experts for classification tasks using generative AI can easily be extended to other media such as image, audio, and video, paving the way for versatile and powerful applications across many domains. As the field of generative AI continues to evolve, we expect to see more innovations and applications in classification tasks and beyond.

References

Abbasi A, Li J, Adjeroh D, Abate M, Zheng W (2019) Don't mention it? Analyzing user-generated content signals for early adverse event warnings. *Information Systems Research* 30(3):1007-1028. https://doi.org/10.1287/isre.2019.0847

Akiba T, Sano S, Yanase T, Ohta T, Koyama M (2019) Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623-2631.

Ananthakrishnan U, Proserpio D, Sharma S (2023) I hear you: Does quality improve with customer voice? *Marketing Science* 0(0). https://doi.org/10.1287/mksc.2023.1437

Archak N, Ghose A, Ipeirotis PG (2011) Deriving the pricing power of product features by mining consumer reviews. *Management Science* 57(8):1485-1509. https://doi.org/10.1287/mnsc.1110.1370

Avramov D, Cheng S, Metzker L (2023) Machine learning vs. Economic restrictions: Evidence from stock return predictability. *Management Science* 69(5):2587-2619. https://doi.org/10.1287/mnsc.2022.4449

Ban G-Y, Rudin C (2019) The big data newsvendor: Practical insights from machine learning. *Operations Research* 67(1):90-108. https://doi.org/10.1287/opre.2018.1757

Berger J, Humphreys A, Ludwig S, Moe WW, Netzer O, Schweidel DA (2020) Uniting the tribes: Using text for marketing insight. *Journal of Marketing* 84(1):1-25. https://doi.org/10.1177/0022242919873106

Brand J, Israeli A, Ngwe D (2023) Using gpt for market research. *Available at SSRN 4395751*. https://dx.doi.org/10.2139/ssrn.4395751

Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, Lee P, Lee YT, Li Y, Lundberg S (2023) Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*. https://doi.org/10.48550/arXiv.2303.12712

Burnap A, Hauser JR, Timoshenko A (2023) Product aesthetic design: A machine learning augmentation. *Marketing Science* Forthcoming. https://doi.org/10.1287/mksc.2022.1429

Busch KE (2023) Generative artificial intelligence and data privacy: A primer. Report, R47569, Congressional Research Service.

Chakraborty I, Kim M, Sudhir K (2022) Attribute sentiment scoring with online text reviews: Accounting for language structure and missing attributes. *Journal of Marketing Research* 59(3):600-622. https://doi.org/10.1177/00222437211052500

Chen X, Liu Q, Wang Y (2023) Active learning for contextual search with binary feedback. *Management Science* 69(4):2165-2181. https://doi.org/10.1287/mnsc.2022.4473

Chui M, Roberts R, Yee L (2022) Generative ai is here: How tools like ChatGPT could change your business. *Quantum Black AI by McKinsey*.

Cohn D, Atlas L, Ladner R (1994) Improving generalization with active learning. *Machine Learning* 15:201-221. https://doi.org/10.1007/BF00993277

Daniels J (2023) How generative ai can affect your business' data privacy. *Forbes*, Retrieved May 15, 2023, https://www.forbes.com/sites/forbesbusinesscouncil/2023/05/01/how-generative-ai-can-affect-your-business-data-privacy/.

Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. https://doi.org/10.48550/arXiv.1810.04805

Dew R, Ansari A, Toubia O (2022) Letting logos speak: Leveraging multiview representation learning for data-driven branding and logo design. *Marketing Science* 41(2):401-425. https://doi.org/10.1287/mksc.2021.1326

Dodge J, Ilharco G, Schwartz R, Farhadi A, Hajishirzi H, Smith N (2020) Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*. https://doi.org/10.48550/arXiv.2002.06305

Frankel R, Jennings J, Lee J (2022) Disclosure sentiment: Machine learning vs. Dictionary methods. *Management Science* 68(7):5514-5532. https://doi.org/10.1287/mnsc.2021.4156

Grootendorst M (2022) Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*. https://doi.org/10.48550/arXiv.2203.05794

Guo B, Zhang X, Wang Z, Jiang M, Nie J, Ding Y, Yue J, Wu Y (2023) How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*. https://doi.org/10.48550/arXiv.2301.07597

Hartmann J, Heitmann M, Siebert C, Schamp C (2023) More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing* 40(1):75-87. https://doi.org/10.1016/j.ijresmar.2022.05.005

Hartmann J, Huppertz J, Schamp C, Heitmann M (2019) Comparing automated text classification methods. *International Journal of Research in Marketing* 36(1):20-38. https://doi.org/10.1016/j.ijresmar.2018.09.009

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

Homburg C, Ehm L, Artz M (2015) Measuring and managing consumer sentiment in an online community environment. *Journal of Marketing Research* 52(5):629-641. https://doi.org/10.1509/jmr.11.0448

Horton JJ (2023) Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv preprint arXiv:2301.07543*. https://doi.org/10.3386/w31122

Howard J, Ruder S (2018) Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*. https://doi.org/10.48550/arXiv.1801.06146

Hutto C, Gilbert E (2014) Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the international AAAI conference on web and social media*, 216-225.

Kiritchenko S, Zhu X, Mohammad SM (2014) Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research* 50:723-762. https://doi.org/10.1613/jair.4272

Klayman A (2022) White hot: The rise & fall of abercrombie & fitch. USA: Netflix.

Kotler P, Calder BJ, Malthouse EC, Korsten PJ (2012) The gap between the vision for marketing and reality. *MIT Sloan Management Review* 53(1):13-14.

Lawrence J, Reed C (2020) Argument mining: A survey. *Computational Linguistics* 45(4):765-818. https://doi.org/10.1162/coli_a_00364

Li P, Castelo N, Katona Z, Sarvary M (2022) Language models for automated market research: A new way to generate perceptual maps. *Available at SSRN 4241291*. https://dx.doi.org/10.2139/ssrn.4241291

Liaukonytė J, Tuchman A, Zhu X (2023) Frontiers: Spilling the beans on political consumerism: Do social media boycotts and buycotts translate to real sales impact? *Marketing Science* 42(1):11-25. https://doi.org/10.1287/mksc.2022.1386

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. https://doi.org/10.48550/arXiv.1907.11692

Luangrath AW, Xu Y, Wang T (2023) Paralanguage classifier (PARA): An algorithm for automatic coding of paralinguistic nonverbal parts of speech in text. *Journal of Marketing Research* 60(2):388-408. https://doi.org/10.1177/00222437221116058

Mallipeddi RR, Kumar S, Sriskandarajah C, Zhu Y (2022) A framework for analyzing influencer marketing in social networks: Selection and scheduling of influencers. *Management Science* 68(1):75-104. https://doi.org/10.1287/mnsc.2020.3899

Manning CD (2022) Human language understanding & reasoning. *Daedalus* 151(2):127-138. https://doi.org/10.1162/daed_a_01905

McInnes L, Healy J (2018) Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*. https://doi.org/10.48550/arXiv.1802.03426

McInnes L, Healy J, Astels S (2017) Hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* 2(11):205. https://doi.org/10.21105/joss.00205

Park S, Shin W, Xie J (2021) The fateful first consumer review. *Marketing Science* 40(3):481-507. https://doi.org/10.1287/mksc.2020.1264

Peres R, Schreier M, Schweidel D, Sorescu A (2023) On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice. *International Journal of Research in Marketing* 40(2):269-275. https://doi.org/10.1016/j.ijresmar.2023.03.001

Puranam D, Kadiyali V, Narayan V (2021) The impact of increase in minimum wages on consumer perceptions of service: A transformer model of online restaurant reviews. *Marketing Science* 40(5):985-1004. https://doi.org/10.1287/mksc.2021.1294

Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training. Report, OpenAI Technical Report, San Francisco, CA.

Reimers N, Gurevych I (2019) Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*. https://doi.org/10.48550/arXiv.1908.10084

Rocklage MD, He S, Rucker DD, Nordgren LF (2023) Beyond sentiment: The value and measurement of consumer certainty in language. *Journal of Marketing Research* 1:19. https://doi.org/10.1177/00222437221134802

Rust RT, Lemon KN, Zeithaml VA (2004) Return on marketing: Using customer equity to focus marketing strategy. *Journal of Marketing* 68(1):109-127. https://doi.org/10.1509/jmkg.68.1.109.24030

Rust RT, Rand W, Huang M-H, Stephen AT, Brooks G, Chabuk T (2021) Real-time brand reputation tracking using social media. *Journal of Marketing* 85(4):21-43. https://doi.org/10.1177/0022242921995173

Schoenmueller V, Netzer O, Stahl F (2023) Frontiers: Polarized america: From political polarization to preference polarization. *Marketing Science* 42(1):48-60. https://doi.org/10.1287/mksc.2022.1408

Şeref MM, Şeref O, Abrahams AS, Hill SB, Warnick Q (2023) Rhetoric mining: A new text-analytics approach for quantifying persuasion. *INFORMS Journal on Data Science,* Forthcoming. https://doi.org/10.1287/ijds.2022.0024

Shapiro BP (1985) Rejuvenating the marketing mix. *Harvard Business Review* 63(5):28-34.

Snow R, O'connor B, Jurafsky D, Ng AY (2008) Cheap and fast–but is it good? Evaluating non-expert annotations for natural language tasks. *Proceedings of the 2008 conference on empirical methods in natural language processing*, 254-263.

Suslava K (2021) "Stiff business headwinds and uncharted economic waters": The use of euphemisms in earnings conference calls. *Management Science* 67(11):7184-7213. https://doi.org/10.1287/mnsc.2020.3826

Tirunillai S, Tellis GJ (2012) Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science* 31(2):198-215. https://doi.org/10.1287/mksc.1110.0682

Van Dis EA, Bollen J, Zuidema W, van Rooij R, Bockting CL (2023) ChatGPT: Five priorities for research. *Nature* 614(7947):224-226. https://doi.org/10.1038/d41586-023-00288-7

Van Noorden R, Perkel JM (2023) Ai and science: What 1,600 researchers think. *Nature* 621(7980):672-675. https://doi.org/10.1038/d41586-023-02980-0

Van Waterschoot W, Van den Bulte C (1992) The 4p classification of the marketing mix revisited. *Journal of Marketing* 56(4):83-93. https://doi.org/10.1177/002224299205600407

Wu L, Hitt L, Lou B (2020) Data analytics, innovation, and firm productivity. *Management Science* 66(5):2017-2039. https://doi.org/10.1287/mnsc.2018.3281

Xia S, Liu C (2022) Applying machine learning to study the marketing mix's effectiveness in a social marketing context: Fashion brands' twitter activities in the pandemic. *International Journal of Business Analytics (IJBAN)* 9(6):1-17. https://doi.org/10.4018/IJBAN.313416

Yoganarasimhan H (2020) Search personalization using machine learning. *Management Science* 66(3):1045-1070. https://doi.org/10.1287/mnsc.2018.3255

Zhang Z, Yang K, Zhang JZ, Palmatier RW (2023) Uncovering synergy and dysergy in consumer reviews: A machine learning approach. *Management Science* 69(4):2339-2360. https://doi.org/10.1287/mnsc.2022.4443

Zhong N, Schweidel DA (2020) Capturing changes in social media content: A multiple latent changepoint topic model. *Marketing Science* 39(4):827-846. https://doi.org/10.1287/mksc.2019.1212