A. Marketing literature using sentiment analysis

We sorted the table by method group and then year of publication. We added a column for model comparison to denote whether the authors simply used a model (N), compared multiple models to see which was best for their task (Y/N), or whether the model comparison was the main goal of the study (Y). The table also has a final section reserved for hybrid approaches; hybrid approaches use 2 or more classifiers in tandem to improve accuracy.

Authors	#Classes	Model(s)	Model Comparison	(Sentiment Analysis related) Goal / Research Question	Results
			Dictionary Methods		
Joyce and Kraut (2006)	2	DICT (LIWC)	N	Does emotional valence of responses to newcomers on newsgroups influence their continued participation?	Neither quality nor emotional tone influenced the continued participation of newcomers.
King et al. (2006)	-	DICT (LIWC)	N	Do customers of different weight experience different levels of discrimination?	Obese customers face greater levels of discrimination than average-weighted customers. Comments and interactions are more negative.
Herhausen et al. (2019)	2	DICT (LIWC) + Customization	N	Developed a framework to detect (potential) online firestorms in brand communities and how to mitigate them.	Negative sentiment is an obvious precursor to online firestorms and must be addressed adequately. Especially high-arousel level negative emotions from frequent customers can lead to online firestorms.
Netzer et al. (2019)	-	DICT(LIWC) + Binary Logit, NB	N	How do words in loan application relate to loan repayment or defaults?	While several linguistic properties indicate future loan default, emotional states do not consistently point to default or repayment.
Zhou et al. (2021)	-1 to 1	DICT (VADER)	N	Developed a framework to extract features from videos to predict viewing behaviour, which uses (among others) text sentiment.	According to SHAP feature importances, positive sentiment has a positive impact on viewing behaviour i.e. completing a video or watching another video.

Woolley and Sharif (2021)	-	DICT (LIWC)	N	Which effect do incentives have on review sentiment?	Incentivised reviews have a higher ratio of positive to negative reviews. The positive affect of an incentive on a reviewer translates to the sentiment of the review.
Lin et al. (2021)	-	DICT (LIWC)	N	How do streamer emotions influence viewer emotions and subsequent tipping behavior?	A more happy streamer positively influences their viewers, subsequently increasing the number and quantity of monetary gifts they send.
Crolic et al. (2022)	1	DICT (LIWC)	N	How does anthropomorphism affect customer response and satisfaction during chatbot interaction.	Angry customers interacting with chatbots (or becoming angry during) show less satisfaction when experiencing anthropomorphisms.
Lacka et al. (2022)	-1 to 1	DICT (VADER)	N	How does firm-generated content impact their stock price?	Both valence and tweet subject can affect stock prices temporarily or permanently. E.g. Negative tweets about competitors lead to permanent negative stock prices developments.
Shi et al. (2022)	-1 to 1	DICT (TextBlob)	N	How do news articles, reviews and other text media amplify or spread misinformation following hype news.	News articles (as well as reviews) following hype news have significantly higher sentiment which may spread misinformation.
Guler et al. (2024)	-1 to 1	DICT (VADER)	N	How does the local market react to acquisitions?	Brand acquisitions are immediately followed by a drop in customer sentiment as seen by lower product ratings and lower sales. Lower sales and ratings locally affect the company in the long run. Large business take-overs have more negative reactions.
Mustak et al. (2024)	-1 to 1	DICT (VADER)	N	What insights can be mined from user generated content about brands?	Sentiment analysis, topic modelling and in-depth analysis allows for brand specific understanding of customer likes and dislikes from user generated content. Firms can use

					these to fine-tune their behaviour / products / services.
Liu and Li (2024)	-1 to 1	DICT (VADER)	N	Development of a method for sustainable service product design based off user generated content.	Sentiment can be used for identifying customer requirements for specific product features, determining their weights, and deriving design requirements from them.
Chen et al. (2024)	-1 to 1	DICT (VADER)	N	What effect do review visibility and diagnosticity have on review helpfulness?	Review sentiment is clearly linked to review helpfulness and improves predictive accuracy when trying to predict helpful reviews.
Hung et al. (2024)	-	DICT (LIWC)	N	How do customer emotions influence ratings?	Positive emotions have a significant positive impact on ratings, subsequently negative emotions have a negative impact. Both lead to emotional contagion effects that influence future ratings. Managers can strategically increase or dampen these emotions by replying to reviews.
			Machine Learning Methods		
Homburg et al. (2015)	2	SVM	N	How do consumers react to firm engagement in social media platforms?	Very high levels of interaction with customers by the firm can decrease customer sentiment. Overall engagement has diminishing returns.
Meire et al. (2016)	2	SVM, RF	Y/N	How does leading and lagging information affect sentiment classification accuracy?	RF generally outperform SVM regardless of supplemental information. Adding more information, however, increases performance. They briefly discuss use cases such as customer satisfaction analysis, personalized e-learning, or predict election outcomes.
Chen et al. (2017)	6	SVM, NB, DT, Ensemble	Y/N	Which factors influence emotion classification accuracy?	More advanced preprocessing of data, as well as more complex models (SVM / Ensemble) improve predictive power.

Tirunillai and Tellis (2017)	2	SVM, NB	Y/N	How does consumer sentiment change after advertising campaigns?	Negative consumer sentiment decreases after ad campaigns. This effect is delayed and decreases over time.
Kratzwald et al. (2018)	Various	RF, SVM, LSTM, Bi-LSTM	Y/N	How do pre-trained word-embeddings improve predictive accuracy?	Deep learning models such as LSTM and Bi-LSTM outperform traditional machine learning. Pre-trained embeddings further improve these models. Pre-training the whole model on a related task is also strictly beneficial. They furthermore provide an overview of selected use cases (see table 6).
Sánchez-Franco et al. (2019)	2	NB	N	How can online review data be used to identify pain points in hotel attributes and services?	NB assisted aspect-based SA helps point out specific problems in hotels with greater precision than statistical alternatives at low computational cost.
Liu et al. (2019)	2	SVM, NB, LSTM, Recursive NN, CNN	Y/N	Analysed the impact of content specific (n=6) sentiment (e.g. price, aesthetics) on conversion rates across many categories.	Content with positive content about aesthetics and price have a strong positive impact on conversion rate. Reranking the reviews accordingly can increase conversion rate.
Kazmaier and van Vuuren (2020)	2, 3	DICT (various), NB, SVM, LR, NN, CNN, LSTM	Y	Development of a comprehensive and generic sentiment analysis framework	A comprehensive framework was developed that encompasses the complete sentiment analysis pipeline. This framework allows for feature engineering, preprocessing, hyper-parameter tuning, evaluation and visualization.
Li and Xie (2020)	2	LR, NB, SVM, RF	Y/N	How do image (and text) content drive user engagement in social networks?	Positive text content increases likes and retweets, while negative text content significantly increases shares but has a strong negative impact on likes.

Schoenmueller et al. (2020)	2	LR	N	What are the drivers of online review polarity?	Reviewers that submit a lot of reviews tend to have less polar (1 or 5 star) ratings than those that submit few.
Chatterjee et al. (2021)	5	LR, XGBoost, RF, DT	Y/N	How can reviews be used to identify key drivers to customer satisfaction across various industries, especially healthcare?	Both overall sentiments and attribute specific sentiments can provide valuable insights. These are oftentimes specific to the business model and vary to others. Subsequently sentiment derived from reviews can be used in customer relationship management to improve services in the healthcare industry. Both sentiment and emotions explain customer satisfaction.
Chuah and Yu (2021)	-1 to 1	DT, RF, boosted trees, SVM	Y/N	Which facial attributes have the most impact on overall text sentiment?	A combination of low sadness, and high happiness / surprise / neutral facial features incurs the most positive sentiment in online videos. Tree based methods mostly perform on a similar level.
Chakraborty et al. (2022)	5	DICT(Custom), SVM, NB, LR, CNN, LSTM, CNN-LSTM	Y	Development of a fine-grained attribute level SA model, to provide information both how positive / negative a review is and about what.	A complex CNN-LSTM hybrid model outperforms classical approaches and helps managers rerank various attribute importances.
Kolomoyets and Dickinger (2023)	5	XGBoost	N	Which text attributes are most important for positive hotel reviews? Can machine learning reflect insights from traditional approaches?	High prevalence of topics like cleanliness lead to more negative reviews while staff topics tend to indicate more positive reviews. Machine learning leads to similar attributes in reviews as traditional methods. These can be used for service improvement and tracking business performance.

Wu and Morwitz (2024)	2	NB (Multilabel)	N	How does the tonality of a review influence the recovery process from a negative experience?	Reviewers with negative experiences can recover both affectively and cognitively if they write
			Transfer Learning Methods		
Hartmann et al. (2021)	3	BERT	N	Analyse user response to brand selfies using various visual and textual features.	SA is achieved through fine-tuning a RoBERTa language model on a subset of the data. It comes to similar conclusions as LIWC but achieves better accuracy.
Xiao et al. (2022)	3	CNN, BERT, BiLSTM, CRF	Y/N	Developed and fine-tuned a pre-trained language model to produce more informative embeddings. These embeddings can be used for various tasks, like user preference mining.	Using the more advanced embeddings and model it outperforms all others on various tasks for sentiment analysis. Subsequent user preference mining for air conditioners reveal which features are deemed positive, and which are deemed negative.
Park et al. (2023)	2	BERT	N	Developed a 4-step framework for smart speaker product improvement.	SA can be used in conjunction with network analysis, and topic modelling to identify product attribute related strengths and weaknesses tailored to individual brands.
Ananthakrishnan et al. (2023)	-1 to 1	Google API	N	How do ratings improve for hotels after addressing negative reviews?	Low rated hotels that address negative attributes improve their overall ratings more than high rated hotels.
Ma et al. (2024)	0 to 1	BERT	N	How important are individual service dimensions and what are their importances?	Sentiment can be used in conjunction with topic modelling as an alternative, or enhancement to survey-based importance-performance analysis.

			Few-Shot Prompting / Learning		
Praveen et al. (2024)	2	BERT, Falcon-7B, MPT-7B, GPT-2	Y	Method comparison of LLMs for binary sentiment analysis focusing primarily on prompt design.	Prompt-Engineering can guide LLMs to accurately predict binary sentiment. Since prompts use natural language this reduces the entrance barrier and can be used by people without extensive knowledge of deep-learning.
Krugmann and Hartmann (2024)	2, 3, 5	GPT-3.5, GPT-4, Llama 2-70B	Y	Method comparison and empirical investigation of influence of textul data characteristics for binary and 3-class sentiment using LLMs.	Increasing the number of classes reduces accuracy; few-shot prompting improves performances compared to zero-shot; LLMs can provide useful explanations out of the box without a secondary model like LIME.
Yi et al. (2025)	2	GPT-4	N	Development of a visual aid to identify strengths and weaknesses of services / What are the drivers of satisfaction in mobile trading?	Using BERTopic and subsequent sentiment analysis with ChatGPT-4, service quality methods typically reliant to TAM and similar can be automated with online reviews.
			Method Comparisons		
Hartmann et al. (2019)	2 3	Various	Y	Empirical comparison of traditional machine learning models for text classification in marketing	RF and NB always outperform SVM and dictionary-based approaches across various data ranges.
Alantari et al. (2022)	3 5 10	Various	Y	Empirical comparison of various text classification methods and their explainability	Modern BERT-based approaches outperform more traditional and dictionary-based approaches. Methods like LIME enable explainability for black-box models.
Hartmann et al. (2023)	-	Various	Y	Meta-analysis of sentiment analysis methods and their efficacy	Modern transfer-learning based approaches outperform any other in most cases and tasks.

Zaghloul et al. (2024)	2	LR, SVM, RF, GBC, NN	Y	Method comparison of few machine learning models and one “deep” learning approach	For binary classification with exhaustive sampling and feature engineering, tree based methods outperform “deep” learning approaches.
			Meta		
Humphreys and Wang (2018)	-	-	Y	Provide a comprehensive overview of automated text analysis, their steps, pitfalls, and a methodological framework to choose the type of study to perform.	Automated Text analysis has multiple application areas where it can aid researchers and consumers, but it cannot be used for all text phenomena.
			Hybrid		
Tran et al. (2021)	2-3	NB, DICT (VADER)	Y/N	How do consumers react to online chatbots?	Sentiment towards chatbots is less negative than towards real humans in fashion contexts. Sentiment towards chatbots is more negative than towards real humans in telecommunication.

In the case of Hartmann et al. (2019), Alantari et al. (2022) and Hartmann et al. (2023) we don’t report the models, as they compare more than 10 different models each (21 for Alantari).

B. Marketing literature using explainability for SA

Authors	Year	Main focus of study?	Method
Hartmann et al. (2021)	2021	N	LIME
Sánchez-Franco et al. (2021)	2021	Y	SHAP
Alantari et al. (2022)	2022	Partially	LIME

Kolomoyets and Dickinger (2023)	2023	Y	SHAP
Park et al. (2023)	2023	N	SHAP
Fong et al. (2024)	2024	Y	Intrinsic heatmaps + SHAP

Dictionary Approaches

Dictionary based approaches construct lists of words which correspond to numerical values that indicate their emotional valence (Hutto & Gilbert, 2014; Pennebaker, 2001). A sentence is then matched with the dictionary and a final score is calculated, which then conveys the overall sentiment. Dictionaries are generally considered to be cumbersome in development, since a large number of words must be hand labelled, and often can't detect negations, don't work well with misspellings, overemphasized capitalization (e.g. "AWESOME"), sarcasm or out of dictionary words (Alantari et al., 2022; Hartmann et al., 2019). Although some work has been done to address several of these points (see Hutto and Gilbert (2014)), more modern approaches often outperform dictionaries. Nonetheless, dictionaries remain a popular choice as they are easy to use, interpret and integrate into larger frameworks (Luri et al., 2023; Melumad et al., 2019; Shi et al., 2022).

Machine Learning Approaches

Naïve Bayes is based on the Bayes' theorem (Lindley, 1958) and works by calculating the conditional probability of a class given a set of attributes, then the class with the highest probability is selected as the prediction. In the case of SA the set of attributes corresponds to the words of a sentence, typically represented as a numerical Bag-of-Words (BoW). Tirunillai and Tellis (2017) monitor the change of consumer sentiment after brand advertising campaigns and find that advertising can decrease negative customer sentiment. Sánchez-Franco et al. (2019) classify customer satisfaction of hotels from yelp reviews. They identify multiple pain points expressed by customers, which can be used to improve hotel attributes and services.

Logistic Regression (LR) learns the dependence between the independent features (numerical representation of text) to the outcome variable (sentiment class). This way it learns which words are most important for the dependent variable (Alantari et al., 2022). Schoenmueller et al. (2020) investigate the drivers of online review polarity. They find that the more reviews a customer submits, the less polar their reviews are (very negative / very positive). Gelper et al. (2018) use three multinomial LR to infer the topic, tone and sentiment of WOM during spike events such as prerelease of products. They show that WOM during these spikes is usually more positive and indicate future product sales.

Support Vector Machines (SVM) transform data into a high dimensional non-linear representation, where they are then linearly separated which leads to high generalizability (Cortes & Vapnik, 1995). Homburg et al. (2015) classify the sentiment of social media texts to determine how firm interaction impacts customers. High levels of engagement are found to decrease customer sentiment, while any form of engagement has diminishing returns across subgroups and types of conversations. Li and Xie (2020) analyse how image and text content drive user engagement in social networks. They find that positive text increase both likes and retweets while negative texts only increase shares.

Decision Trees (DT) are a type of non-parametric machine learning model that repeatedly splits the data into child-nodes. Each split uses an independent feature as the decision boundary and aims at producing the most homogonous split possible (each child-node should ideally only contain instances of one target class). Splits only ever produce 2 child-nodes so the decision can be seen as an "either or question" (Breiman, 1984; Bruyn et al., 2008). In the context of sentiment analysis, "simple" DT approaches are generally less common than boosting strategies like XGBoost or Adaboost. Chen et al. (2017) test various text preprocessing approaches as well as ensemble configurations (combining different models) to predict the sentiment of online videos using the title, tags, description and comments. Their results show that DT are competitive even when compared to various ensemble approaches (at most 3 percentage points less), and outperform all alternatives in one preprocessing approach. Chuah and Yu (2021) used various tree based models to predict compound sentiment scores (initially determined by VADER) to determine which facial attributes (happiness, surprise, anger, …) are most relevant. They find that tree-based methods (DT, RF, boosted trees) have the least error and outclass SVM.

Random Forests (RF) are oftentimes superior across multiple application areas (Uddin et al., 2019), while still being inexpensive compared to more modern approaches. They work by implementing multiple decision trees that are trained on subsets of the data and pooling the predictions to a final estimate. This allows individuals trees to specialize on a subset of the data and makes them more robust to overfitting and more robust than simple DTs (Breiman, 2001). Meire et al. (2016) research the impact of leading and lagging information on sentiment classification. They compare SVM and RF and find that RF generally outperforms SVM but strictly benefits from more information.

XGBoost (Extreme Gradient Boosting) builds on the ensemble concept of RF, however, instead of using n independent trees, XGBoost builds trees sequentially (Chen et al., 2015). Each subsequent tree builds on the residuals of the previous chain of trees. XGBoost implements a gradient boosting method which performs additive optimization as well as regularization to avoid overfitting. The final output is determined by building the sum over all trees. Sánchez-Franco et al. (2021) use XGBoost to construct a model that predicts a users' acceptance and satisfaction with intelligent personal assistants based on the topics. Their model achieves high predictive accuracy and is used in a subsequent step to explain which models facilitate which sentiments. Kolomoyets and Dickinger (2023) use XGBoost with a subsequent SHAP interpretation and feature importance study to identify how topic prevalence relates to customer satisfaction / sentiment. For example, they find that topics indicating cleanliness are strongly associated with negative reviews (irrespective of the content), while topics about the staff indicate positive reviews. Chatterjee et al. (2021) explore how user texts can be used to identify key drivers of customer satisfaction across multiple industries. They find that both overall sentiment and attribute specific sentiment provides valuable insights into various business cases. However, an insight does not necessarily hold true across cases.

Artificial Neural Networks (ANN) are a popular tool for a variety of tasks and have found usage in marketing for decades (Thieme et al., 2000; Zhang et al., 1998). ANN are inspired by the structure of the human brain and are made up of nodes and layers. A node represents a computational unit and consists of connections, learnable weights, and an activation function. Layers consist of multiple nodes through which the information must propagate. ANN usually consist of at least 3 layers: an input layer which represents the dimensionality of the variables (e.g. one input node per variable), at least one hidden layer (containing multiple fully connected nodes), and an output layer that maps the information to the decision space (binary, multi-class, regression, …) (Thieme et al., 2000). This can then be used to adjust the weights of each connection by back-propagating the loss (Rumelhart et al., 1986; Werbos, 1974). Due to their non-linear nature ANN are capable of modelling much more complex relationships (Hartmann et al., 2019). Hartmann et al. (2019) chose ANN as part of their method comparison because of their versatility and good performance for a variety of text classification tasks. In their comparison they find that ANN is often among the top performing models and have low variance.

Convolutional Neural Networks (CNN) are a type of deep neural network and fall under the deep learning category and have predominantly been used for image tasks (Chellapilla et al., 2006). They typically consist of convolutional layers and pooling layers.

Convolutional layers capture local (e.g. image) features whereas pooling layers merge similar features (LeCun et al., 2015). Deep CNNs with multiple convolutions, pooling-, and skip-layers have excelled at various image recognition tasks (such as faces, handwritten writing, and animals). Liu et al. (2019) measure the impact of content specific sentiment on conversion rate across categories and compared several models. Their results show that CNN generally outclass other models (Naïve Bayes, SVM, LSTM) by a large margin and can help identify review content that has a strong positive impact on conversion rate. For example, they find that positive texts about aesthetics and price positively improve conversion rate.

The models described so far can identify complex relationships between features (most often even non-linear ones). However, except for CNNs they cannot consider structural dependencies. Long Short-Term Memory (LSTM) account for the sequential dependency inherently present in text. While approaches like DT require bigrams to account for negations (e.g. not-good), LSTM inherently model this relationship (Hochreiter & Schmidhuber, 1997). These networks often rely on so-called embedding layers, that automatically learn a vector representation for the words. Since the weights of the embedding layers are initially random, they are trained alongside the rest of the model. Another option is to use pre-trained embeddings such as GloVe (Global Vectors for Word Representation) (Pennington et al., 2014). Later on, these have been mostly substituted by transformer based embedding strategies, which we discuss under the transfer learning section. Kratzwald et al. (2018) evaluated the impact of pre-trained embeddings on the performance of LSTM (and Bi-LSTM) for affective classification (Joy, Anger, …). They find that deep learning models like LSTM strictly outperform traditional models, additionally using pre-trained embeddings (such as GloVe) improves performance. In a second step they first pre-trained the network on a typical sentiment task and then fine-tuned on the affective task. This pre-training greatly improves performance.

CNN-LSTM leverages strengths from both individual models to achieve better performance. Usually after an embedding layer, CNN layers extract important text features at different levels which can are then fed into an LSTM which models the sequential nature of text (Chakraborty et al., 2022). Chakraborty et al. (2022) compare several classical models (SVM, Naïve Bayes, LR) with more complex deep learning models (CNN, LSTM, CNN-LSTM) to predict both product attributes and their sentiment from reviews to identify which specific issues restaurants should address. They find that CNN-LSTM outperform all other models and show that corrections of attribute mentions can lead to an increase in the overall attribute ratings.

Transfer Learning Approaches

With more recent developments in ML researchers have unlocked pre-training and transfer learning capabilities for NLP thanks to a mechanism called attention (Bahdanau et al., 2015; Luong et al., 2015). The attention mechanism allows models to focus on different parts of an input sequence, when determining the output. This concept has improved tasks like machine translation, and finally led to the transformer architecture which relies solely on attention mechanisms (Vaswani et al., 2017). For a model overview see Figure 1. Models based on the transformer architecture can be pre-trained on vast amounts of text data with the task of masked language modelling and next sentence prediction, which enables abstract language understanding (Devlin et al., 2019). These models are then used in downstream tasks and fine-tuned to fit specific needs, such as SA. Again, vast datasets are commonly used to achieve both high accuracy and generalizability for the new context. After this stage researchers and practitioners can use the pre-trained and fine-tuned model for their own problem, without having to manually label any data, or train a model from scratch (see e.g. Hartmann et al. (2023) for a fine-tuned SA model)[1].

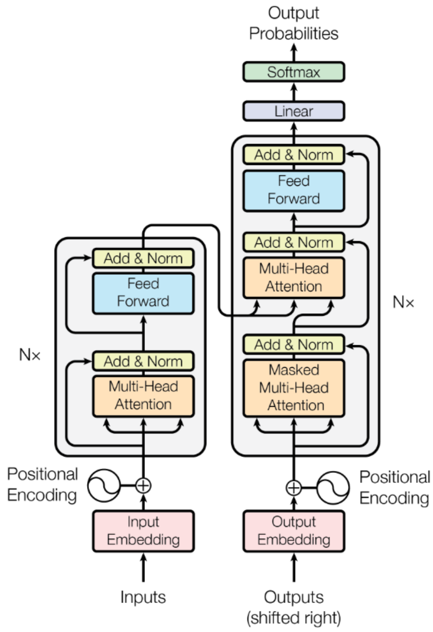

Fig. 1: Transformer model – Source: Vaswani et al. (2017)

BERT consists of multiple stacked encoders (left block of Figure 1) that produces rich text embeddings where each word is enriched with attention from both preceding and succeeding words, which increases the contextual information. BERT and subsequent improvements were trained on large text corpora such as BooksCorpus (a collection of 11,038 free books with 800m words) and English Wikipedia entries (2,500M words) (Devlin et al., 2019). BERT was then pre-trained using the aforementioned masked language modelling and next sentence prediction tasks. Ma et al. (2024) use BERT to substitute survey based importance-performance analysis by mixing topic modelling and SA. In the context of online food delivery services they are able to identify important topics which were overlooked by conventional methods. Park et al. (2023) use BERT as part of a 4-step framework combining SA, topic modelling, and network analysis to identify brand specific product strengths and weaknesses. Both Alantari et al. (2022) and Hartmann et al. (2023) concluded in their method comparison, that BERT (and later improvements) outperform all other models. This is also heavily reflected in the computer science literature.

At this point we note that there is also the possibility to use SA as a service (i.e. using an API). Various cloud providers offer specific tools for e.g. SA that have similarly been trained on large corpora. Ananthakrishnan et al. (2023) use the Google Cloud SA API to conduct entity level SA to extract which aspects hotel chains should improve. They find that lower rated hotels that listen to reviews have a larger potential to improve than higher rated ones. While APIs offer a cheap and uncomplicated option to SA, because of a pay as you go system and no hardware requirements, we will not provide an overview of API usage. Instead, as part of the few-shot learning approaches, we show ways to use ChatGPT in a similar way.

Few-Shot Learning Approaches

Few-Shot Learning (FSL) is a concept where few examples are given to a model to learn from, and usually excels in scenarios where only little data is available, labelling is

[1] Models are often shared free of charge and under permissive (i.e. commercially usable) licenses on online platforms such as HuggingFace: https://huggingface.co/models

expensive or more data is hard to come by (Brown, T. et al., 2020; Tunstall et al., 2022). Few-Shot Prompting (FSP), as a subset of FSL, uses few examples inside of prompts to LLMs such as GPT3.5 to infer the class of an unknown instance without having to train, or fine-tune the model further. This is usually referred to as in-context-learning (Wies et al., 2024). By sampling N known instances and crafting a Q&A prompt from it with the last question having no answer, the LLM is then tasked to complete the prompt (text completion task). See Figure 2 for a simple zero-shot example; the expected output in this case is {“label”: “neutral”}. Usually, the prompt is designed to mimic the interaction that is widely known from Web-Interfaces such as OpenAI’s ChatGPT and should use a JSON format (see e.g. Kondrashchenko and Kostromin (2023) for a framework). Brown, T. B. et al. (2020) have demonstrated that GPT3 performs strongly across a variety of tasks and datasets, which heavily reduces the amount of data needed to infer new and unknown data. Despite its novelty, FSL and FSP has seen some use in marketing literature. Reisenbichler et al. (2022) use it to draft content to support search engine optimization. Krugmann and Hartmann (2024) use FSP on a small dataset for binary and 3-class sentiment, and find that FSP with ChatGPT-4 is on par, or even exceeds the performance of fine-tuned transfer learning models. Praveen et al. (2024) show similar results for a binary sentiment dataset where they fine-tuned several LLMs. However, as shown by Brown, T. et al. (2020) and Tunstall et al. (2022), these results generally do not hold for more diverse datasets, and more classes.

````
BASIC_SENTIMENT = """
What is the most likely sentiment (using these labels: {labels}) for the following review.
(Provide your response in a JSON format containing a single key `label`):

Review: ```{x}```
"""
FEW_SHOT_SENTIMENT = """
What is the most likely sentiment (using these labels: {labels}) for the following review.
You will be given training data with their assigned sentiment, use this to classify the review.
(Provide your response in a JSON format containing a single key `label`):

Training data:
{training_data}

Review: ```{x}```

Your JSON response:
"""
````

Fig. 2: Examples of Zero-Shot and Few-Shot prompts for a sentiment task

One major downside of FSP is, that researchers have to oftentimes craft prompts manually, and that the output quality heavily depends on the quality of the supplied prompt (Tunstall et al., 2022; Zhou et al., 2023). Similarly, the output is dependent on the autoregressive nature of the LLM and may diverge from the expected output format. This can even happen with sophisticated and high parameter LLMs such as Llama2 (Krugmann & Hartmann, 2024). As an alternative to FSP Tunstall et al. (2022) propose a framework (SetFit) that leverages sentence transformers (ST), which are often used for sentence similarity tasks. In a first step SetFit trains a Siamese ST model by learning the text similarity of positive and negative examples. Afterwards a classifier is trained using the embeddings learned from the first step. Tunstall et al. (2022) report impressive accuracies with as little as 8 examples, when compared to models like GPT-3 or a fine-tuned RoBERTA (n=3000). The substantially smaller model allows for inference on consumer grade GPUs and offers the same strengths of FSP without having to craft prompts. SetFit has not found any application in relevant marketing literature yet. Similarly to FSP, SetFit’s performance deteriorates with an increasing number of classes and is generally outclassed by a fine-tuned model trained on the full training dataset.

Tab. 1 Strengths and weaknesses of evaluated models

<table>
<tr><th>Model</th><th>Strengths</th><th colspan="2">Weaknesses</th></tr>
<tr><td>Logistic regression</td><td rowspan="3">• Easy to train / tune / optimize
• Fast inference
• Intrinsically interpretable
• Good when interpretability is main focus, and some labelled data is available</td><td rowspan="3">• Can only model limited complexity
• Performance heavily relies on feature representation
• Typically struggles with high dimensional data
• Sensitive to noise
• Does not account for sentence structure
• Usually, no pre-trained models</td><td>• Linear assumption</td></tr>
<tr><td>Complement / Multinomial Naïve Bayes</td><td>• Independence assumption</td></tr>
<tr><td>Decision tree</td><td>• Prone to overfitting</td></tr>
<tr><td>Random forest</td><td>• Easy to train / tune / optimize
• Less prone to overfitting
• Accounts for non-linear relationships
• Fast inference</td><td colspan="2">• Prone to overfitting with little data
• Loses interpretability
• Robust against noise
• Does not account for sentence structure
• Usually, no pre-trained models</td></tr>
<tr><td>Convolutional neural networks</td><td>• Can account for sentence structure locally
• Efficient at extracting features
• Robust to noise</td><td colspan="2">• Requires specialized hardware
• Requires careful fine-tuning
• Less effective at long contexts
• Little to no pre-trained models</td></tr>
<tr><td>DistilBERT</td><td rowspan="2">• Highly versatile and adaptable, making it suitable for diverse tasks and contexts.
• Excels at handling complex and long data, including tasks with long sentences or extensive context.
• Performs best with ample labeled data, leveraging its ability to capture nuanced relationships.</td><td rowspan="2">• Requires specialized hardware
• Requires further fine-tuning to specific task
• Reduced inference speed
• Data requirements
• Not interpretable by design</td><td>• Less accurate than BERT
• Less robust than BERT</td></tr>
<tr><td>BERT</td><td>• Requires more training
• Higher latency than distilled versions</td></tr>
<tr><td>NLPTown</td><td rowspan="2">• Usable without further training
• Extremely high performance
• Further training usually improves performance</td><td rowspan="2">• Requires specialized hardware
• Slow training and inference speed
• Requires tasks to match
• Not interpretable by design
• May require adjustment</td><td>• Primarily trained for 5-class sentiment</td></tr>
<tr><td>SieBERT</td><td>• Primarily trained for binary sentiment</td></tr>
<tr><td>SetFit</td><td>• Requires little labelled data
• Works well with few classes
• Reasonably fast inference</td><td colspan="2">• Requires specialized hardware
• Requires tuning to specific tasks
• Usually outperformed by fully trained models</td></tr>
<tr><td>ChatGPT-3.5</td><td>• Requires little labelled data
• Works well with few classes
• Can explain and interpret classifications</td><td colspan="2">• Requires expensive hardware and takes very long for training and inference
• Predictions are not deterministic
• Interpretability hinges on non-deterministic interactions
• Usually outperformed by fully trained models</td></tr>
</table>

References

Alantari, H. J., Currim, I. S., Deng, Y., & Singh, S. (2022). An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. *International Journal of Research in Marketing*, *39*(1), 1–19. https://doi.org/10.1016/j.ijresmar.2021.10.011

Ananthakrishnan, U., Proserpio, D., & Sharma, S. (2023). I Hear You: Does Quality Improve with Customer Voice? *Marketing Science*, *42*(6), 1143–1161. https://doi.org/10.1287/mksc.2023.1437

Bahdanau, D., Cho, K. H., & Bengio, Y. (Eds.) (2015). *Neural machine translation by jointly learning to align and translate*.

Breiman, L. (1984). *Classification and regression trees*. Routledge.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & others (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). *Language Models are Few-Shot Learners.* https://doi.org/10.48550/arXiv.2005.14165

Bruyn, A. de, Liechty, J. C., Huizingh, E. K. R. E., & Lilien, G. L. (2008). Offering Online Recommendations with Minimum Customer Input Through Conjoint-Based Decision Aids. *Marketing Science*, *27*(3), 443–460. https://doi.org/10.1287/mksc.1070.0306

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

Chatterjee, S., Goyal, D., Prakash, A., & Sharma, J. (2021). Exploring healthcare/health-product ecommerce satisfaction: A text mining and machine learning application. *Journal of Business Research*, *131*, 815–825. https://doi.org/10.1016/j.jbusres.2020.10.043

Chellapilla, K., Puri, S., & Simard, P. (2006). High performance convolutional neural networks for document processing. In *Tenth international workshop on frontiers in handwriting recognition.* Symposium conducted at the meeting of Suvisoft.

Chen, K., Tsai, C.-F., Hu, Y.-H., & Hu, C.-W. (2024). The effect of review visibility and diagnosticity on review helpfulness – An accessibility-diagnosticity theory perspective. *Decision Support Systems*, *178*, 114145. https://doi.org/10.1016/j.dss.2023.114145

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., & others (2015). Xgboost: extreme gradient boosting. *R Package Version 0.4-2*, *1*(4), 1–4.

Chen, Y.-L., Chang, C.-L., & Yeh, C.-S. (2017). Emotion classification of YouTube videos. *Decision Support Systems*, *101*, 40–50. https://doi.org/10.1016/j.dss.2017.05.014

Chuah, S. H.-W., & Yu, J. (2021). The future of service: The power of emotion in human-robot interaction. *Journal of Retailing and Consumer Services*, *61*, 102551. https://doi.org/10.1016/j.jretconser.2021.102551

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018
Crolic, C., Thomaz, F., Hadi, R., & Stephen, A. T. (2022). Blame the Bot: Anthropomorphism and Anger in Customer–Chatbot Interactions. *Journal of Marketing*, *86*(1), 132–148. https://doi.org/10.1177/00222429211045687
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423
Fong, H., Kumar, V., & Sudhir, K. (2024). A Theory-Based Explainable Deep Learning Architecture for Music Emotion. *Marketing Science*, Article mksc.2022.0323. Advance online publication. https://doi.org/10.1287/mksc.2022.0323
Gelper, S., Peres, R., & Eliashberg, J. (2018). Talk Bursts: The Role of Spikes in Prerelease Word-of-Mouth Dynamics. *Journal of Marketing Research*, *55*(6), 801–817. https://doi.org/10.1177/0022243718817007
Guler, A. U., Misra, K., & Singh, V. (2024). Local Market Reaction to Brand Acquisitions: Evidence from the Craft Beer Industry. *Marketing Science*, *43*(5), 1102–1131. https://doi.org/10.1287/mksc.2022.0383
Hartmann, J., Heitmann, M., Schamp, C., & Netzer, O. (2021). The Power of Brand Selfies. *Journal of Marketing Research*, *58*(6), 1159–1177. https://doi.org/10.1177/00222437211037258
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
Herhausen, D., Ludwig, S., Grewal, D., Wulf, J., & Schoegel, M. (2019). Detecting, Preventing, and Mitigating Online Firestorms in Brand Communities. *Journal of Marketing*, *83*(3), 1–21. https://doi.org/10.1177/0022242918822300
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
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
Humphreys, A., & Wang, R. J.-H. (2018). Automated Text Analysis for Consumer Research. *Journal of Consumer Research*, *44*(6), 1274–1306. https://doi.org/10.1093/jcr/ucx104
Hung, H.-Y., Hu, Y., Lee, N., & Tsai, H.-T. (2024). Exploring online consumer review-management response dynamics: A heuristic-systematic perspective. *Decision Support Systems*, *177*, 114087. https://doi.org/10.1016/j.dss.2023.114087
Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*.
Joyce, E., & Kraut, R. E. (2006). Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, *11*(3), 723–747.
Kazmaier, J., & van Vuuren, J. H. (2020). A generic framework for sentiment analysis: Leveraging opinion-bearing data to inform decision making. *Decision Support Systems*, *135*, 113304. https://doi.org/10.1016/j.dss.2020.113304

King, E. B., Shapiro, J. R., Hebl, M. R., Singletary, S. L., & Turner, S. (2006). The stigma of obesity in customer service: a mechanism for remediation and bottom-line consequences of interpersonal discrimination. *Journal of Applied Psychology*, *91*(3), 579.

Kolomoyets, Y., & Dickinger, A. (2023). Understanding value perceptions and propositions: A machine learning approach. *Journal of Business Research*, *154*, 113355. https://doi.org/10.1016/j.jbusres.2022.113355

Kondrashchenko, I., & Kostromin, O. (2023). *Scikit-LLM: Scikit-Learn Meets Large Language Models*. beastbyte.ai. https://github.com/iryna-kondr/scikit-llm

Kratzwald, B., Ilić, S., Kraus, M., Feuerriegel, S., & Prendinger, H. (2018). Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision Support Systems*, *115*, 24–35. https://doi.org/10.1016/j.dss.2018.09.002

Krugmann, J. O., & Hartmann, J. (2024). Sentiment Analysis in the Age of Generative AI. *Customer Needs and Solutions*, *11*(1). https://doi.org/10.1007/s40547-024-00143-4

Lacka, E., Boyd, D. E., Ibikunle, G., & Kannan, P. K. (2022). Measuring the Real-Time Stock Market Impact of Firm-Generated Content. *Journal of Marketing*, *86*(5), 58–78. https://doi.org/10.1177/00222429211042848

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

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. https://doi.org/10.1177/0022243719881113

Lin, Y., Yao, D., & Chen, X. (2021). Happiness Begets Money: Emotion and Engagement in Live Streaming. *Journal of Marketing Research*, *58*(3), 417–438. https://doi.org/10.1177/00222437211002477

Lindley, D. V. (1958). Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society. Series B (Methodological)*, 102–107.

Liu, X., Lee, D., & Srinivasan, K. (2019). Large-Scale Cross-Category Analysis of Consumer Review Content on Sales Conversion Leveraging Deep Learning. *Journal of Marketing Research*, *56*(6), 918–943. https://doi.org/10.1177/0022243719866690

Liu, X., & Li, M.-Y. (2024). Sustainable service product design method: Focus on customer demands and triple bottom line. *Journal of Retailing and Consumer Services*, *80*, 103935. https://doi.org/10.1016/j.jretconser.2024.103935

Luong, T., Pham, H., & Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics. https://doi.org/10.18653/v1/d15-1166

Luri, I., Schau, H. J., & Ghosh, B. (2023). Metaphor-Enabled Marketplace Sentiment Analysis. *Journal of Marketing Research*, 00222437231191526.

Ma, B., Wong, Y. D., Teo, C.-C., & Wang, Z. (2024). Enhance understandings of Online Food Delivery's service quality with online reviews. *Journal of Retailing and Consumer Services*, *76*, 103588. https://doi.org/10.1016/j.jretconser.2023.103588

Meire, M., Ballings, M., & van den Poel, D. (2016). The added value of auxiliary data in sentiment analysis of Facebook posts. *Decision Support Systems*, *89*, 98–112. https://doi.org/10.1016/j.dss.2016.06.013

Melumad, S., Inman, J. J., & Pham, M. T. (2019). Selectively Emotional: How Smartphone Use Changes User-Generated Content. *Journal of Marketing Research*, *56*(2), 259–275. https://doi.org/10.1177/0022243718815429

Mustak, M., Hallikainen, H., Laukkanen, T., Plé, L., Hollebeek, L. D., & Aleem, M. (2024). Using machine learning to develop customer insights from user-generated content.

Journal of Retailing and Consumer Services, *81*, 104034. https://doi.org/10.1016/j.jretconser.2024.104034
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. https://doi.org/10.1177/0022243719852959
Park, J., Yang, D., & Kim, H. Y. (2023). Text mining-based four-step framework for smart speaker product improvement and sales planning. *Journal of Retailing and Consumer Services*, *71*, 103186. https://doi.org/10.1016/j.jretconser.2022.103186
Pennebaker, J. W. (2001). Linguistic inquiry and word count: LIWC 2001.
Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.
Praveen, S. V., Gajjar, P., Ray, R. K., & Dutt, A. (2024). Crafting clarity: Leveraging large language models to decode consumer reviews. *Journal of Retailing and Consumer Services*, *81*, 103975. https://doi.org/10.1016/j.jretconser.2024.103975
Reisenbichler, M., Reutterer, T., Schweidel, D. A., & Dan, D. (2022). Frontiers: Supporting Content Marketing with Natural Language Generation. *Marketing Science*, *41*(3), 441–452. https://doi.org/10.1287/mksc.2022.1354
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0
Sánchez-Franco, M. J., Arenas-Márquez, F. J., & Alonso-Dos-Santos, M. (2021). Using structural topic modelling to predict users' sentiment towards intelligent personal agents. An application for Amazon's echo and Google Home. *Journal of Retailing and Consumer Services*, *63*, 102658. https://doi.org/10.1016/j.jretconser.2021.102658
Sánchez-Franco, M. J., Navarro-García, A., & Rondán-Cataluña, F. J. (2019). A naive Bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services. *Journal of Business Research*, *101*, 499–506. https://doi.org/10.1016/j.jbusres.2018.12.051
Schoenmueller, V., Netzer, O., & Stahl, F. (2020). The Polarity of Online Reviews: Prevalence, Drivers and Implications. *Journal of Marketing Research*, *57*(5), 853–877. https://doi.org/10.1177/0022243720941832
Shi, Z., Liu, X., & Srinivasan, K. (2022). Hype News Diffusion and Risk of Misinformation: The Oz Effect in Health Care. *Journal of Marketing Research*, *59*(2), 327–352. https://doi.org/10.1177/00222437211044472
Thieme, R. J., Song, M., & Calantone, R. J. (2000). Artificial Neural Network Decision Support Systems for New Product Development Project Selection. *Journal of Marketing Research*, *37*(4), 499–507. https://doi.org/10.1509/jmkr.37.4.499.18790
Tirunillai, S., & Tellis, G. J. (2017). Does Offline TV Advertising Affect Online Chatter? Quasi-Experimental Analysis Using Synthetic Control. *Marketing Science*, *36*(6), 862–878. https://doi.org/10.1287/mksc.2017.1040
Tran, A. D., Pallant, J. I., & Johnson, L. W. (2021). Exploring the impact of chatbots on consumer sentiment and expectations in retail. *Journal of Retailing and Consumer Services*, *63*, 102718. https://doi.org/10.1016/j.jretconser.2021.102718
Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). *Efficient Few-Shot Learning Without Prompts*. arXiv.
Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1), 281. https://doi.org/10.1186/s12911-019-1004-8

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Werbos, P. (1974). Beyond regression: New tools for prediction and analysis in the behavioral sciences. *PhD Thesis, Committee on Applied Mathematics, Harvard University, Cambridge, MA*.

Wies, N., Levine, Y., & Shashua, A. (2024). The learnability of in-context learning. *Advances in Neural Information Processing Systems*, *36*.

Woolley, K., & Sharif, M. A. (2021). Incentives Increase Relative Positivity of Review Content and Enjoyment of Review Writing. *Journal of Marketing Research*, *58*(3), 539–558. https://doi.org/10.1177/00222437211010439

Wu, A. Y., & Morwitz, V. G. (2024). Digital Therapy for Negative Consumption Experiences: The Impact of Emotional and Rational Reviews on Review Writers. *Journal of Consumer Research*, Article ucae037. Advance online publication. https://doi.org/10.1093/jcr/ucae037

Xiao, Y., Li, C., Thürer, M., Liu, Y., & Qu, T. (2022). User preference mining based on fine-grained sentiment analysis. *Journal of Retailing and Consumer Services*, *68*, 103013. https://doi.org/10.1016/j.jretconser.2022.103013

Yi, J., Oh, Y. K., & Kim, J.-M. (2025). Unveiling the drivers of satisfaction in mobile trading: Contextual mining of retail investor experience through BERTopic and generative AI. *Journal of Retailing and Consumer Services*, *82*, 104066. https://doi.org/10.1016/j.jretconser.2024.104066

Zaghloul, M., Barakat, S., & Rezk, A. (2024). Predicting E-commerce customer satisfaction: Traditional machine learning vs. deep learning approaches. *Journal of Retailing and Consumer Services*, *79*, 103865. https://doi.org/10.1016/j.jretconser.2024.103865

Zhang, G., Eddy Patuwo, B., & Y. Hu, M. (1998). Forecasting with artificial neural networks. *International Journal of Forecasting*, *14*(1), 35–62. https://doi.org/10.1016/S0169-2070(97)00044-7

Zhou, M., Chen, G. H., Ferreira, P., & Smith, M. D. (2021). Consumer behavior in the online classroom: Using video analytics and machine learning to understand the consumption of video courseware. *Journal of Marketing Research*, *58*(6), 1079–1100.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2023). Large Language Models are Human-Level Prompt Engineers. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=92gvk82DE-