Fuzzy Systems and Data Mining IX
A.J. Tallón-Ballesteros and R. Beltrán-Barba (Eds.)
© 2023 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA231006

An End-to-End Solution for Net Promoter Score Estimation and Explanation from Social Media Using Natural Language Processing

Xinghua LIANG [a,1], Yingwei YU [a], Diego SOCOLINSKY [a], Tianyi MAO [a], Carlos ROJAS [a], Alex GORYAINOV [a], Fang WANG [a] and Xin CHEN [a]
[a] *Machine Learning Solutions Lab, Amazon Web Services, WA, USA*

Abstract. The Net Promoter Score (NPS) is often used in customer experience programs for measuring customer loyalty. Increasingly more companies seek to automatically process millions of pieces of customer feedback from social media per month in order to estimate their NPS, leveraging advanced analytics like machine learning (ML) and natural language processing (NLP). Discovering trends and themes in customer interactions helps explain the NPS, empowering companies to improve products and customer experience. In this paper, we describe an end-to-end solution for NPS estimation and explanation from social media. The process includes sentiment analysis on user comments, estimating product information based on text semantics, grouping and tagging user comments for text discovery, and NPS explanation. The solution gives companies the capability to identify overall customer sentiment and common topics in a unified platform, allowing faster analysis and insights on NPS based on customer feedback.

Keywords. Business insights, net promoter score, social media, sentiment analysis, product prediction

1. Introduction

1.1. Motivation

The Net Promoter Score (NPS) [1, 2, 3, 4] is widely used as a measure of customer loyalty, customer satisfaction and potential for growth, and is obtained by aggregating the single-digit answers of a closed question survey. However, for rapidly growing businesses, the number of people answering surveys generally represents a small portion of the overall customer base. Additionally, the NPS provides only a coarse measure and does not capture the reasons behind customer opinions. On the other hand, people often spontaneously voice their opinions via social networks and customer-centric forums, providing insight into how they feel about a product and why. These large and heterogeneous text datasets may include the answers to open questions that are part of the survey used to calculate the NPS. Systematically analyzing customer comments at

[1] Corresponding Author, Xinghua Liang, 227 W Monroe St #1400, Chicago, IL 60606, United States; E-mail: lxinghua@amazon.com.

scale, we can enable businesses to listen to more customer voices and understand the root causes behind trends in NPS.

In this paper we outline a solution realizing that vision, along with a case study using a subset of the Amazon Customer Reviews Dataset (ACR dataset) [5]. We show how we can apply natural language processing and machine learning techniques in a unified platform that enables fast analysis and insights into changes in NPS over time, based on customer feedback.

1.2. Net Promoter Score Computation

The NPS is usually computed from answers to a single survey question with a 0-10 scale: "How likely is it that you would recommend [company X] to a friend or colleague?". The answers are mapped into three classes: promoters (score of 9-10), the passively satisfied (score of 7-8), and detractors (score of 0-6). The NPS is defined as the percentage of promoters minus the percentage of detractors. An additional open question asking to explain the rated answer is also included as part of the survey [1].

1.3. NPS Explanation

Despite being widely adopted, the NPS does not provide any insight into the contributing factors leading to the score. Analysis of the replies to the associated open-ended question may provide such insights. In particular, this may be achieved by identifying specific products people discuss, how they feel about them and any recurring themes connected with them. These goals may be mapped directly into standard NLP and ML problems, namely text categorization, sentiment analysis and topic modeling, respectively. This high-level breakdown facilitates reuse of a solution for similar datasets, such as tweets or customer comments, providing the means to aggregate and handle disparate sources in a unified manner.

1.4. Related Work

The NPS is discussed in [1, 2, 3, 4]. A critical view which addresses overreliance on the score, is presented in [6]. A use case of the application of NLP to extract insights from comments from NPS surveys is presented in [7].

Details of NLP and ML problems and techniques are outside the scope of this work. Sentiment analysis has been a topic of research for over two decades – [8] presents a survey of approaches based on deep learning, and [9] covers earlier techniques in detail. Text classification is also a well-studied problem; in particular, the advantages of the approach we use, fine tuning from embeddings, is discussed in [10]. The survey in [11] covers topic modeling, including approaches that leverage embeddings, as ours does.

Extracting business insights from customers opinions has been studied over the years. It was advised in [12], that the original NPS survey include at least one open question: "What is the primary reason for your score?". Including the correct open question is critical to understanding customer motivations, as studied in [13]. Since many customers are reluctant to complete surveys, one possible mechanism to augment survey data is to include posts and comments spontaneously voiced by the customers. Because the manual analysis of large numbers of comments is infeasible, it has been common to apply machine learning and natural language processing techniques. For instance, the reviews on [14] and [15] focus on healthcare and food delivery services scenarios. Our

work extends this line of work by proposing a mechanism to estimate an analog of the NPS using sentiment analysis. For simplicity, we refer to the original NPS and our proposed method by the same name.

2. Methodology

2.1. Solution Overview

We aim to accomplish the following goals:

- Estimate the NPS for each relevant product via sentiment analysis of customer comments from multiple sources
- Group user comments into clusters to aid in root cause analysis
- Create a unified dashboard to facilitate data exploration

The proposed workflow is shown in Figure 1 and includes four separate modules.

- **Module 1 – Sentiment Analysis:** We use Amazon Comprehend to estimate customer sentiment for all comments.
- **Module 2 – Product Alignment:** We train a supervised deep learning model to predict the most likely relevant product given the comment text. We use a pre-trained multilingual Sentence-BERT model [16] for comment encoding.
- **Module 3 – Comment Clustering:** We develop an unsupervised topic clustering model to group and rank comments for each dataset.
- **Module 4 – Comment Tagging:** We apply part-of-speech tagging to extract keywords from comments. The tags can be used as a filter for comment clusters.

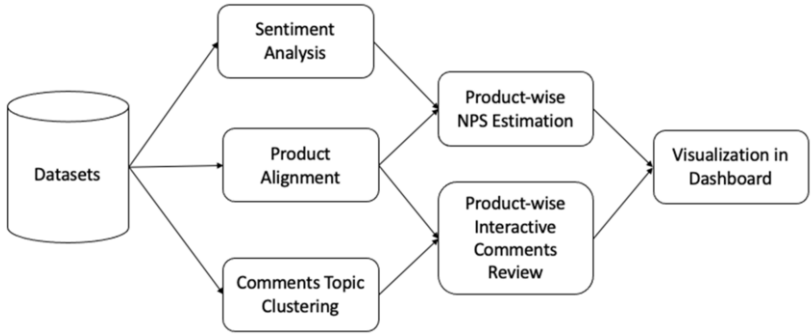


Figure 1. Solution Overview.

2.2. Exploratory Data Analysis

We use online customer feedback datasets to build the solution. Data from social media, review sites and customer service interactions can also be used. Some of the datasets may only contain parts of the relevant information needed for NPS estimation and explanation. Two primary components are needed to build the desired solution. First, we need a tool to estimate the dominant sentiment for millions of customer comments across multiple datasets. Second, in order to identify common trends and customer sentiment about specific products, we need to know which product is being discussed in each

comment. Consequently, at least one dataset must have ground truth product names associated with customer comments. This dataset is used for training a supervised learning model (see Section 2.3 Product Alignment). Other datasets may have no explicit product labels at all. The trained model is used to predict products for any comments missing product information.

2.3. Product-wise NPS Estimation with Sentiment Analysis

2.3.1. Module 1: Sentiment Analysis of User Comments

We estimate comment sentiment using Amazon Comprehend, which is a machine learning service with many NLP features, including sentiment analysis. It has been pre-trained and supports multiple languages including English, making it a good fit for all relevant datasets in our case study.

The sentiment analysis is treated as a classification task. For each comment, Amazon Comprehend will classify the input into one of the following four sentiment classes with confidence scores [17]:

- Positive: The text expresses an overall positive sentiment
- Negative: The text expresses an overall negative sentiment
- Mixed: The text expresses both positive and negative sentiments
- Neutral: The text does not express either positive or negative sentiments

2.3.2. ***Module 2: Product Alignment***

We are interested in estimating NPS for a specific product or product category. Since some relevant datasets may not include such product information, we developed a machine learning model to predict the product based on the text in the comments.

We use a dataset with both user comments and product labels for training a supervised learning model, where the input is the comment text and the output is one out of several product labels. The workflow for training the product alignment model includes comment cleaning, comment embedding, and model training. We start by extracting input comments from the training dataset and cleaning them by removing special characters such as emojis. We then embed comments into a high-dimensional vector space using Sentence-BERT. Finally, we use AutoGluon Tabular [18] to train a classifier that predicts product labels from the embedding vector corresponding to each user comment.

2.3.3. Net Promoter Score Estimation

We estimate the net promoter score using the predicted comments' sentiment for each product in a rolling time window via the Eq. (1) as defined below:

$$NPS_p = \frac{C_p^+ - C_p^-}{T_p} \times 100 \quad (1)$$

where p is the product, C_p^+ is the number of positive comments for product p and C_p^- is the number of negative comments of the same product. The normalizing factor T_p is the total number of comments for product p in the given time window, which can be refreshed daily, weekly, or monthly as a user input. The resulting statistic ranges between

-100 and 100. This is an analog of the original NPS definition, where comments with positive sentiment are taken as promoters and those with negative sentiment as detractors.

2.4. Explaining NPS Changes

*2.4.1. **Module 3: Unsupervised Topic Clustering***

We group comments into topics for ease of browsing. The workflow for topic clustering is illustrated in Figure 2. We encode comments in each dataset using Sentence-BERT and then apply the UMAP algorithm [19] to project the resulting embedding vectors onto a 2-dimensional feature space. Finally, we use the HDBScan algorithm [20], a density-based hierarchical clustering method, to cluster vectors on the feature space.

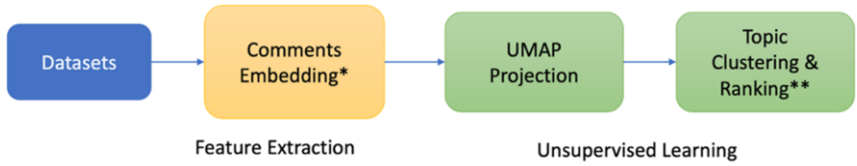


Figure 2. Workflow for Topic Clustering. (*) Comments are embedded with Sentence-BERT; (**) HDBScan is the clustering algorithm.

For example, Figure 3 shows the UMAP projection of comment embeddings for the Electronics product in the ACR dataset described below. Each point in the figure represents the UMAP projection of a single comment. Points are subsequently aggregated into clusters, ranked by size and color-coded. We extract the top five most frequent tags from comments in each cluster and use them to represent the cluster, as discussed in the following section.

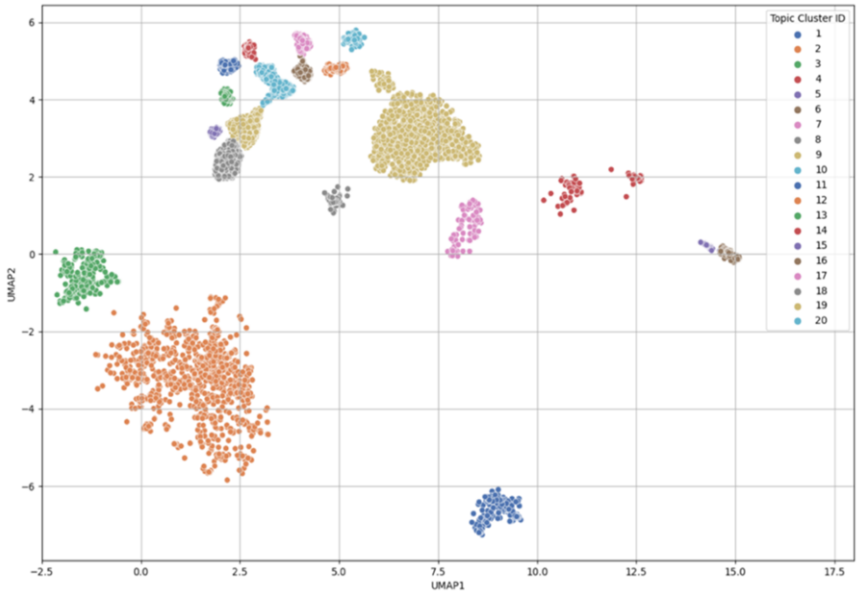


Figure 3. Example comment UMAP projections and clustering results with HDBScan for a single product category. Clusters are color-coded by ID.

2.4.2. Module 4: Comment Tagging

We assign tags to each comment to allow search and facilitate business insights. Tags are words extracted from comments through a three-step process:

- **Lemmatization:** for each word in the comment, we extract the root of the word to minimize variation due to word forms. For example, the word "shirts" is lemmatized to its base form "shirt".
- **Part of Speech (POS) analysis:** we keep only nouns, verbs and adjectives, and discard all other words.
- **Tagging deduplication:** we remove duplicate words and combine the remaining ones into a tag list.

Table 1 below shows the process of extracting tags for the comment "Shirts are too long, with heavy hem".

Table 1. Example of lemmatization and POS analysis. The input is "Shirts are too long, with heavy hem.". The final tag set for this input comment is [shirt, long, heavy, hem].

Input:	Shirts are too long, with heavy hem.	
Word	**Lemma**	**POS**
Shirts	shirt	NOUN
are	be	AUX
too	too	ADV
long	long	ADJ
,	,	PUNCT
with	with	ADP
heavy	heavy	ADJ
hem	hem	NOUN
.	.	PUNCT
Final tags list:	[shirt, long, heavy, hem]	

The survey paper "The Evolution of Topic Modeling" [11] extensively discusses the evolution and challenges of topic modeling methods. It highlights that many newer topic models still suffer from noise pollution in topics and are not scalable to the size of modern data sets. There are very few topic models designed to model on a temporal aspect, which is a significant issue given the speed with which topics on a platform can change. In addition, unsupervised topic models have limitations and that semi-supervised models may be an important future direction that balances the computational cost and the cost of labeling training data.

In contrast, our approach to topic modeling, uses an end-to-end solution for Net Promoter Score (NPS) estimation and explanation from social media. Our solution includes sentiment analysis on user comments, estimating product information based on text semantics, grouping and tagging user comments for text discovery, and NPS explanation. We use Sentence-BERT for comment encoding and apply the UMAP algorithm to project the resulting embedding vectors onto a 2-dimensional feature space. We then use the HDBScan algorithm, a density-based hierarchical clustering method, to cluster vectors on the feature space. The advantages of our approach are as follows:

- **Handling Noise**: our approach uses Sentence-BERT for encoding and HDBScan for clustering, which can handle noise in the data more effectively.
- **Scalability**: our approach is designed to handle large-scale social media data, making it more scalable for modern data sets.
- **Temporal Aspect**: our approach is designed to work with real-time social media data, implicitly taking into account the temporal nature of the data.
- **Supervised Learning**: our approach uses supervised deep learning for product alignment, which can potentially provide more accurate results than unsupervised models.
- **Practical Application**: our approach is designed with a specific application in mind (NPS estimation and explanation), making it more practical and directly useful for businesses.

3. Case Study

We used the ACR dataset to demonstrate the proposed solution for NPS estimation and explanation. This dataset contains over 130+ million customer reviews and 40+ products from 1995 through 2015. For simplicity, we restrict our case study to the ten most popular product categories: automotive, beauty, books, digital software, electronics, gift card, grocery, jewelry, kitchen and music. We sample 1,000 data points per category for each month from January to December, 2014.

3.1. Sentiment Analysis

Figure 4 shows the estimated sentiment for the top ten product categories in the ACR dataset. Positive sentiment comprises the largest percentage. Gift cards have the least negative sentiment, which is understandable since buying and redeeming a gift card is simple and straightforward. In contrast, products in the digital software category have the most negative sentiment. We will show below how the proposed solution can identify the source of this negative sentiment and how it can be addressed.

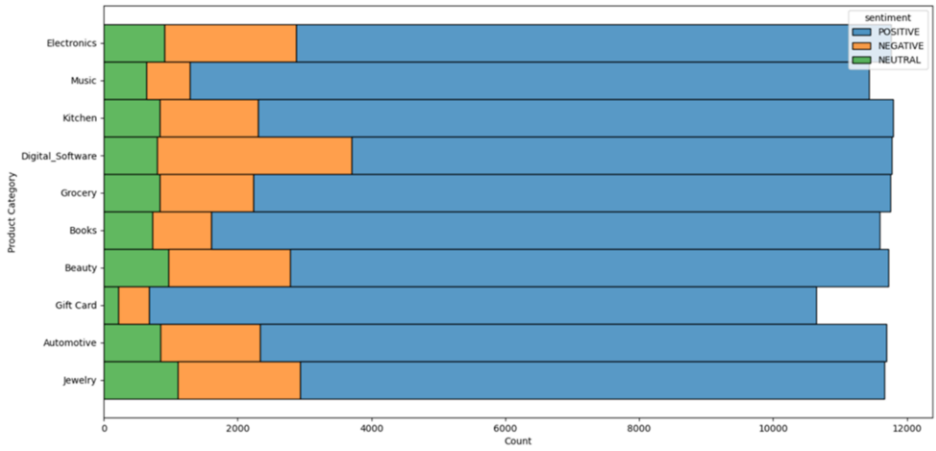


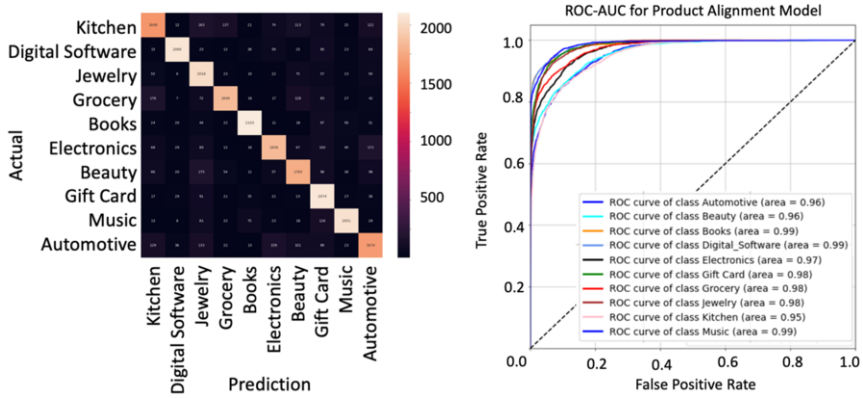
Figure 4. Sentiment result for the Amazon dataset.

3.2. Product Alignment

As described above, in many relevant cases such as social media posts there is no explicit product information. Since the ACR dataset has product information for all user comments, we simulated this situation by using separate sets for product prediction training and inference. We split the ACR dataset into training, validation and testing sets with a ratio of 70:10:20 across all product categories, encoded comments in the ACR dataset into vectors using Sentence-BERT, and trained a multi-class classifier acting on those vectors to predict product categories, achieving 79.5% accuracy for product estimation from the comment text. The confusion matrix is shown in Figure 5(a). Among the ten product categories, the "books" has the highest prediction accuracy, while kitchen is the most difficult one to predict. The model also achieved high ROC-AUC scores, ranging from 95% to 99% for each of the ten product categories. The ROC-AUC scores are shown in Figure 5(b).

3.3. Interactive Dashboard

Amazon QuickSight is a business analytics platform that powers data-driven organizations with unified business intelligence. We created a QuickSight Dashboard to visualize all insights extracted from the ACR dataset (see Figure 6). This dashboard includes 120,000 comments from January through December of 2014. In the Overview panel, besides showing sentiment and product distribution, the bottom left chart presents the NPS trend for 2014, with the score remaining around 70 throughout the year. The bottom right chart shows product-level NPS trends. We see that gift card has the highest NPS and digital software has the lowest NPS, with significant temporal fluctuation.



(a) (b)

Figure 5. (a) Confusion matrix for product category prediction. (b) ROC-AUC curve for product category prediction.

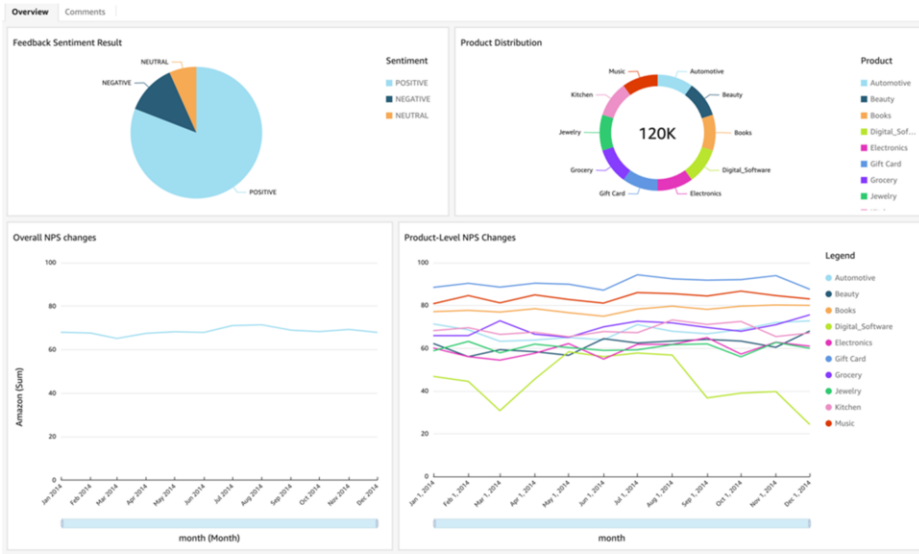


Figure 6. QuickSight dashboard overview.

The dashboard also includes a comments panel which allows business analysts to review individual user comments. By clicking on the interactive sentiment chart and product chart, users can access a filtered view showing only comments about a particular product with a particular sentiment (see Figure 7).

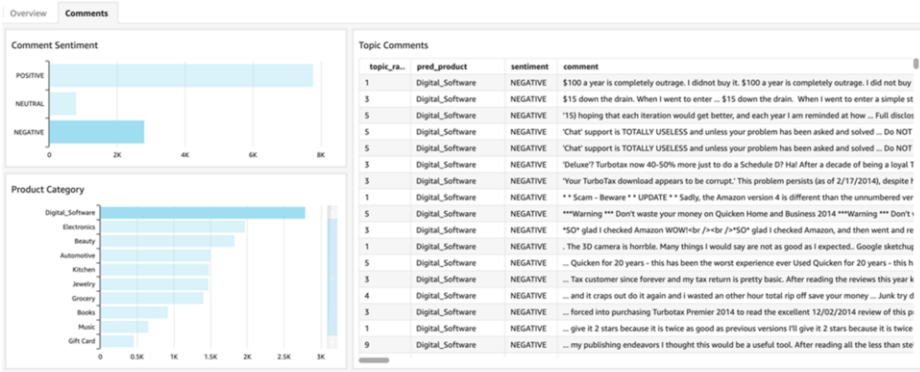


Figure 7. Topic ranking panel in the QuickSight dashboard.

Other visualizations provide insight on the primary themes for each comment cluster. An interactive chart shows topic clusters ordered by number of comments and tags extracted as described above (see Figure 8). By clicking on a particular topic cluster and a particular keyword, users can access a filtered view showing only comments of that cluster which contain the selected keyword. This helps uncover reasons underlying customer sentiment for a particular product and topic, enabling business intelligence and improving customer satisfaction. For example, looking at Figure 8, we see that the main complain about digital software is that it is difficult to download and/or install the software, which makes it unusable for many consumers.

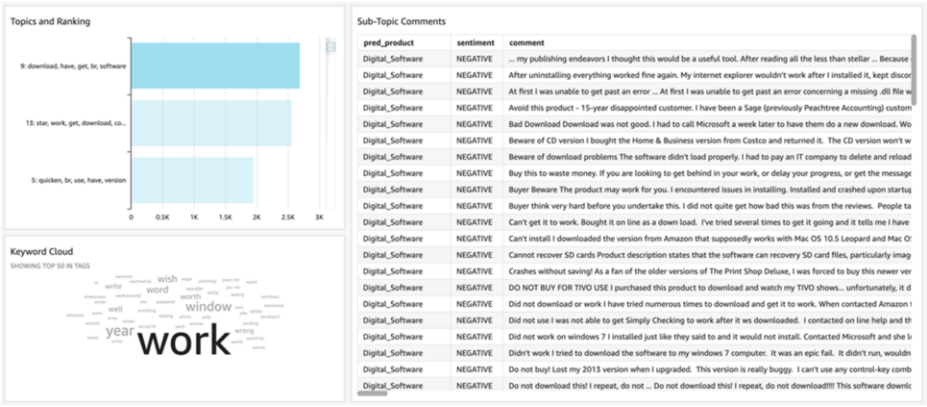


Figure 8. Topic ranking and keyword cloud panel in the dashboard.

4. Deployment

We developed an automated end-to-end pipeline implementing the solution on Amazon Web Services (AWS). We use multiple services including Amazon SageMaker (a machine learning platform), Amazon Simple Storage Service (a cloud storage service), AWS Lambda (an event-driven compute service), Amazon Simple Queue Service (a message queuing service), Amazon Elastic Container Registry (a container storage service) and Amazon QuickSight. The solution architecture is illustrated in Figure 9.

The architecture is designed to have low-overhead, be cost-effective, user-friendly and robust. It can initialize models with historical data, train and perform inference on new incremental data and update insights on a daily basis. It can also allocate computing resources dynamically, either using idle time on existing compute instances or creating dedicated instances and releasing them once processing is complete.

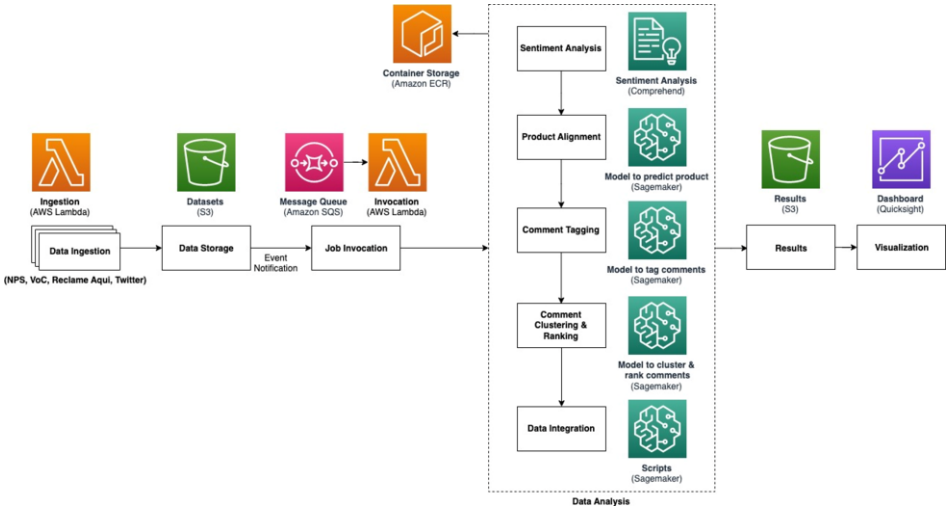


Figure 9. Solution Architecture.

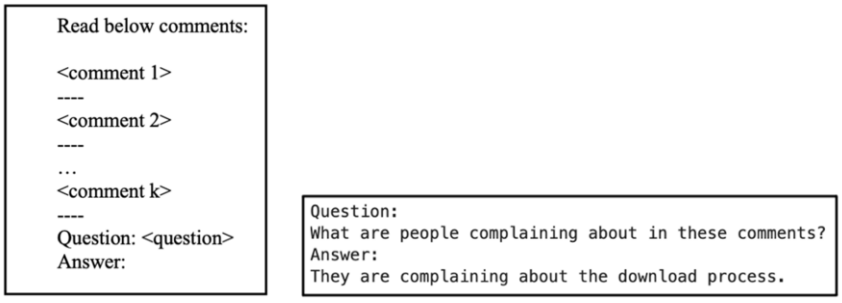
Our pipeline sequentially runs through the following modules:

- A data ingestion module that fetches new daily data and saves it to a designated Amazon S3 bucket.
- An invoking module that receives data update events from a message queue and kickstarts the inference.
- A processing module that retrieves a prebuilt Docker image from Amazon ECR containing all custom models and processes new incremental data, saving the result to another designated Amazon S3 bucket.
- A trigger to refresh the QuickSight dashboard and display the updated results.

Our architecture can be extended to incorporate new models in the processing module, supporting parallelized inference based on different types of data feeds and more frequent update of the insights at additional computation cost.

5. Opportunities for Improvement

One strength of the proposed solution is the ability to seamlessly incorporate additional modules. Emerging large language models (LLMs) present an opportunity to improve this solution, given their wide-ranging capabilities in natural language processing and generation. We performed a simple experiment incorporating the Flan-T5 [21] (flan-t5-xl) language model to offer insights into customers' comments. Since the model supports a limited number of tokens as input, it is not possible to use it for summarization of all comments in a cluster. Only some of the most relevant comments can be included into the prompt template shown in Figure 10a.



(a) Prompt Template (b) Using LLM to gain Insights

Figure 10. Improving Customer Experience with LLM

Fortunately, our solution already provides topic clustering and ranking. We can include the top-k comments for a particular topic cluster into the prompt template. For example, Figure 8 shows comments with negative sentiment for the 9th topic cluster for digital software. As a test, we included the top 27 comments in the prompt template, with the question “What are people complaining about in these comments?”. Using this prompt, Flan-T5 generates the answer “They are complaining about the download process.”, as shown in Figure 10b. The model’s answer matches our own conclusion based on manual review. This experiment suggests that our solution can easily incorporate a LLM to automatically identify valuable insights from customers' feedback

and better understand their concerns. Our future work includes studying mechanisms to use LLMs in other modules and to identify the best way to integrate them with the solution.

6. Conclusion

In this study, we developed a comprehensive customer comment analytics platform that integrates sentiment analysis, product category prediction, common topic clustering, and keyword extraction, along with an interactive dashboard. Our solution enables business analysts to identify overall customer sentiment, feedback trends, and themes in a unified environment, thereby facilitating faster insights into why net promoter scores (NPS) rise and fall. This, in turn, empowers companies to adjust their strategies for specific product categories in a timely manner.

Our approach is unique in its application of natural language processing and machine learning techniques in a unified platform to analyze changes in NPS over time, based on customer feedback. We validated our methodology using the Amazon Customer Reviews Dataset, which provided us with a rich source of customer feedback across a wide range of product categories.

One of the key findings of our study is the potential of our solution to incorporate large language models (LLMs) to automatically identify valuable insights from customers' feedback and better understand their concerns. This was demonstrated through a test case where the LLM's answer matched our own conclusion based on manual review. Looking ahead, our future work includes studying mechanisms to use LLMs in other modules and identifying the best way to integrate them with our solution. We believe that by scaling our solution, we can enable businesses to listen to more customer voices and understand the root causes behind trends in NPS.

In conclusion, our study contributes to the field by providing a robust and scalable solution for NPS estimation and explanation from social media. We believe that our work will be valuable for businesses seeking to leverage advanced analytics to improve their products and customer experience.

Acknowledgement

The authors would like to thanks to Aldo Arizmendi and Hellen Rosa for many helpful discussions.

References

[1] Reichheld F. The one number you need to grow. Harvard Business Review. 2003; 81(12): 46–124
[2] Reichheld F. The Ultimate Question: Driving Good Profits and True Growth. Harvard Business School Press. 2006
[3] Reichheld F. Darnell D. and B. Maureen, Net Promoter 3.0. Harvard Business Review. Nov – Dec 2021.
[4] Reichheld F. and B. Markey, The Ultimate Question 2.0: How Net Promoter Companies Thrive in a Customer-Driven World. Harvard Business School Press. 2011

[5] Amazon Customer Reviews Dataset. https://s3.amazonaws.com/amazon-reviews-pds/readme.html
[6] Safdar K, Pacheco I: CEOs Embrace a Dubious Metric. The Wall Street Journal (May 16, 2019). Print edition. Online as "The Dubious Management Fad Sweeping Corporate America" at https://www.wsj.com/articles/the-dubious-management-fad-sweeping-corporate-america-11557932084
[7] Zaki M, McColl-Kennedy JR, Neely A. Using AI to Track How Customers Feel — In Real Time. Harvard Business Review. 2021. Available online at https://hbr.org/2021/05/using-ai-to-track-how-customers-feel-in-real-time
[8] Zhang L, Wang S, Liu B. Deep Learning for Sentiment Analysis: A Survey. Preprint arXiv:1801.07883. Wiley Interdisciplinary Reviews - Data Mining and Knowledge Discovery. 2018; 8(4). DOI: 10.1002/widm.1253, (invited paper).
[9] Liu B. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press. June 2015 June.
[10] Howard J, Ruder S. Universal language model fine-tuning for text classification. Association for Computational Linguistics. 2018.
[11] Churchill R, Singh L. The Evolution of Topic Modeling. ACM Comput. Surv. 2022 Jan; 54, 10s, Article 215. DOI: https://doi.org/10.1145/3507900
[12] Reichheld F, Markey R. The Ultimate Question 2.0: How Net Promoter Companies Thrive in a Customer-Driven. Harvard Business Review Press. 2020.
[13] Bone SA, Lemon KN, Voorhees CM, Liljenquist KA, Fombelle PW, Detienne KB, Money RB. "Mere Measurement Plus": How Solicitation of Open-Ended Positive Feedback Influences Customer Purchase Behavior. Journal of Marketing Research. 2017; 54(1):156–170. https://doi.org/10.1509/jmr.14.0232
[14] Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. BMJ health & care informatics. 2021; 28(1), e100262. https://doi.org/10.1136/bmjhci-2020-100262
[15] Adak A, Pradhan B, Shukla N. Sentiment Analysis of Customer Reviews of Food Delivery Services Using Deep Learning and Explainable Artificial Intelligence: Systematic Review. Foods. 2022 Nov; 1500. https://doi.org/10.3390/foods11101500
[16] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BET-Networks. In 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Hong Kong, China. 2019. DOI: https://doi.org/10.48550/arXiv.1908.10084
[17] Amazon Comprehend Documentation, URL: https://docs.aws.amazon.com/comprehend/latest/dg/how-sentiment.html
[18] Erickson N, Mueller J, Shirkov A, Zhang H, Larroy , Li M, Smola A. Autogluon-tabular: Robust and accurate automl for structured data. 2020. https://doi.org/10.48550/arXiv.2003.06505.
[19] McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018. DOI: https://doi.org/10.48550/arXiv.1802.03426
[20] Campello RJGB, Moulavi D, Sander J. Density-Based Clustering Based on Hierarchical Density Estimates. In: Pei J, Tseng VS, Cao L, Motoda H, Xu G. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science. 2013; vol 7819. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14
[21] Chung HW, Hou L, Longpre S, et al. Scaling Instruction-Finetuned Language Models. DOI: https://doi.org/10.48550/arXiv.2210.11416