NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Enhancing Customer Feedback Analysis in the Insurance Sector through Generative AI and NLP

Fatma Zahra Boumelala

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Enhancing Customer Feedback Analysis in the Insurance Sector through Generative AI and NLP

by

Fatma Zahra Boumelala

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics

Supervised by

Augusto Santos

Ph.D., NOVA Information Management School

July, 2025

Acknowledgments

I would like to express my sincere gratitude to Generali Tranquilidade for the opportunity to carry out this master's internship and for providing access to the dataset and resources that made this project possible. I am thankful for the chance to collaborate with the team and contribute to a real-world application in the insurance domain.

I would also like to thank my academic supervisor, Professor Augusto Santos, for his continuous guidance, valuable feedback, and support throughout the development of this thesis.

A special thanks to Francisco Nogueira, my supervisor at Generali Tranquilidade, whose mentorship, insights, and encouragement were fundamental to the progress of this work.

Lastly, I am deeply grateful to my parents, family, and friends for their constant support, patience, and encouragement throughout this journey. Their belief in me has been a constant source of strength and motivation.

Abstract

This study presents an automated text classification system for categorizing customer reviews from Generali Tranquilidade, an insurance company, into 11 predefined topics, such as Atendimento, Preço, Documentação, Burocracia e Processo and Ferramentas Digitais. The primary objective is to enhance the efficiency and reliability of review analysis to support data-driven decision-making. The methodology integrates Large Language Models (LLMs) for synthetic data generation to address class imbalance, Extreme Gradient Boosting (XGBoost) for classification, Conformal Prediction for uncertainty quantification, and SHAP values to interpret the model by identifying the words that most influence each classification. The dataset comprises 1,943 Portuguese customer reviews, augmented with 1,727 synthetic reviews generated by Claude 3.5. Key preprocessing steps include error correction, text cleaning, stop word removal, lemmatization, and Term Frequency-Inverse Document Frequency (TF-IDF) vectorization with custom term weighting. The XGBoost model, optimized via Bayesian optimization, achieved a 59% accuracy on real data, with improved performance on minority classes when trained with synthetic data. Conformal Prediction ensured 95.6% coverage for ambiguous reviews, while SHAP analysis identified key terms driving classifications. Despite challenges like multi-topic reviews and class imbalance, the system offers a scalable, interpretable solution for insurance feedback analysis, with implications for improving customer satisfaction and operational efficiency. Future work should explore multi-label classification and enhanced annotation strategies to better handle real-world review complexity.

Keywords

Natural Language Processing (NLP), Large Language Models (LLMs), conformal prediction, SHAP (Model Interpretability)

Sustainable Development Goals (SDG)

Decent work and economic growth

TABLE OF CONTENTS

1. Introduction

1.1 Context and Motivation

Automated text classification, a key technique in Natural Language Processing (NLP), facilitates the efficient categorization and analysis of large-scale textual data, enabling businesses to derive technical insights from unstructured feedback. In the context of customer feedback, accurately categorizing reviews by their subject matter is critical for businesses to understand customer needs and enhance service quality. For insurance companies like Generali Tranquilidade, customer reviews provide a lot of information about customer experiences across various touchpoints, such as Atendimento, Produto, Preço, Sinistros, Pagamentos, Comunicação central and Ferramentas Digitais. However, manually categorizing these reviews is labor-intensive, time-consuming, and prone to human bias, necessitating automated solutions that are both efficient and reliable.

Recent advancements in machine learning, particularly in supervised learning models like Extreme Gradient Boosting (XGBoost), have shown remarkable success in text classification tasks. These models excel at handling the complexities of natural language, such as nuanced expressions, and multi-topic content. In this study, Large Language Models (LLMs) were utilized to generate synthetic reviews to address dataset imbalances, enhancing the training data's diversity and robustness, while XGBoost served as the primary classifier.

However, challenges such as class imbalance, ambiguous language, and the need for interpretable predictions require innovative approaches, such as uncertainty quantification and explainability techniques. This project is driven by the goal of developing an efficient, reliable, and interpretable classification system tailored to the insurance domain, leveraging advanced NLP techniques to support data-driven decision-making.

1.2 Objectives of the Project

The primary objective of this project is to develop an automated text classification system for categorizing customer reviews from Generali Tranquilidade into 11 predefined topics: *Atendimento, Não Especificado - Neutro, Produto, Preço, Sinistros, Pagamentos, Não Especificado - Reclamações, Documentação, burocracia e processo, Tempo de resposta e resolução, Comunicação central, Ferramentas Digitais.*

Specific objectives include:

1. **Accurate Classification**: Implementation of a machine learning pipeline using Extreme Gradient Boosting (XGBoost) as the primary classifier to accurately categorize reviews.

2. **Data Augmentation with LLMs**: Utilize Large Language Models (LLMs) to generate synthetic reviews, addressing class imbalances in the dataset to improve model performance across minority categories.
3. **Handling Uncertainty**: Apply Conformal Prediction to manage ambiguous or multi-topic reviews.
4. **Ensuring Interpretability**: Apply SHAP (SHapley Additive exPlanations) values to interpret model predictions, identifying the most influential features in the model's output.

These objectives aim to enhance the efficiency and consistency of review categorization, enabling Generali Tranquilidade to derive targeted insights for improving customer satisfaction and operational processes.

1.3 Overview of the Approach

The approach combines advanced machine learning and NLP techniques to address the challenges of classifying customer reviews. The methodology includes:

- **Synthetic Data Generation**: Generating synthetic reviews to balance minority categories.
- **Text Preprocessing**: Cleaning and transforming raw customer reviews through tokenization, stop word removal, lemmatization and TF-IDF vectorization to prepare data for modeling.
- **Model Selection and Training**: Using XGBoost as the primary classifier due to its effectiveness with sparse text data.
- **Hyperparameter Tuning**: Using Bayesian optimization to fine-tune model parameters, optimizing performance across the 11 topic categories.
- **Uncertainty Quantification**: Apply Conformal Prediction to handle ambiguous or overlapping reviews.
- **Explainability**: Applying SHAP values to interpret model predictions, identifying key words or phrases driving classifications.
- **Evaluation**: Using k-fold cross-validation and metrics like accuracy, precision, recall, and F1-score to assess model performance, ensuring generalizability and reliability.

This comprehensive approach ensures a robust, interpretable, and scalable classification system tailored to the nuances of customer feedback in the insurance domain.

2

Figure1: Overview of the classification pipeline for customer reviews

2. Related Work and Literature Review

Automated text classification has long relied on traditional models like Naive Bayes and Support Vector Machines (SVMs), which perform well on smaller datasets but tend to underperform with sparse, high-dimensional inputs such as Term Frequency–Inverse Document Frequency (TF-IDF) representations (Fotache, 2023).To address these limitations, ensemble methods such as Random Forests and gradient-boosted models like eXtreme Gradient Boosting (XGBoost) have become widely used due to their ability to manage noisy features and improve performance through boosted decision trees (Collado-Montañez et al., 2024). XGBoost, in particular, has shown strong results across various text classification domains, including healthcare, finance, and insurance (Chen et al., 2024).

Although deep learning models like Convolutional Neural Networks (CNNs) (Kim, 2014) and Transformer-based models like BERT (Devlin et al., 2019) achieve excellent performance in text classification tasks, they typically require large labeled datasets and significant computational resources. Since this project focused on a relatively small and imbalanced dataset, deep learning methods were not adopted, though they remain a powerful direction for future work.

A significant challenge in review classification is class imbalance, where certain topics are underrepresented in the data. In recent years, Large Language Models (LLMs) have emerged as effective tools for generating synthetic data to address this issue. For example, the paper “Synthetic Data Generation with Large Language Models for Text Classification,” (Li et al., 2023) demonstrated that LLM-generated reviews improved classifier generalization and performance on underrepresented classes, especially when controlled for tone, vocabulary, and sentiment. However, their findings also highlight that synthetic data may fail to capture the full nuance and ambiguity of real customer feedback, particularly in subjective domains.

Additionally, many customer reviews reference multiple topics, even though they are often labeled with a single category during training. This presents a multi-label classification challenge, which has been widely studied in NLP.
To address label uncertainty, Conformal Prediction (CP) (Molnar, 2023) was used, a method that produces prediction sets rather than single label outputs, providing a principled way to quantify

uncertainty and handle ambiguous or multi-topic reviews (Angelopoulos et al., 2022; Shafer & Vovk, n.d.2008).

Lastly, interpretability is essential in regulated domains like insurance. While ensemble models such as XGBoost deliver strong performance, understanding how individual predictions are made can be challenging. To offer transparency, SHapley Additive exPlanations (SHAP) (SHAP, n.d.) values were used to identify which words or features most influenced each classification. SHAP offers both local explanations (for individual reviews) and global insights (across the dataset), making the model's decision-making process more understandable (Lubo-Robles et al., 2020).

3. Methodology

3.1. Dataset Description

The dataset consists of 1,943 customer reviews written in Portuguese, collected from Generali Tranquilidade through surveys, service evaluations, and complaint forms. Each review is a free-text response and is labeled with one of 11 predefined categories.

The label distribution is as follows:

- Atendimento: 465 (23.94%)
- Tempo de resposta e resolução: 262 (13.49%)
- Preço: 221 (11.38%)
- Documentação, burocracia e processo: 220 (11.33%)
- Sinistros: 210 (10.81%)
- Não Especificado - Neutro: 176 (9.06%)
- Comunicação central: 94 (4.84%)
- Pagamentos: 93 (4.79%)
- Não Especificado - Reclamações: 90 (4.63%)
- Produto: 57 (2.94%)
- Ferramentas Digitais: 54 (2.78%)

This distribution shows that categories like *Atendimento* dominate the dataset, while others such as *Produto* and *Ferramentas Digitais* are underrepresented, posing challenges for model training.

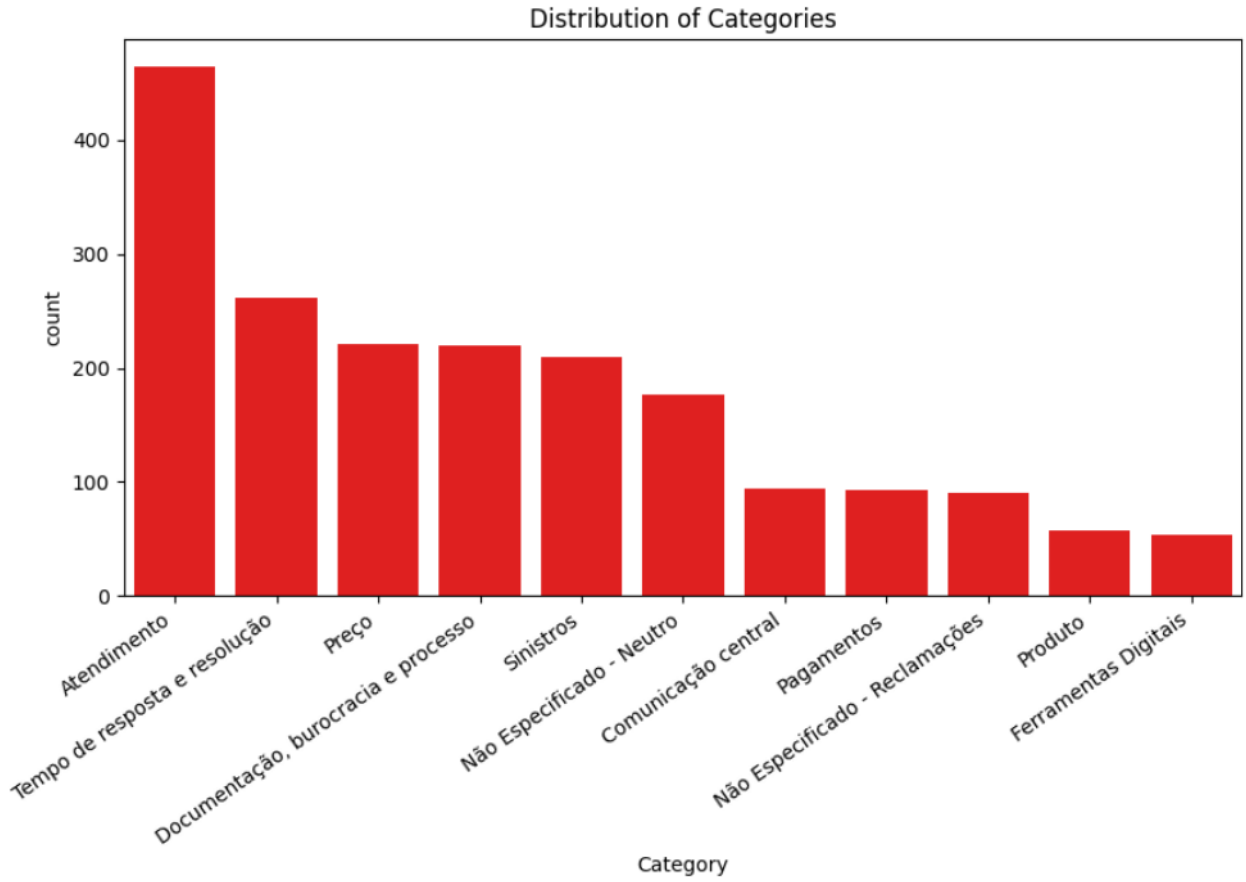


Figure2: Distribution of customer reviews

3.2. Challenges of the dataset

The dataset presents several challenges that impact the classification process:

- **Class Imbalance**: The distribution of review categories is uneven. For example, *Atendimento* (customer service) alone accounts for nearly 24% of all reviews, while categories such as *Ferramentas Digitais* (digital tools) and *Produto* (product-related feedback) make up less than 3% each. This imbalance can bias the model toward predicting the more frequent classes and lead to poor performance on minority classes, necessitating techniques like synthetic data generation to increase these minority categories.
- **Multi-Topic Reviews**: Many reviews cover multiple topics making it difficult to assign a single clear label.
- **Labeling Errors**: Manual labeling may introduce inconsistencies or errors, particularly for ambiguous reviews, affecting model training and evaluation.

- Spelling Errors: Since the reviews are written by customers, frequent spelling mistakes and informal expressions were present.

To address these challenges, the methodology incorporates synthetic data generation to balance the dataset, Conformal Prediction to manage uncertainty, and advanced preprocessing techniques tailored to Portuguese text, ensuring a robust classification system.

Understanding these challenges was key to selecting appropriate preprocessing, modeling, and evaluation strategies, which are discussed in the following sections.

3.3. Synthetic Data Generation Using LLMs

3.3.1. Understanding Large Language Models (LLMs):

Large Language Models (LLMs) (*Using LLMs for Synthetic Data Generation*, n.d.), such as GPT-4, Claude 3.5, and LLaMA, are advanced machine learning systems trained on vast and diverse text data to understand and generate human-like language. Because of how they are designed, these models excel at capturing context, semantics, and structure, making them produce coherent, diverse, and contextually relevant text when given a suitable prompt. Their ability to generate high quality text and code makes them valuable tools for synthetic data generation, particularly in scenarios where obtaining real data is costly or infeasible. For instance, LLMs can be prompted to create labeled examples for training classifiers. This synthetic data has shown promise in enhancing model performance where it helps in reducing annotation costs, and supporting data augmentation for improved robustness.

In this project, synthetic data generation using LLMs was applied to improve text classification of customer reviews in the insurance domain. The following sections describe the selected model, generation process, and evaluation strategy (Nadas et al., 2025) (Whitfield, n.d.).

3.3.2. Benefits of Using LLMs for Data Augmentation:

For tasks like text classification, where labeled data is often limited or imbalanced, LLMs offer an efficient solution by generating realistic synthetic data (Li et al., 2023). This has multiple benefits:

- **Improved Class Balance**: LLMs can generate additional examples for underrepresented categories, helping to mitigate class imbalance.
- **Enhanced Model Generalization**: Exposure to more varied and expressive examples improves the ability of models to generalize to unseen data.
- **Time Efficiency**: Reduces manual effort in data annotation and expansion.
- **Domain Adaptation**: LLMs can be guided to write in a specific tone, vocabulary, or perspective.

3.3.3. Model Used:

For this project, the Claude 3.5 model by Anthropic was used (*Anthropic's Claude 3.5 Sonnet*, 2024). This model was selected due to its advanced natural language understanding capabilities, control over output diversity, and alignment with human-like reasoning. Claude 3.5 demonstrated a high ability to generate text that mimics authentic customer reviews in both tone and content.

3.3.4. Integration with the Company's Internal Library:

To manage the generation process efficiently and ensure reproducibility, the company (Generali Tranquilidade) provided access to an internal Python-based LLM library. This library was used to:

- Define and create a prompts for generating reviews
- Generate synthetic customer reviews across different categories
- Adjust model parameters (such as temperature or token limits)

With this infrastructure in place, the next step was to design effective prompts to guide the generation of realistic and category-aligned synthetic reviews.

3.3.5. Prompt Design and Customization:

The prompt engineering process played a central role in ensuring that the generated reviews aligned with the requirements of each category. Unlike traditional approaches where LLMs are prompted to classify or analyze existing text, this method focused on generating synthetic reviews from scratch. The prompt was carefully designed to guide the model in producing realistic, category-specific feedback. Key elements included a clear and simple instruction asking the model to "*write as if you were a customer giving an insurance-related review*". To promote diversity in the output, the model was instructed to vary the structure and wording of each review.

The prompt also provided detailed context about each category along with a few real customer reviews as examples to keep the tone and vocabulary.

To guide the generation process, a single carefully designed system prompt was used across all categories. This prompt instructed the model to take on the role of a customer leaving feedback in European Portuguese and to base each review on the provided topics and their real examples. Reviews were constrained to focus on one specific issue at a time, replicate the tone and sentiment of real feedback, and follow the typical length and structure of actual customer comments. These constraints helped ensure the generated reviews were realistic, varied, and contextually aligned with genuine customer experiences (Amanatullah, 2024)(Halterman, 2025).

3.3.6. Temperature and Sampling Control:

Temperature is a hyperparameter that controls the randomness and creativity in a language model's output. Adjusting this value influences how predictable or varied the generated text will be. (Peeperkorn et al., 2024)

- At **temperature = 0**, the model behaves in a deterministic way, always choosing the most likely next word.
  For example, if prompted with *"The claims process was…"*, it might consistently respond with *"…slow and frustrating,"* because that's the most statistically probable response. This setting ensures consistency and coherence, but the outputs tend to be repetitive and lack variation.

- At **temperature = 1**, the model introduces maximum randomness, leading to more creative but less predictable responses.
  Using the same prompt *"The claims process was…"*, the model might generate responses like *"…a complete nightmare,"* *"…handled decently,"* or even *"…confusing but manageable."* While this can lead to creative and varied responses, it also increases the risk of incoherence or off-topic text, making it less reliable for structured tasks like synthetic data generation.

In this project, a temperature of 0.7 was selected, striking a balance between consistency and diversity. It was low enough to ensure that the generated reviews stayed relevant and coherent, but high enough to introduce natural variation and avoid repetitive phrasing. This allowed the model, Claude 3.5, to produce realistic and distinct customer reviews that aligned with the tone and style of real feedback.

3.3.7. Category Explanations:

To generate realistic content, it was important that the model understood the explanation and the key concerns for each category. For example:

- **Sinistros (Claims)**: This category covers difficulties associated with submitting and resolving a claim, from the moment it occurs, until the final decision by Generali Tranquilidade, including decisions by the company or intervention by third parties (experts, assistance, etc.).

- **Atendimento: (Service):** This category covers difficulties in service, including the attitude of the call center operator, empathy, information provided, knowledge and professionalism.

- **Preço (Pricing): T**his category covers the price of the insurance (increases, different amounts, incorrect amounts, etc,)

- **Pagamentos (Payments)**: This category covers late payments, charges earlier than expected, incorrect bank details.

- **Ferramentas Digitais (Digital Tools)**: This category covers comments about digital tools (website, customer area, WhatsApp, chatbot…).

Each category prompt included a brief definition and pain point list, ensuring the LLM generated text aligned with real-world concerns and customer language.

3.3.8. Example Review Usage

To ensure that synthetic reviews closely mirror the language and tone of real customer feedback, a diverse set of examples of real reviews was included in the prompt. These examples were carefully selected to represent a wide range of expressions, maintain clarity and structure, and align with each specific review category. By incorporating at least one example per category, the model was better equipped to replicate the vocabulary, sentiment, and stylistic patterns typical of real insurance customers.

3.3.9. Evaluation of Generated Reviews

To ensure the quality and category alignment of the synthetic reviews, an additional validation step using a separate LLM Claude 3.5 Haiku was conducted. This model was prompted to classify each generated review into one of the predefined categories, and only the reviews that were correctly categorized were retained, representing approximately 70% of the total synthetic ones. The evaluation prompt was designed to enforce strict classification, instructing the model to return only the category name or “not sure” if the review did not clearly fit any category. The prompt read:

“Classify the following review into one of the predefined categories. The output should only be the category name, with no explanation needed. There should be only one category for each review. If the review does not clearly belong to any category, respond with 'not sure'. Predict the reviews based on the common words and explanation of each category below.”

This evaluation step helped ensure that the synthetic reviews were both realistic and accurately categorized, ultimately improving the reliability of the augmented dataset for training classification models.

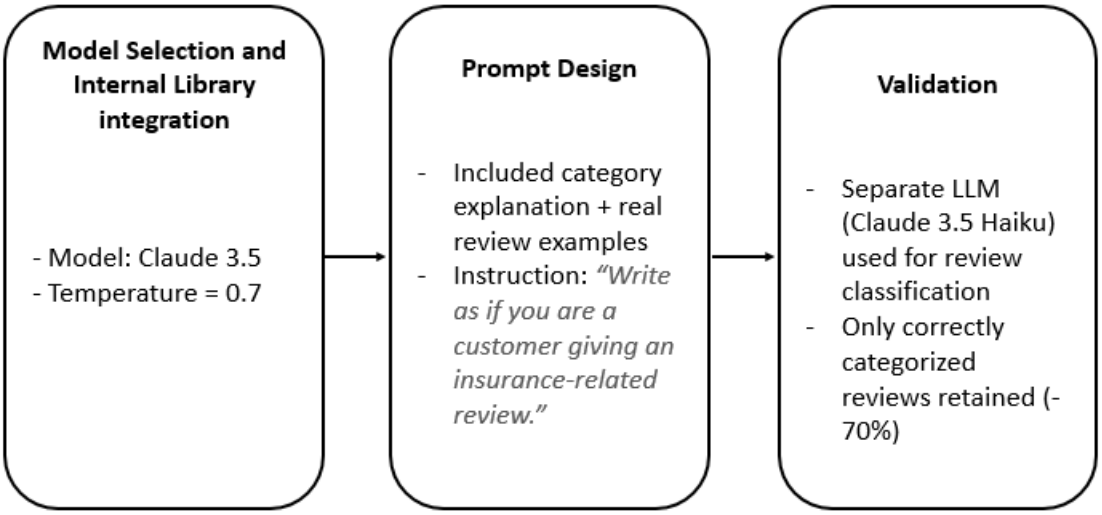


Figure 3: Synthetic Review Generation Workflow

3.3.10. Limitations and Considerations

4. While synthetic data helped balance the dataset, it comes with some limitations. The generated reviews may reflect biases from the prompt and examples, potentially missing the full range of natural language variation. There's also a risk that the model could learn patterns specific to the synthetic data and not perform well on real reviews. Additionally, since these reviews aren't written by real users, it can be difficult to know how realistic or representative they truly are.

This section describes the full pipeline used to classify customer reviews into predefined categories. The process includes four main components: text preprocessing, feature extraction, model training for classification, and uncertainty estimation using Conformal Prediction. Each step is designed to handle specific challenges, such as text variability, data imbalance, and label ambiguity. The following subsections explain each of these components in detail.

3.4. Text Preprocessing

After combining the real and synthetic reviews into a unified dataset, the next essential step was preprocessing the text. Text preprocessing is a crucial step in preparing data for Natural Language Processing (NLP) tasks, ensuring that the text is clean, consistent, and structured for analysis. This process involves several key steps:

3.4.1. Error Correction

Some customer reviews contained typos or shortened words that could affect the quality of the analysis. To improve the consistency of the data, common mistakes were corrected, and abbreviations were expanded, for example, "app" was replaced with "aplicação" and “doc” for “documento”.

3.4.2. Text Cleaning

This step involves removing irrelevant elements such as punctuation, special characters, and extra spaces, as well as converting all text to lowercase to maintain consistency.

Lowercasing helps treat words like “Seguro” and “seguro” as the same, reducing variability in the data. By minimizing unique token counts, where a token is typically a word or part of a word used by the model to process language, and by standardizing the format, this step makes the data easier to handle and improves the model’s efficiency.

3.4.3. Stop Word Removal

Stop-words are common words that frequently appear in a language but add little value in classification tasks. Examples in Portuguese include *"e," "de,"* and *"o."* Removing these words helps reduce noise in the data, allowing the model to focus on more meaningful content. Stop-word removal can be done using predefined lists or adaptive algorithms that dynamically filter out frequent but uninformative words (Kadhim, 2019) (GeeksforGeeks, 2025).

3.4.4. Lemmatization

Lemmatization reduces words to their root or base form, helping to decrease vocabulary size and simplify text analysis. This process ensures that different word variations are treated as a single entity. For example, the word *"pagadora"* would be lemmatized to *"pagar."* By standardizing word forms, lemmatization improves the consistency of textual data (Abidin et al., 2024).

3.4.5. Tokenization

Tokenization breaks text into smaller units called tokens, typically individual words or phrases, which is essential for subsequent NLP tasks such as vectorization and classification. By segmenting full reviews and sentences into separate tokens, we create a structured representation that machine learning models can process. For example, the review "O atendimento foi rápido e eficiente" would be tokenized into ["O", "atendimento", "foi", "rápido", "e", "eficiente"]. This step ensures that the text is broken

down into meaningful units, facilitating accurate feature extraction and classification(Mielke et al., 2021).

With the text now cleaned, standardized, lemmatized and tokenized, the next step involved converting it into a numerical format that machine learning models can understand.

3.5. Feature Extraction

To make textual data understandable to machine learning models, it is necessary to transform it into a numerical representation. This was achieved using a two-step approach: TF-IDF vectorization and custom term weighting.

3.5.1. TF-IDF Vectorization

The **TF-IDF (Term Frequency-Inverse Document Frequency)** method was first used to convert textual information into numerical format. TF-IDF helps assess the importance of words by assigning scores to words based on how frequently they appear in a document (TF) and how rare they are across all documents in the dataset (IDF) (Feldges, 2022).

TF-IDF is based on two key components:

1. **Term Frequency (TF):** Measures how often a word appears in a document. Words that occur more frequently within a document are considered more significant.
2. **Inverse Document Frequency (IDF):** Evaluates the importance of a word across the entire corpus. Words that appear in many documents receive lower importance, whereas words that appear in fewer documents are considered more informative.

By balancing word frequency within a document against its rarity across all documents, TF-IDF ensures that common words like *"e"* or *"a"* do not dominate the analysis and words like **"reembolso"** (refund) or **"assistência"** (assistance) receive a higher score if they appear in fewer documents but they're more central to a specific context (Polpinij & Luaphol, 2021).

To further enrich the representation of textual content, **n-grams** were incorporated. An n-gram is a sequence of *n* words that appear together in the text. For example, a **bigram** (n=2) like *"serviço rápido"* retains more semantic meaning than analyzing *"serviço"* and *"rápido"* separately, as it captures the context and relationships between words. This is particularly useful for short texts like customer reviews, where small phrases can carry significant meaning. Including n-grams allows the model to better understand frequent patterns and expressions specific to each category, thereby improving classification accuracy (Cavnar & Trenkle, 2001).

3.5.2. Custom term Weighting

While TF-IDF gives general importance to words based on their frequency, it doesn't consider how important certain words are for specific categories. To improve this, **Custom Term Weighting** was used. This means some important words were given extra weight manually because they are especially more relevant to certain types of topics (Benjamin et al., 2008)(Soucy & Mineau, 2005)(AC, 2019). For instance:

- In the category **"Ferramentas Digitais"**, words like *"internet"*, *"WhatsApp"*, and *"virtual"* were given much higher weights because they often appeared in reviews about problems with mobile tools and digital services.
- In the category **"Preço"**, words such as *"valor"*, *"aumento"*, and *"preço"* were given more importance since they were more linked to complaints about rising costs.
- For **"Documentação, Burocracia e Processo"**, words like *"burocracia", "documento",* and *"processo"* were given more importance because they showed up frequently in reviews about delays and administrative issues.
- In the **"Sinistros"** category, terms such as *"reboque", "assistência",* and *"sinistro"* helped identify reviews related to claims.

The same approach was used for other categories like **"Atendimento"**, **"Pagamentos"**, and **"Tempo de Resposta e Resolução"**, where specific words were highlighted to help the model better understand the topic of each review.
Rather than relying solely on TF-IDF, Custom Term Weighting helps refine classification, making the model more sensitive to key terms that better define each category. (Rathi & Mustafi, 2023)

3.6. Text Classification Pipeline

Automatic text classification is a supervised machine learning problem where a model learns from labeled data to assign new texts to predefined categories. In this study, this technique was implemented to classify customer reviews efficiently.

3.6.1. Dataset Splitting Strategy

To train and evaluate the models, the labeled dataset was split as follows:

- **20%** of the **real customer reviews** (388 out of 1942 reviews) were used as the **test set** to evaluate the model's performance on real, unseen data.

- The **remaining 80%** of real reviews (1,554 reviews) were used for training.

- In addition, **1,727 synthetic reviews** generated using a LLM were included in the **training set** to improve performance and increase the diversity of training examples.

This combination of real and synthetic data in training aimed to help the model generalize better to different types of reviews while keeping the test set entirely composed of real customer inputs for reliable evaluation.

3.6.2. Baseline Model Comparison

To select the most appropriate model, several supervised learning algorithms were tested on the same dataset with TF-IDF features. Performance was evaluated using two main metrics:

- **Accuracy**: the proportion of correct predictions among all predictions made.
- **F1-score**: a metric that combines precision (how many predicted labels were actually correct) and recall (how many of the true labels the model correctly found). It balances both.

The table below summarizes their performance based on the accuracy.

Model	Accuracy
Logistic Regression	0.50
Naive Bayes	0.39
Random Forest	0.52
Support Vector Machine (SVM)	0.55
XGBoost	**0.57**

Table 1: Accuracy Comparison of Baseline Models

XGBoost was selected as the final model due to its superior performance, robustness to overfitting, and compatibility with imbalanced data. Additionally, XGBoost trains quickly and efficiently, which makes it a good choice for running many experiments. As an ensemble learning method, XGBoost builds multiple decision trees in sequence, where each tree aims to correct the errors of the previous one by optimizing a loss function through gradient boosting. These characteristics make XGBoost especially well suited for high dimensional, sparse data such as TF-IDF representations of text (Rusjayanthi et al., 2023)(Fotache, 2023)(Collado-Montañez et al., 2024).

3.6.3. Hyperparameter Tuning with Bayesian Optimization

To improve the performance of the XGBoost model, Bayesian optimization was applied to fine-tune its hyperparameters efficiently. Unlike random search methods (Hassan, 2023), Bayesian

optimization (Putatunda & Rama, 2019) uses past evaluations to guide the search toward the most promising parameter combinations, reducing computational cost while maximizing model accuracy.

The search focused on key parameters:

- **n_estimators:** the number of trees in the model, which affects the complexity and learning capacity.
- **max_depth:** the maximum depth of each tree, controlling how detailed the splits can be and affecting model complexity.
- **learning_rate:** the step size at each boosting iteration, balancing how much each tree contributes to the final model.
- **subsample:** the fraction of training samples used to grow each tree, which helps prevent overfitting by adding randomness.
- **colsample_bytree:** the fraction of features (columns) randomly selected for each tree, also used to reduce overfitting.

The best configuration found included 435 estimators, a learning rate of 0.048, a maximum depth of 6, and carefully selected sampling ratios, resulting in an accuracy of approximately 59% on the test set.

3.6.4. Cross-Validation for Model Evaluation

To ensure the model's robustness and reduce the risk of overfitting, we applied cross-validation during the hyperparameter optimization phase. This involved dividing the training data into 5 folds, allowing the model to be trained and validated on different subsets. This approach helped assess the model's ability to generalize and ensured balanced performance across all review categories before any evaluation on the test set. Once the best hyperparameters were identified, the final XGBoost model was retrained on the full training set and then tested on unseen data. The optimized model achieved an accuracy of 0.59, reflecting a 0.02 improvement over the baseline. Notably, several categories showed performance gains, including Preço (F1-score: 0.72), Atendimento (F1-score: 0.64), and Sinistros (F1-score: 0.69). While some lower-frequency categories remained challenging, the overall classification report demonstrated more balanced results than before.(Chen et al., 2024)

3.7. Model Interpretability with SHAP Values

Understanding why a machine learning model makes certain predictions is particularly important in domains like insurance, where interpretability is essential for transparency, trust, and compliance. To address this, SHapley Additive exPlanations (SHAP) (SHAP n.d.) were used to

interpret the output of the XGBoost model. SHAP provides a principled framework for explaining predictions by assigning each feature a contribution value for a given output.

The following subsections describe how SHAP values work and how they were applied in this project.

3.7.1. Understanding SHAP Values

SHAP values are based on Shapley values, a concept from cooperative game theory that fairly distributes the "payout" among players based on their contributions to the total outcome. In the context of machine learning, each “player” is a feature (in this case, a word or token), and the “payout” is the model’s predicted probability for a specific class.

SHAP calculates how much each feature contributes to moving the model output from the base value (the average prediction) to the actual prediction for a given instance. It does this by checking how much each word (feature) changes the model’s prediction when it is included or left out, across different combinations of words. This helps fairly show how important each word really is.

One key advantage of SHAP is that it provides both local explanations (for individual predictions) and global explanations (summarizing feature importance across the entire dataset) in a unified framework. This makes it suitable for high-stakes applications where both individual decisions and overall model behavior need to be interpreted (Zhao et al., 2021) (Lubo-Robles et al., 2020).

3.7.2. Interpretation of SHAP Values

In this project, SHAP was applied to explain the predictions made by the XGBoost classifier on customer review data. For each review, SHAP values identified which words had the greatest positive or negative impact on the final predicted category.

- At the **local level**, SHAP helped explain why a specific review was assigned to a certain category. For example, in a review labeled as "Preço" (Pricing), SHAP showed that words like “aumento”, “preço”, and “valor” contributed most to the classification.
- At the **global level**, SHAP summaries revealed which words were most important overall for each category. This allowed for the evaluation of whether the model was relying on meaningful, domain-relevant terms or being influenced by irrelevant or misleading features.

SHAP visualizations highlighted that terms like “sinistro” were strongly associated with the sinistros category, while word like “valor” was mostly linked to the ‘Pagamentos’ and ‘Preço’ categories.

These insights improve transparency and support error analysis, helping to better understand the model's behavior and increasing confidence in how it makes predictions. A limitation of SHAP is that it can be computationally intensive on large datasets, but in this case, its ability to provide clear and actionable explanations justified its use.

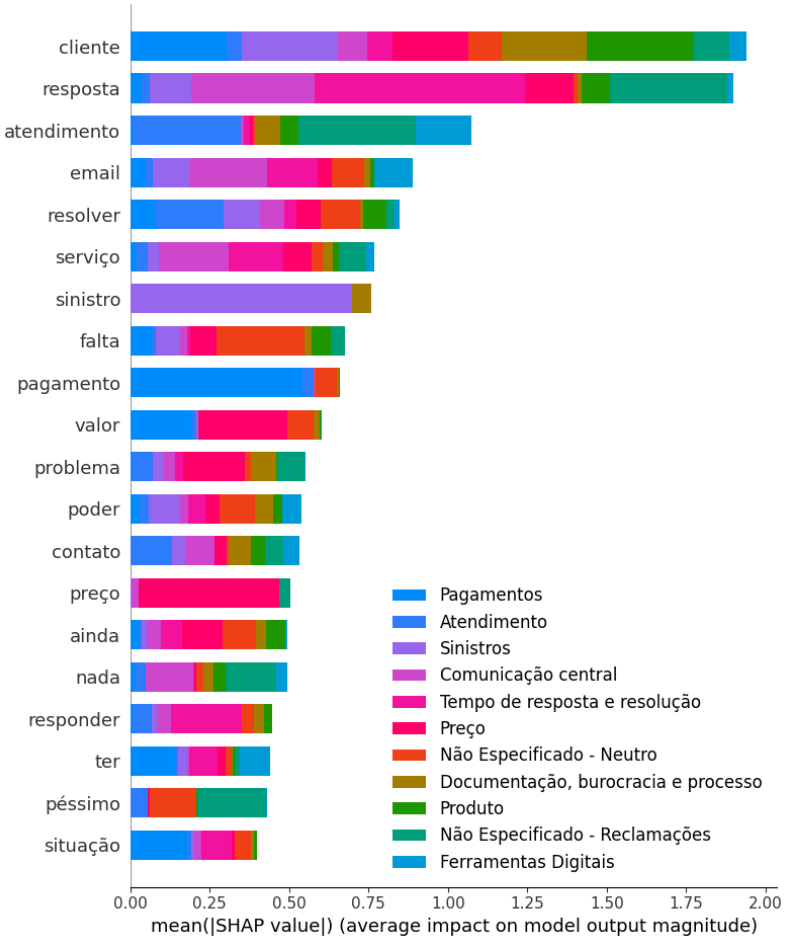


Figure 3: SHAP summary plot across all categories

3.8. Handling Uncertainty with Conformal Prediction

One of the challenges in classifying customer reviews is that some reviews cover multiple topics, making it difficult to assign a single definitive label. Additionally, there is uncertainty in some classifications, either due to ambiguous wording in the reviews or potential errors in the initial labeling process. To address these issues, we opted for Conformal Prediction instead of relying solely on classic machine learning models. (Angelopoulos et al., 2022) (Molnar, 2023)

Conformal Prediction is a framework designed to quantify uncertainty in machine learning predictions. Unlike traditional classifiers that provide a single label, conformal prediction generates prediction sets, which include multiple possible labels with a certain confidence level.

This ensures that the true label is within the prediction set with a specified probability, set to **95% confidence** in our case. (Shafer & Vovk, 2008)

The implementation involved wrapping the XGBoost classifier in a conformal prediction framework, using a non-conformity measure (Kato et al., 2023) to construct prediction sets. This allowed us to take advantage of XGBoost's strong predictive performance while incorporating a measure of uncertainty into our predictions.
This approach was particularly valuable for ambiguous reviews, where a single-label prediction might be misleading. For example, a review stating "great app but slow response" might yield a prediction set {ferramentas digitais, tempo de resposta e resolusao}, reflecting uncertainty in categorization.
Two **key metrics were used to evaluate this method:**

i. **Coverage:** Coverage represents the proportion of true labels that are included in the prediction sets. For example, if the coverage is **95%**, it means that in 95% of cases, the correct label is included in the predicted set. Ensuring high coverage means the model is less likely to miss the correct label.

ii. **Set Size:** The **average set size** indicates how many labels, on average, are included in each prediction set. A **smaller set size** makes predictions more precise. For instance, if the average set size is **2**, it means that each review is typically assigned **two possible labels** rather than a single one, reflecting the uncertainty in classification.

This method enhanced the trustworthiness of the classification system, aligning with the research goal of providing reliable subject-based analysis for business decision-making.

Category	Coverage	Average Set Size
Atendimento	0.96	2.70
Comunicação central	0.37	2.84
Documentação, burocracia e processo	0.86	2.82
Ferramentas Digitais	0.18	2.91
Não Especificado - Neutro	0.91	2.83
Não Especificado - Reclamações	0.72	2.72
Pagamentos	0.74	2.63
Preço	0.86	2.00
Produto	0.36	2.82

Sinistros	0.76	2.48
Tempo de resposta e resolução	0.79	2.57

Table 2: Coverage and Average set size per category

Overall, the model achieved a coverage of 79.9% with an average set size of 2.62, striking a balance between reliability and precision.

Category level results showed high coverage for Atendimento (0.96), Preço (0.86), and Documentação, burocracia e processo (0.86), while categories like Ferramentas Digitais (0.18) and Produto (0.36) were more challenging. The average set size ranged from 2.00 for Preço to 2.91 for Ferramentas Digitais, indicating that some categories required broader prediction sets to capture uncertainty.

By integrating **Conformal Prediction**, we improved the model's ability to handle ambiguous or multi-topic reviews while maintaining reliable classification performance. This approach ensures that when uncertainty exists, the model provides a broader set of likely categories rather than forcing a single, potentially incorrect label.

4. Experimental Results and Evaluation

a. Performance on Real vs. Synthetic Data

To assess the impact of synthetic data, the XGBoost classifier was evaluated across three configurations:

1. **Training and testing only on real data.**
2. **Training on both real and synthetic reviews and then testing only on real data.**
3. **Training and testing only on synthetic reviews.**

These experiments were designed to evaluate how synthetic data generation using LLMs could address challenges such as data imbalance, low coverage in minority categories, and label ambiguity.

Initially, training and testing were done using only real customer reviews. However, this resulted in poor performance, particularly for minority classes, with several categories having zero F1-score. These limitations prompted the generation of synthetic reviews using Claude 3.5, aiming to balance the dataset and enhance generalization. Finally, we tested performance using only synthetic reviews to assess how well the model could learn from this artificial but structured data.

1. Training and testing only on real data

Accuracy: **49%**

This configuration showed the weakest performance. Minority categories such as “Ferramentas Digitais” and “Produto” had F1-scores close to 0, demonstrating the model’s limited capacity to generalize from insufficient and imbalanced real data.
This scenario highlights the limitations of working with a small, imbalanced dataset. The lack of sufficient training data, especially for minority categories, caused the model to overfit to the majority classes and fail to generalize.

Classification report:

Category	**Precision**	**Recall**	**F1-Score**	**Support**
Atendimento	0.51	0.68	0.58	93
Comunicação central	0.23	0.15	0.18	19
Documentação, burocracia e processo	0.31	0.32	0.31	44
Ferramentas Digitais	0.10	0.09	0.09	11
Não Especificado - Neutro	0.28	0.40	0.33	35
Não Especificado - Reclamações	0.24	0.28	0.26	18
Pagamentos	0.29	0.32	0.30	19
Preço	0.52	0.65	0.58	44
Produto	0.00	0.00	0.00	11
Sinistros	0.66	0.38	0.48	42
Tempo de resposta e resolução	0.44	0.41	0.42	53
Accuracy			**0.49**	389

Table 3: Classification Report for Model Trained and Tested only on Real Data

2. Training on real and synthetic data, testing only on real data

Accuracy: **59%**
The model performed better than when trained only on real data, especially in minority classes, thanks to the synthetic examples. However, the classifier still struggled with ambiguity and mixed topic reviews typical in real world data.

Synthetic data helped reduce class imbalance and provided clearer examples for low-frequency classes. This improved overall F1-scores, especially for categories like Produto and Pagamentos. For instance, Produto had an F1-score of 0.00 when trained and tested on real data, which increased to 0.33 with synthetic augmentation. Similarly, Ferramentas Digitais improved from 0.09 to 0.15, and Pagamentos from 0.30 to 0.55. These improvements indicate that the synthetic

reviews helped the model better understand minority categories. Still, the real world complexity of customer reviews introduced noise that synthetic examples couldn't fully replicate.

Classification Report:

Category	**Precision**	**Recall**	**F1-Score**	**Support**
Atendimento	0.58	0.71	0.64	93
Comunicação central	0.33	0.21	0.26	19
Documentação, burocracia e processo	0.47	0.45	0.46	44
Ferramentas Digitais	0.50	0.09	0.15	11
Não Especificado - Neutro	0.40	0.54	0.46	35
Não Especificado - Reclamações	0.39	0.39	0.39	18
Pagamentos	0.52	0.58	0.55	19
Preço	0.67	0.77	0.72	44
Produto	0.43	0.27	0.33	11
Sinistros	0.86	0.57	0.69	42
Tempo de resposta e resolução	0.62	0.55	0.58	53
Accuracy			**0.59**	389

Table 4: Classification Report for Model Trained on real and synthetic data and Tested only on real data

3. Training and testing only on synthetic data

Accuracy: **97.4%**
The model performed exceptionally well, achieving near-perfect F1-scores across most categories. This is expected, as the synthetic reviews were clean, clearly labeled, and closely aligned with the prompt constraints.

This result confirms that the model can easily learn from highly structured synthetic data. However, it also highlights that strong performance on synthetic reviews does not guarantee reliable performance on real-world reviews, which often contain more noise, spelling mistakes, and overlapping topics.

Classification Report:

Category	Precision	Recall	F1-Score	Support
Comunicação central	0.92	0.92	0.92	24
Documentação, burocracia e processo	1.00	1.00	1.00	17
Ferramentas Digitais	1.00	0.95	0.97	20
Não Especificado - Neutro	0.95	0.98	0.96	53
Não Especificado - Reclamações	0.98	1.00	0.99	49
Pagamentos	1.00	0.96	0.98	50
Preço	1.00	0.94	0.97	32
Produto	0.99	0.99	0.99	72
Sinistros	0.94	1.00	0.97	29
Accuracy			**0.97**	346

Table 5: Classification Report for Model Trained and Tested only on Synthetic Data

These results demonstrate the valuable role of synthetic data in boosting classification performance, particularly for minority classes. However, even with augmentation, the model still struggled with real-world ambiguity and overlapping themes.

b. Reliability and Interpretability

The classifier's average accuracy on real reviews with synthetic augmentation was 59%. Majority categories such as Atendimento (F1-score: 64%) consistently outperformed minority ones like Ferramentas Digitais (F1-score: 15%).

To enhance model reliability under ambiguity, Conformal Prediction was applied, generating prediction sets instead of single labels. This approach achieved a coverage of **79.9%** with an average set size of 2.62, meaning the correct label appeared in the prediction set in most cases.

For interpretability, SHAP analysis revealed the most influential terms behind each classification. For example, "app" strongly contributed to Ferramentas Digitais predictions, while "expensive" influenced Preço. These insights help identify customer pain points but are computationally expensive when scaling to larger datasets.

Compared to a baseline Logistic Regression model (which achieved only 45% accuracy), XGBoost showed clear improvements. However, its limitations with minority classes persist due to scarce training data.

5. Conclusion

This study successfully developed an automated text classification system tailored to the needs of Generali Tranquilidade, providing a scalable and interpretable solution for analyzing customer feedback in the insurance domain. By leveraging Large Language Models (LLMs) for synthetic data generation, the system addressed class imbalance, enhancing the model's ability to handle underrepresented categories. The integration of Extreme Gradient Boosting (XGBoost) as the primary classifier, combined with Conformal Prediction for uncertainty quantification and SHAP values for interpretability, created a robust framework that effectively managed the complexities of real-world customer reviews, including ambiguity and multi-topic content.

The approach demonstrated significant improvements over traditional methods, particularly in handling minority classes and providing reliable uncertainty estimates, which are critical for decision-making in the regulated insurance industry. The use of SHAP values offered actionable insights into customer pain points, enabling Generali Tranquilidade to prioritize operational and service improvements effectively. However, challenges such as the inherent complexity of multi-topic reviews and limitations of synthetic data in capturing real-world variability highlight areas for further refinement.

The system's practical implications include enabling faster, data driven insights into customer feedback, which can inform targeted strategies for enhancing customer satisfaction and operational efficiency. Its scalability and interpretability make it a valuable tool for insurance companies facing similar challenges with unstructured feedback.

Future research should focus on adopting multi-label classification to better capture reviews spanning multiple topics, improving the quality of real-world data annotations, and exploring advanced LLMs or domain-specific fine-tuning to generate more representative synthetic data. Additionally, integrating real-time feedback mechanisms and expanding the dataset could further enhance the system's robustness and generalizability, ensuring its adaptability to evolving customer expectations and industry demands.

References

Abidin, Z., Junaidi, A., & Wamiliana. (2024). Text stemming and lemmatization of regional languages in Indonesia: A systematic literature review. *Journal of Information Systems Engineering and Business Intelligence, 10*(2), 217–231. https://doi.org/10.20473/jisebi.10.2.217-231

AC. (2019, May 26). Beyond feature selection: A term weighting approach in text classification. *Data Folks Indonesia*. https://medium.com/data-folks-indonesia/beyond-feature-selection-a-term-weighting-approach-in-text-classification-24821599bb2c

Amanatullah. (2024, December 19). Introducing the synthetic data generator—Build datasets with natural language. *Medium.* https://medium.com/@amanatulla1606/introducing-the-synthetic-data-generator-build-datasets-with-natural-language-897e5c00b204

Angelopoulos, A., Bates, S., Malik, J., & Jordan, M. I. (2022). Uncertainty sets for image classifiers using conformal prediction. https://doi.org/10.48550/arXiv.2009.14193

Anthropic's Claude 3.5 Sonnet: A new era of generative AI—Blogs on AI Technology. (2024, July 20). https://www.dailyaiblogs.com/claude-3-5-sonnet/

Benjamin, C. M. X., Woon, W. L., & Wong, K. S. D. (2008). Customized term weighting scheme for document classification. *2008 International Conference on Computer and Communication Engineering*, 294–299. https://doi.org/10.1109/ICCCE.2008.4580615

Cavnar, W. B., & Trenkle, J. M. (1994, April). N-gram based text categorization. https://www.researchgate.net/publication/2375544_N-Gram-Based_Text_Categorization

Chen, M., Wu, Y., Wingerd, B., Liu, Z., Xu, J., Thakkar, S., Pedersen, T. J., Donnelly, T., Mann, N., Tong, W., Wolfinger, R. D., & Bao, W. (2024). Automatic text classification of drug-induced liver injury using document-term matrix and XGBoost. *Frontiers in Artificial Intelligence, 7*, 1401810. https://doi.org/10.3389/frai.2024.1401810

Collado-Montañez, J., López-Úbeda, P., Chizhikova, M., Díaz-Galiano, M. C., Ureña-López, L. A., Martín-Noguerol, T., Luna, A., & Martín-Valdivia, M. T. (2024). Automatic text classification of prostate cancer malignancy scores in radiology reports using NLP models. *Medical & Biological Engineering & Computing, 62*(11), 3373–3383. https://doi.org/10.1007/s11517-024-03131-x

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*. https://arxiv.org/abs/1810.04805

Feldges, C. (2022, April 2). Text classification with TF-IDF, LSTM, BERT: A quantitative comparison. *Medium.* https://medium.com/@claude.feldges/text-classification-with-tf-idf-lstm-bert-a-quantitative-comparison-b8409b556cb3

Fotache, C. (2023, November 29). Text classification in Python: Pipelines, NLP, NLTK, transformers, XGBoost and more. *Medium.* https://chrisfotache.medium.com/text-classification-in-python-pipelines-nlp-nltk-tf-idf-xgboost-and-more-b83451a327e0

GeeksforGeeks. (2025, July 26). Removing stop words with NLTK in Python. *GeeksforGeeks.* https://www.geeksforgeeks.org/nlp/removing-stop-words-nltk-python/

Halterman, A. (2025). Synthetically generated text for supervised text analysis. *Political Analysis,* 1–14. https://doi.org/10.1017/pan.2024.31

Hassan, M. H. (2023, November 25). Tuning model hyperparameters with random search. *Medium.* https://medium.com/@hammad.ai/tuning-model-hyperparameters-with-random-search-f4c1cc88f528

Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review, 52*(1), 273–292. https://doi.org/10.1007/s10462-018-09677-1

Kato, Y., Tax, D. M. J., & Loog, M. (2023). A review of nonconformity measures for conformal prediction in regression. *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications,* 369–383. https://proceedings.mlr.press/v204/kato23a.html

Kim, Y. (2014). Convolutional neural networks for sentence classification. *EMNLP.* https://arxiv.org/abs/1408.5882

Li, Z., Zhu, H., Lu, Z., & Yin, M. (2023). Synthetic data generation with large language models for text classification: Potential and limitations (arXiv:2310.07849). *arXiv.* https://doi.org/10.48550/arXiv.2310.07849

Lubo-Robles, D., Devegowda, D., Jayaram, V., Bedle, H., Marfurt, K. J., & Pranter, M. J. (2020, September). Machine learning model interpretability using SHAP values: Application to a seismic facies classification task. In *Proceedings of the SEG Annual Conference* (pp. 1460–1464). *Society of Exploration Geophysicists.* https://www.researchgate.net/publication/344504497_Machine_Learning_model_interpretability_using_SHAP_values_application_to_a_seismic_facies_classification_task

Lundberg, S. (2018). An introduction to explainable AI with Shapley values—SHAP latest documentation. (n.d.). *SHAP documentation.* Retrieved March 19, 2025, from https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html

Mielke, S. J., Alyafeai, Z., Salesky, E., Raffel, C., Dey, M., Gallé, M., Raja, A., Si, C., Lee, W. Y., Sagot, B., & Tan, S. (2021). Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. https://doi.org/10.48550/arXiv.2112.10508

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning based text classification: A comprehensive review (arXiv:2004.03705). *arXiv.* https://doi.org/10.48550/arXiv.2004.03705

Molnar, C. (n.d.). Introduction to conformal prediction with Python.

Nadas, M., Diosan, L., & Tomescu, A. (2025). Synthetic data generation using large language models: Advances in text and code. https://doi.org/10.48550/arXiv.2503.14023

Peeperkorn, M., Kouwenhoven, T., Brown, D., & Jordanous, A. (2024). Is temperature the creativity parameter of large language models? https://doi.org/10.48550/arXiv.2405.00492

Polpinij, J., & Luaphol, B. (2021). Comparing multi-class text classification methods for automatic ratings of consumer reviews. In *Multi-Disciplinary Trends in Artificial Intelligence: 14th International Conference, MIWAI 2021* (pp. 164–175). https://doi.org/10.1007/978-3-030-80253-0_15

Putatunda, S., & Rama, K. (2019). A modified Bayesian optimization based hyper-parameter tuning approach for extreme gradient boosting. *2019 Fifteenth International Conference on Information Processing (ICINPRO),* 1–6. https://doi.org/10.1109/ICInPro47689.2019.9092025

Rathi, R. N., & Mustafi, A. (2023). The importance of term weighting in semantic understanding of text: A review of techniques. *Multimedia Tools and Applications, 82*(7), 9761–9783. https://doi.org/10.1007/s11042-022-12538-3

Rusjayanthi, N. K. D., Sudana, A. A. K. O., & Trisna, I. N. P. (2023). Text classification system using text mining with XGBoost method. *Jurnal Ilmiah Merpati (Menara Penelitian Akademika Teknologi Informasi), 11*(2), 61.

Shafer, G., & Vovk, V. (2007, June). A tutorial on conformal prediction. https://arxiv.org/pdf/0706.3188

Soucy, P., & Mineau, G. W. (2005). Beyond TF-IDF weighting for text categorization in the vector space model. https://dl.acm.org/doi/10.5555/1642293.1642474

Vongthongsri, K. (2025, January). Using LLMs for synthetic data generation: The definitive guide. *Confident AI.* https://www.confident-ai.com/blog/the-definitive-guide-to-synthetic-data-generation-using-llms

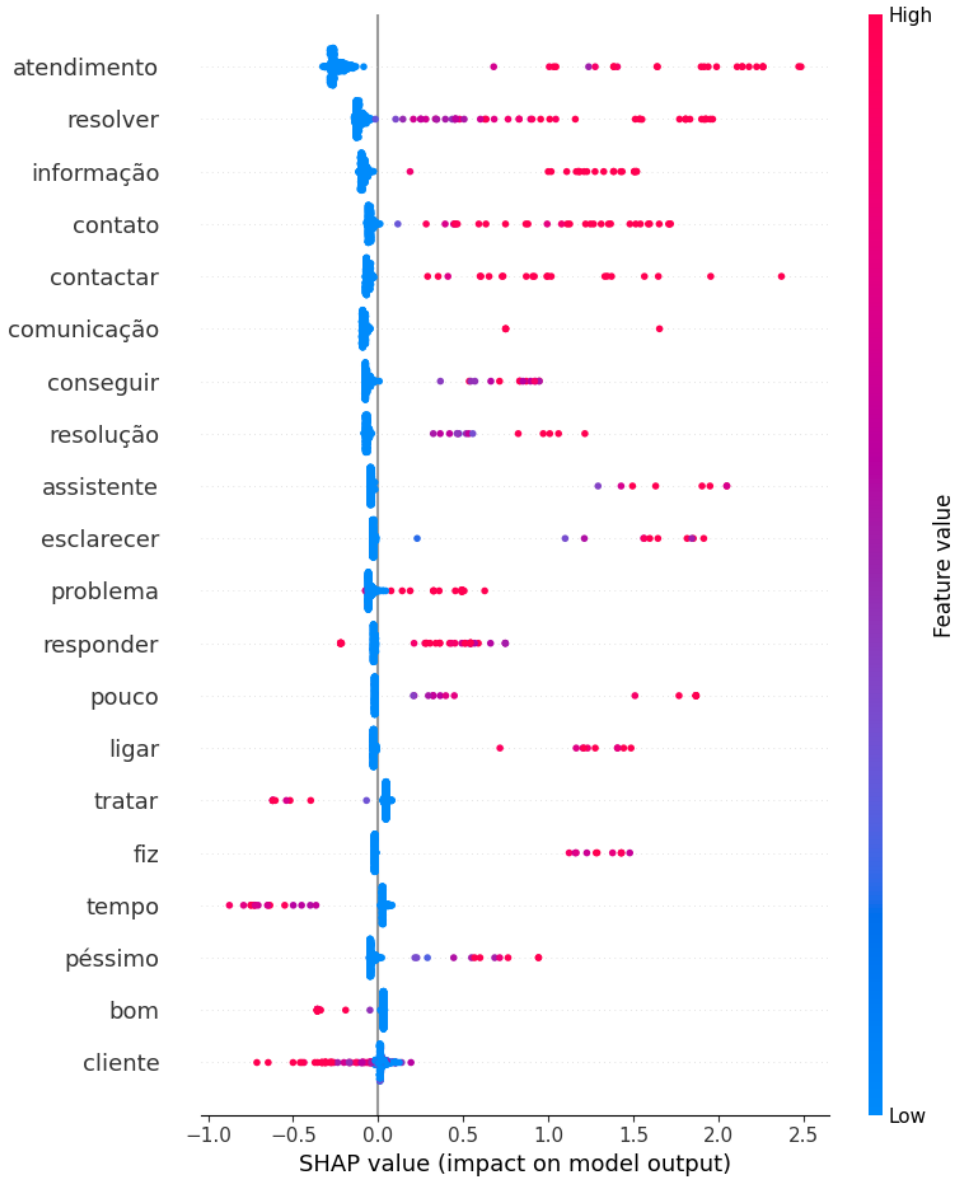
Whitfield, D. (2020, December). Using GPT-2 to create synthetic data to improve the prediction performance of NLP machine learning classification models. https://www.researchgate.net/publication/351046833_Using_GPT2_to_Create_Synthetic_Data_to_Improve_the_Prediction_Performance_of_NLP_Machine_Learning_Classification_Models

Yadav, R. K., Madaan, A., & Janu. (2024). Comprehensive analysis of natural language processing. *Global Journal of Engineering and Technology Advances, 19*(1), 83–90. https://doi.org/10.30574/gjeta.2024.19.1.0058

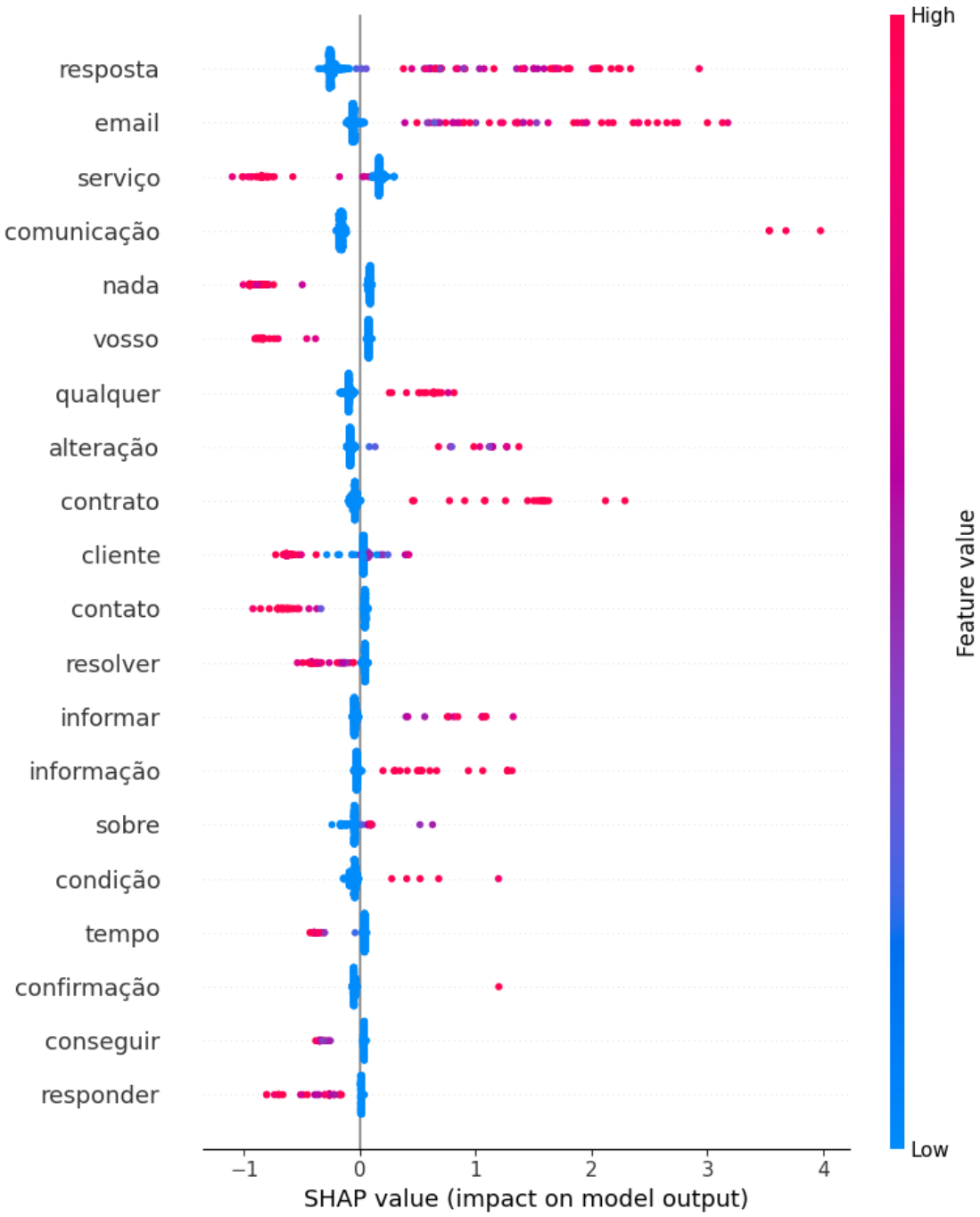
Zhao, W., Joshi, T., Nair, V. N., & Sudjianto, A. (2021). SHAP values for explaining CNN-based text classification models. https://doi.org/10.48550/arXiv.2008.11825

Appendix

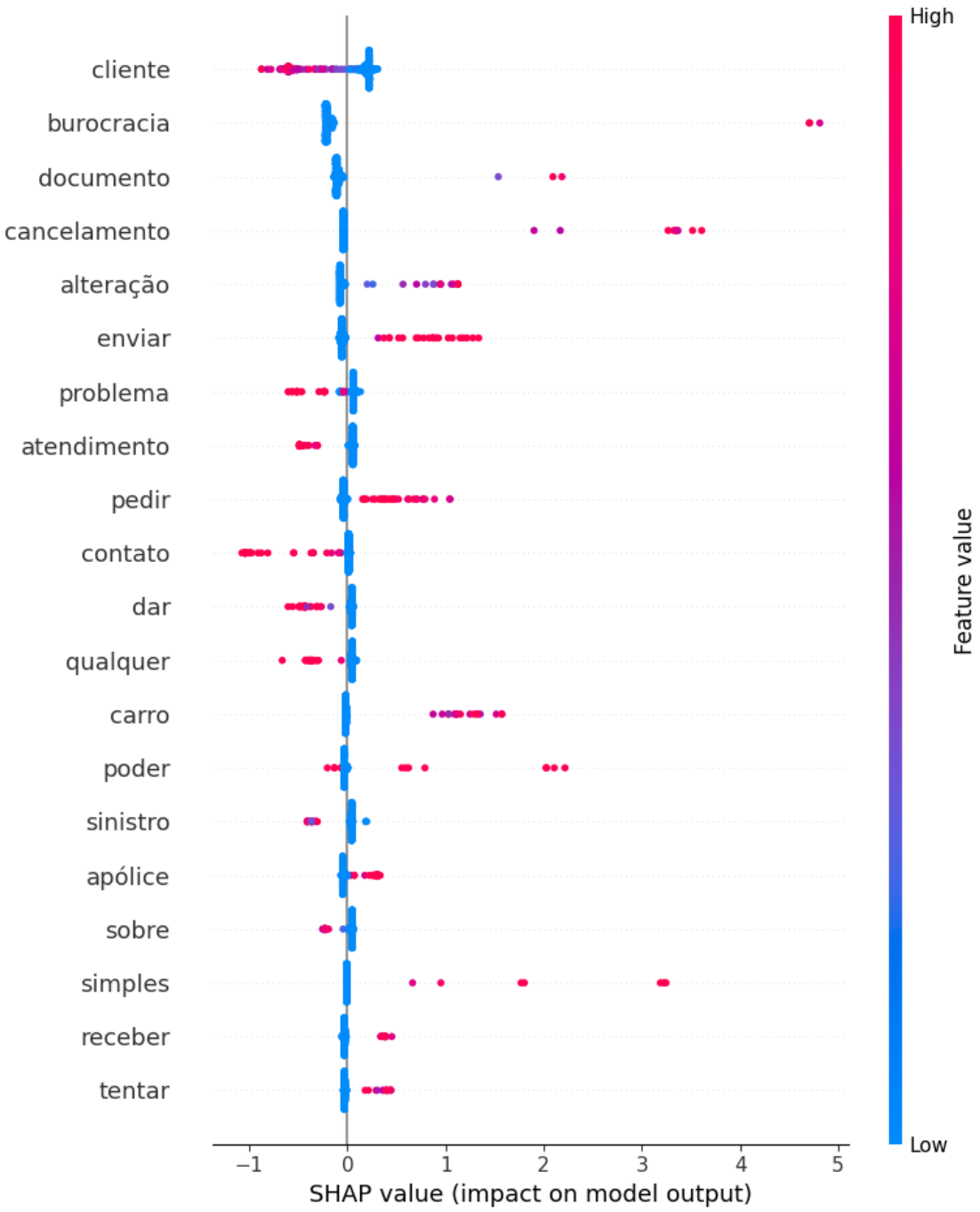
SHAP Value Analysis of Feature Importance Across Customer Review Categories



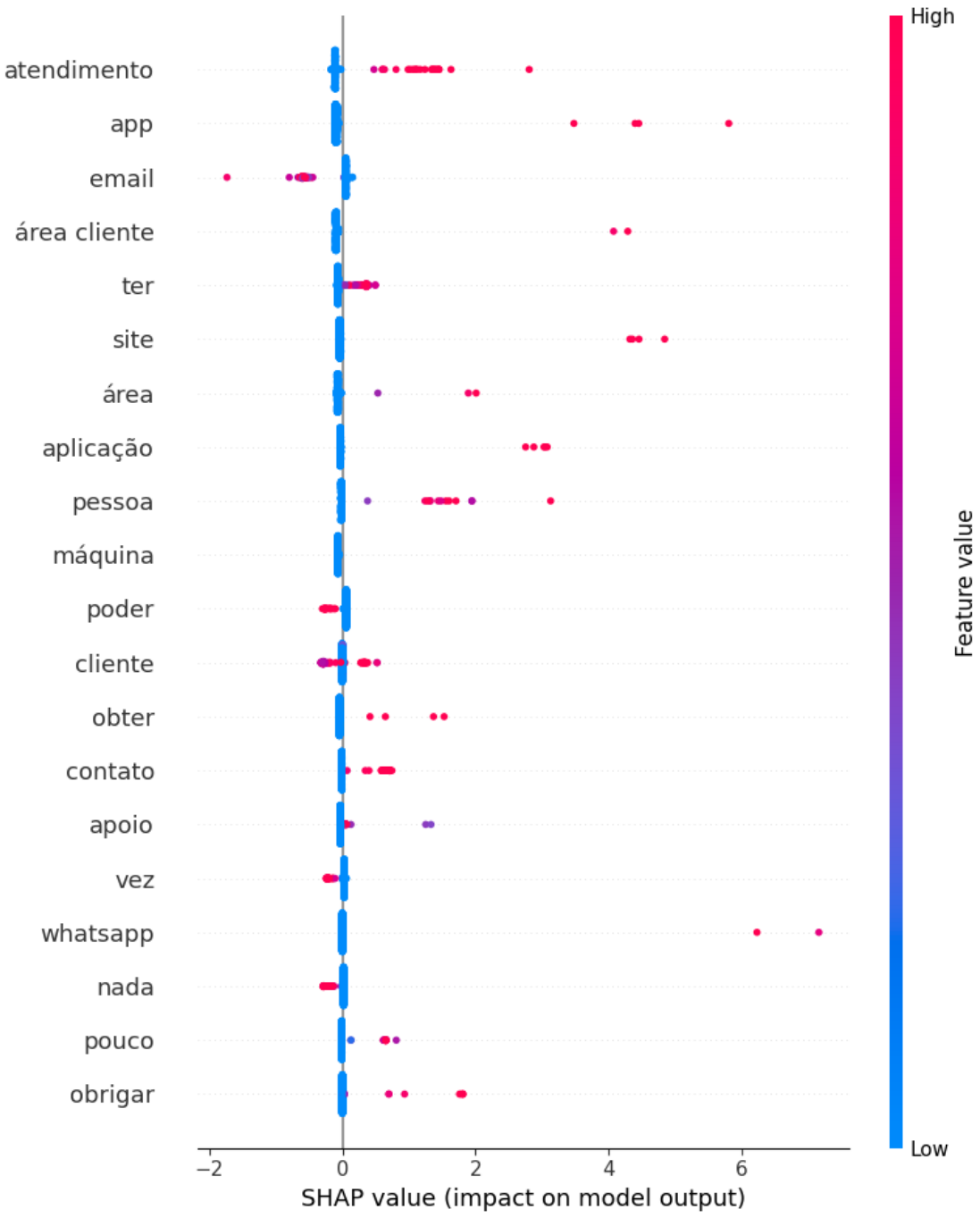
*SHAP Value Analysis of Feature Importance for **Atendimento** Category Predictions*



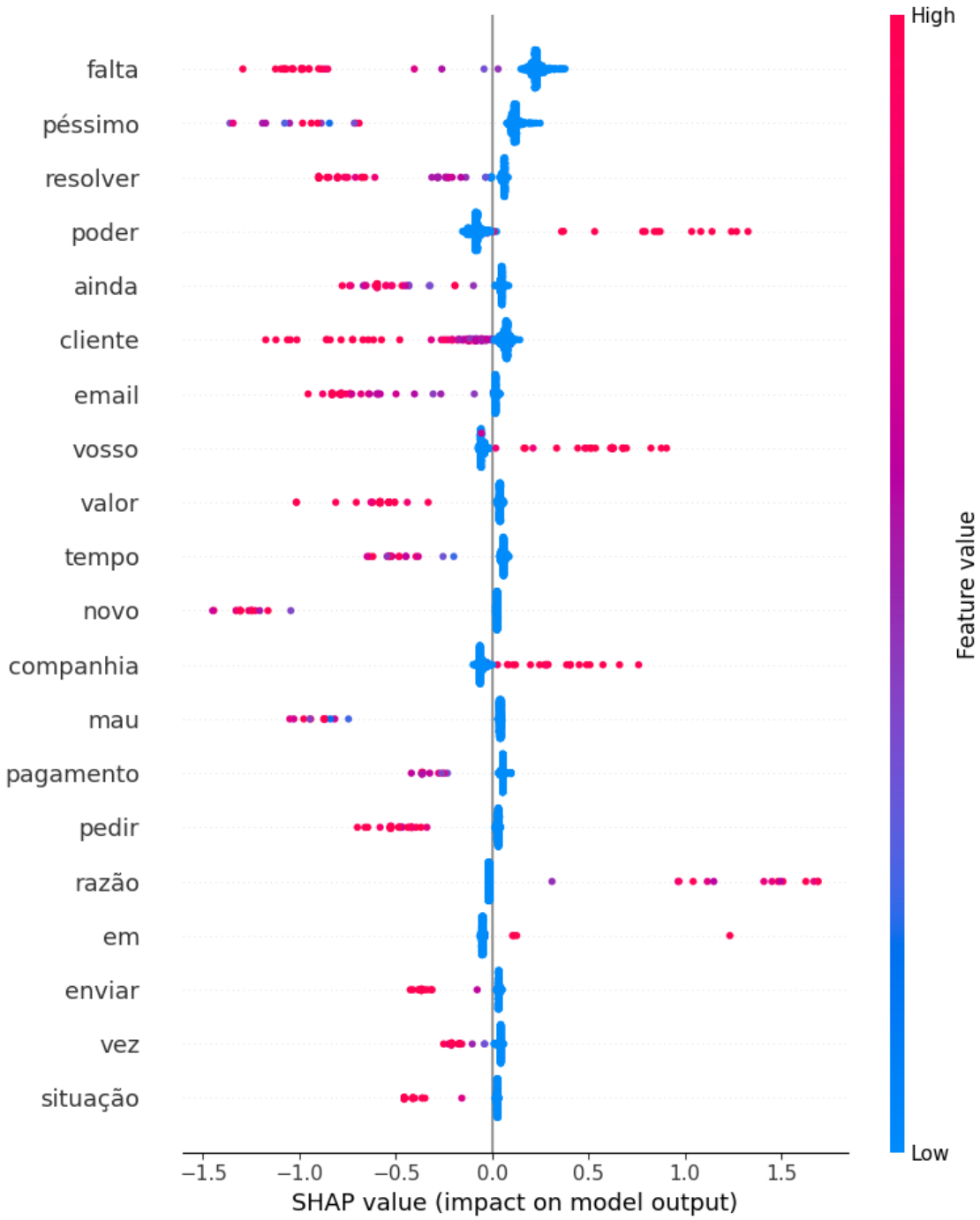
*SHAP Value Analysis of Feature Importance for **Comunicação central** Category Predictions*



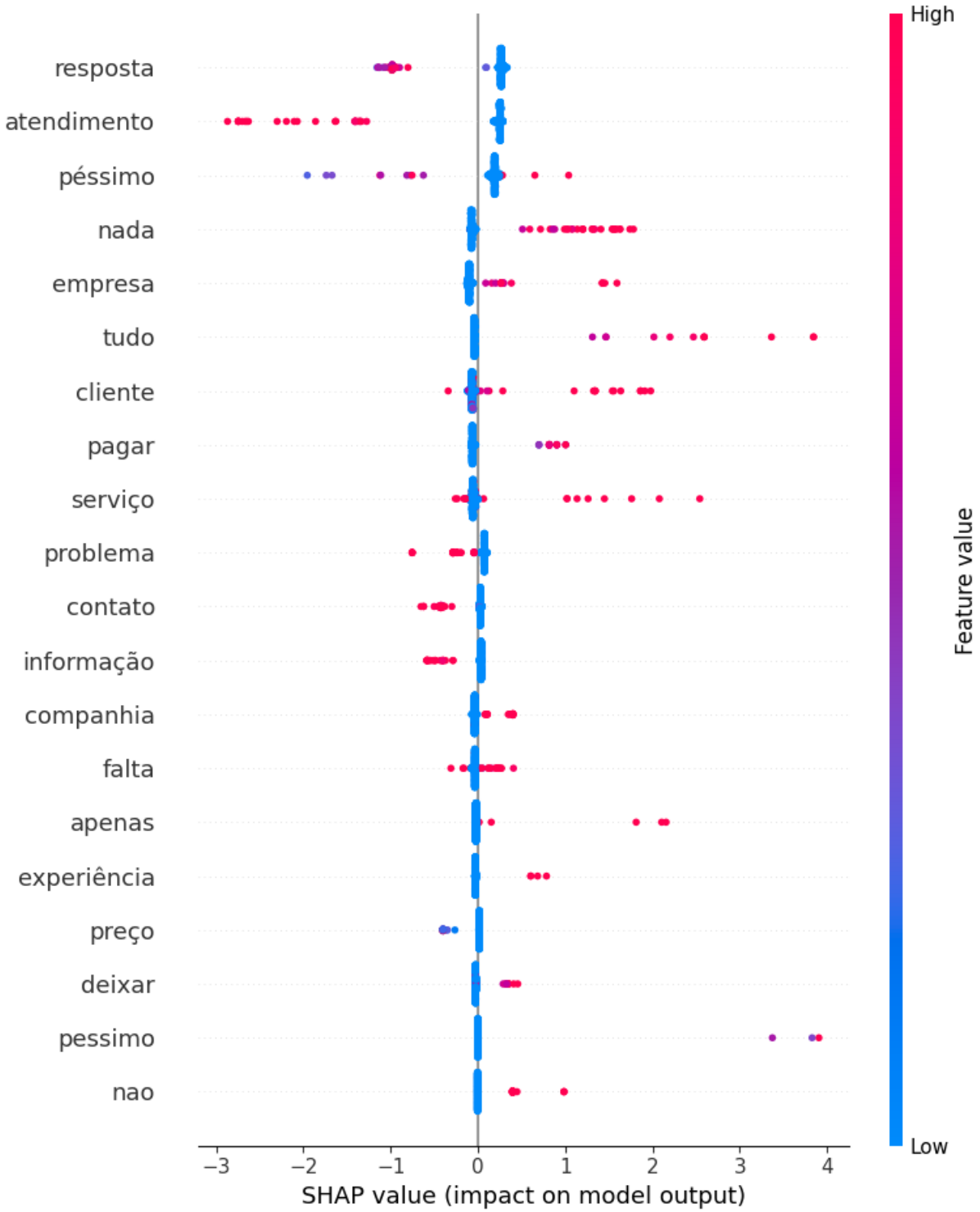
SHAP Value Analysis of Feature Importance for ***Documentação, burocracia e processo*** *Category Predictions*



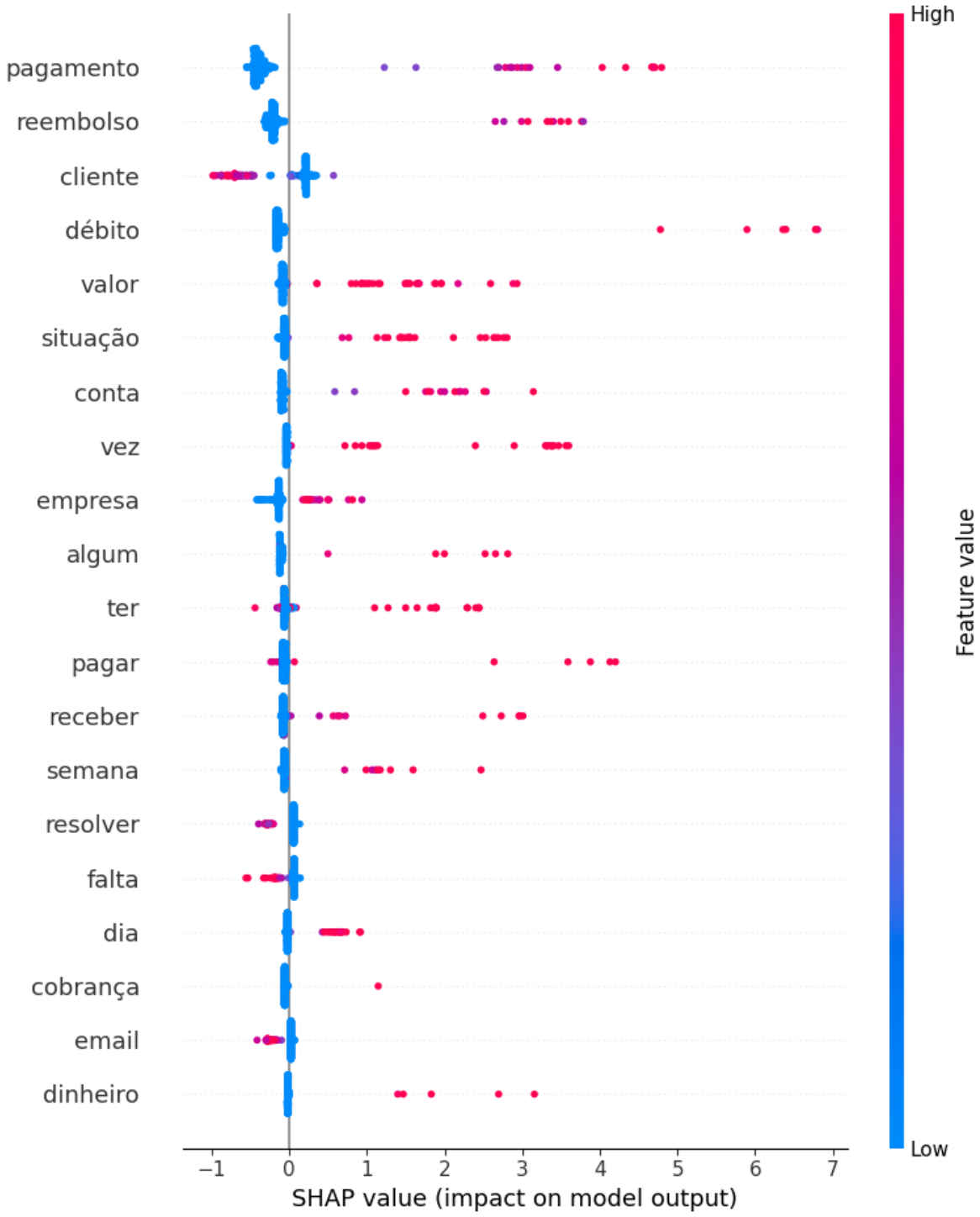
*SHAP Value Analysis of Feature Importance for **Ferramentas Digitais** Category Predictions*



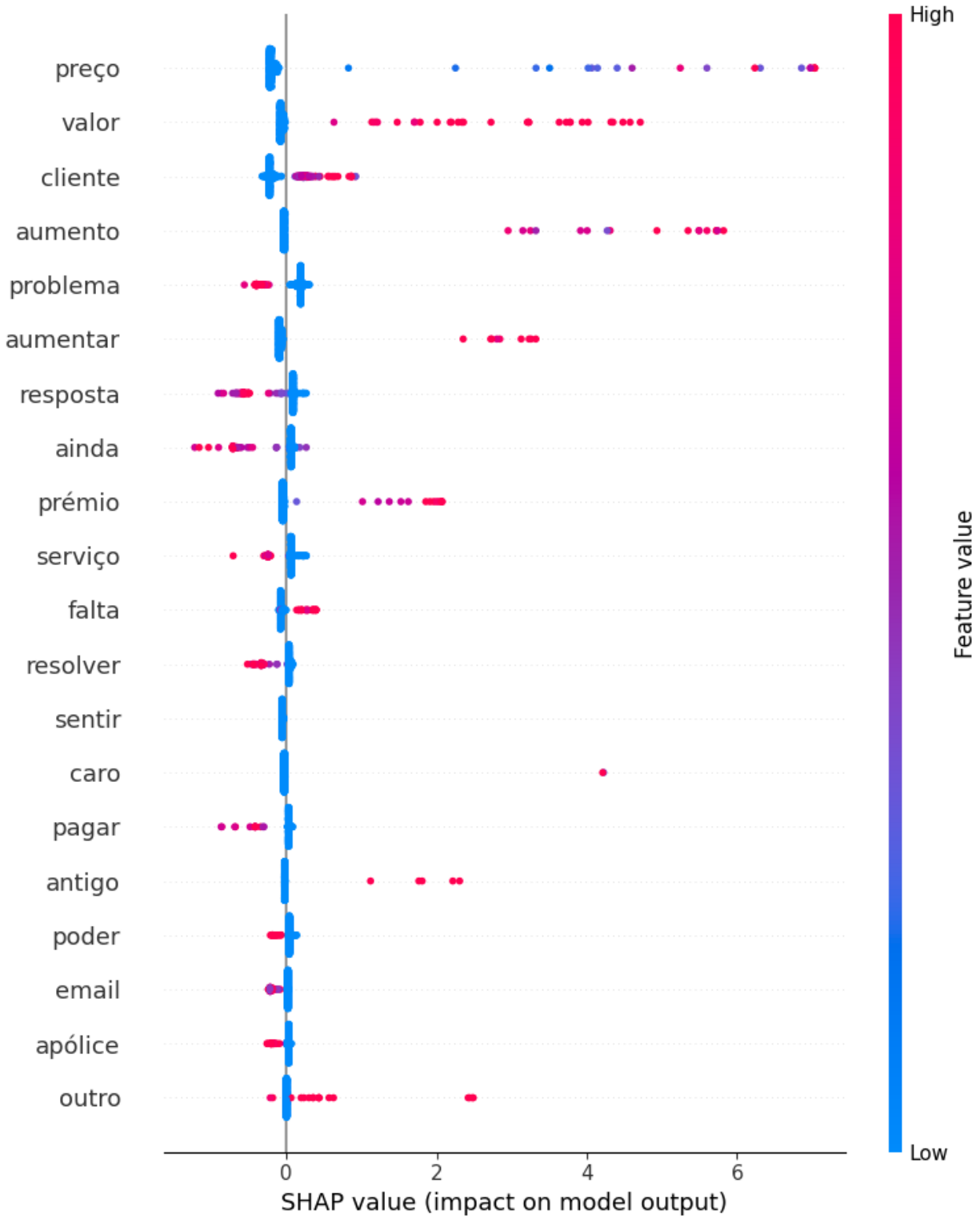
*SHAP Value Analysis of Feature Importance for **Não Especificado - Neutro** Category Predictions*



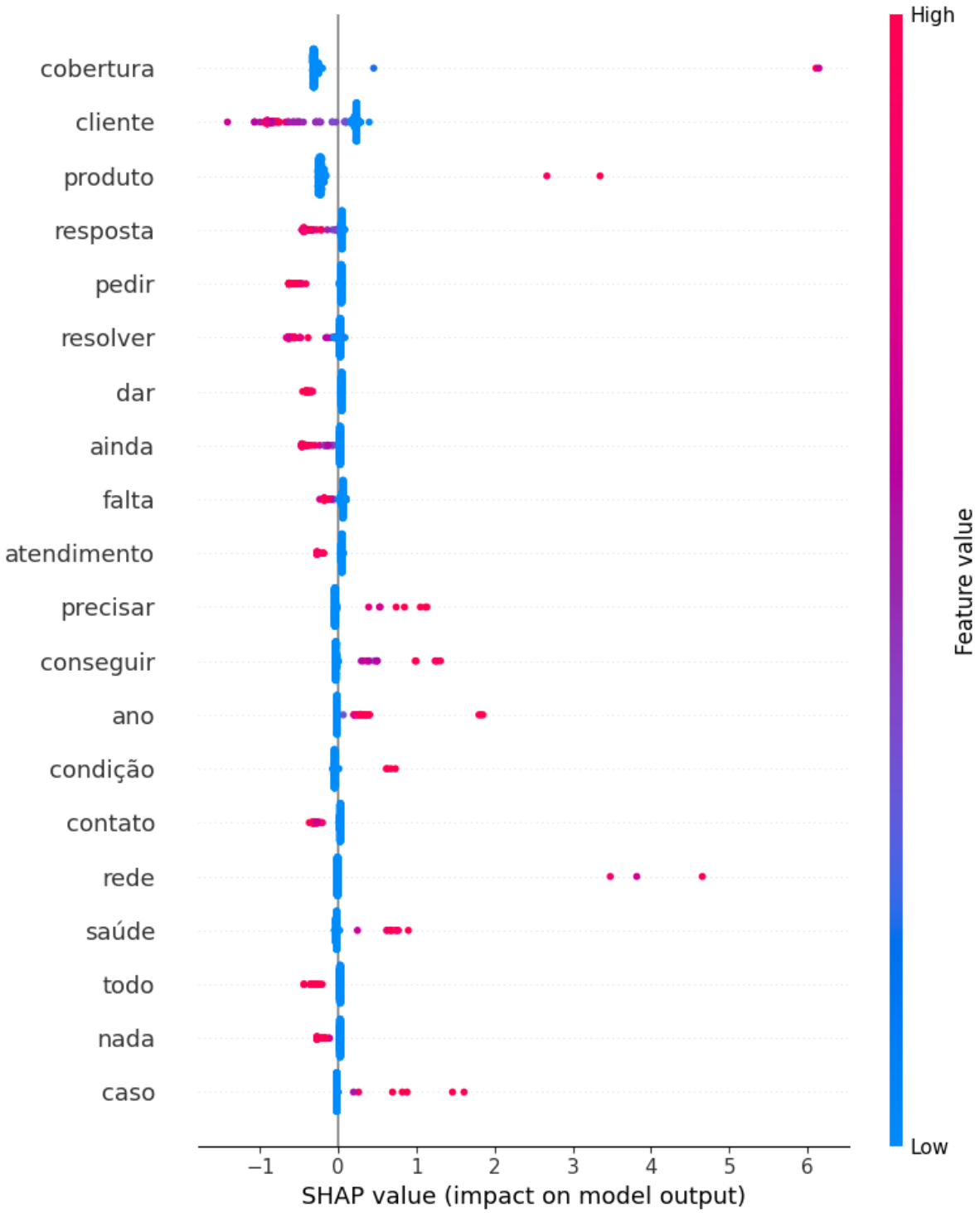
*SHAP Value Analysis of Feature Importance for **Não Especificado** - **Reclamações** Category Predictions*



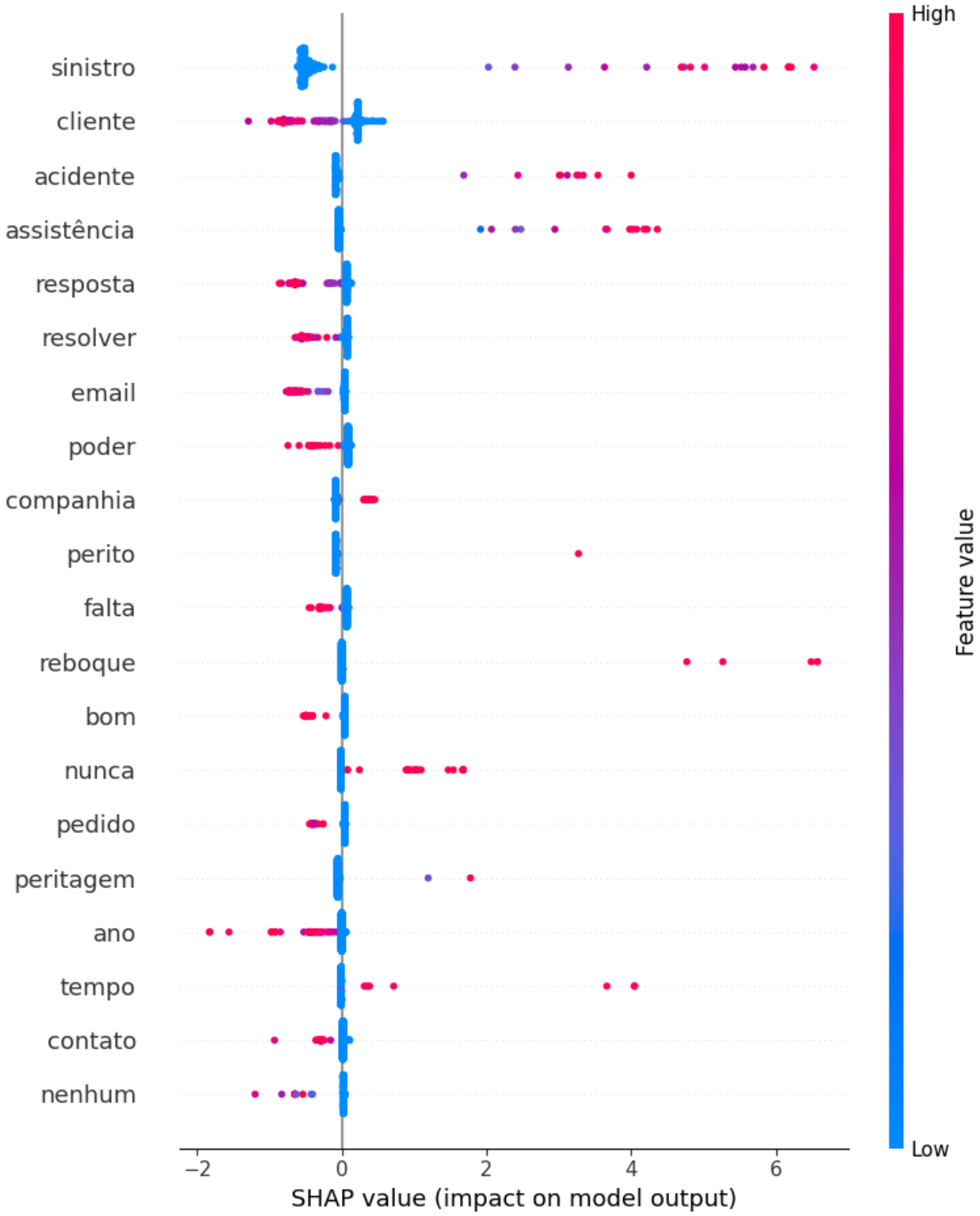
*SHAP Value Analysis of Feature Importance for **Pagamentos** Category Predictions*



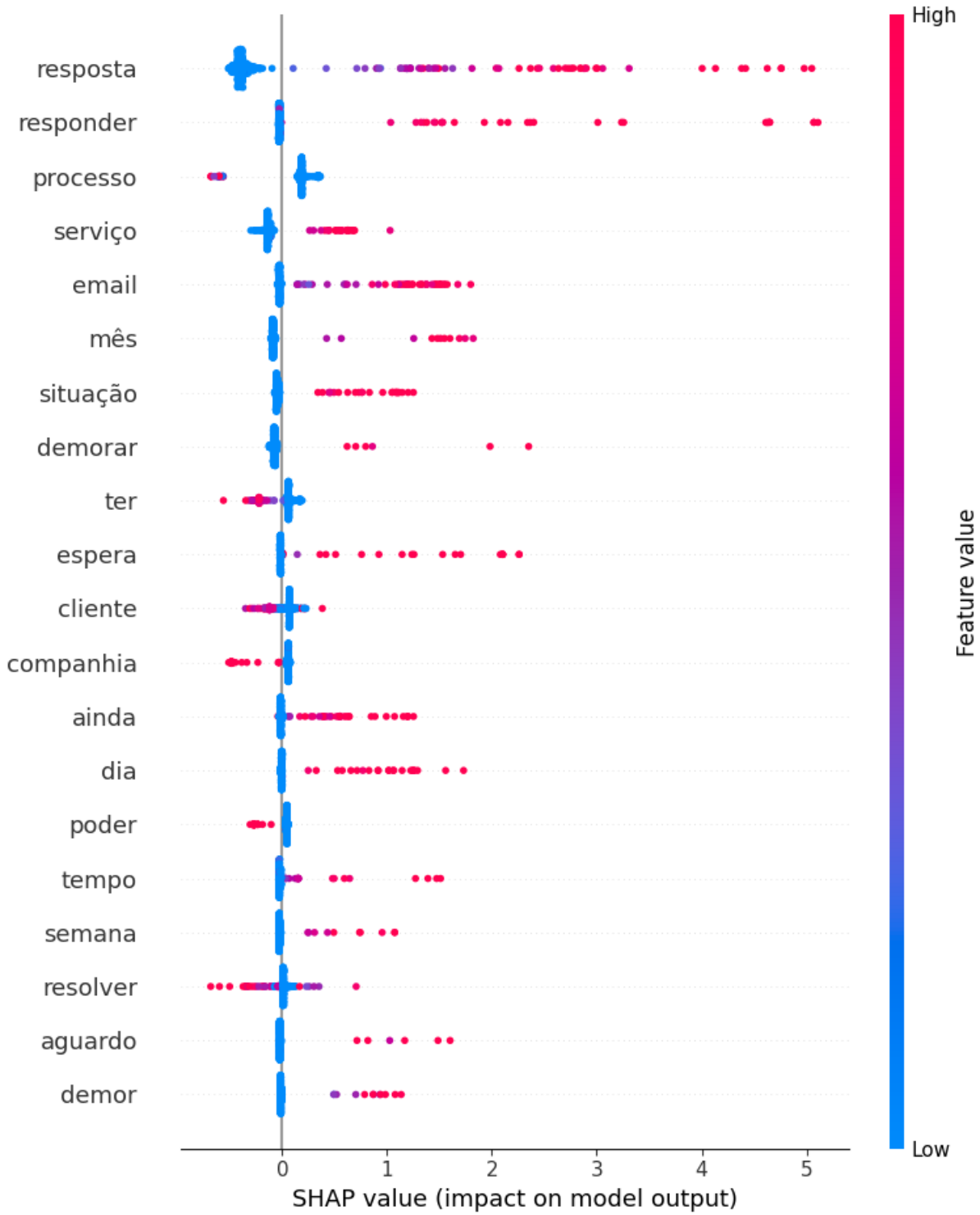
*SHAP Value Analysis of Feature Importance for **Preço** Category Predictions*



*SHAP Value Analysis of Feature Importance for **Produto** Category Predictions*



*SHAP Value Analysis of Feature Importance for **Sinistros** Category Predictions*



SHAP Value Analysis of Feature Importance for Tempo de resposta e resolução Category Predictions