



# Automatic Complaints Classification in E-Commerce: A Case Study Using CRISP-DM


Matheus Konrad Xavier   [ State University of Paraná | [matheuskonradxavier@hotmail.com](mailto:matheuskonradxavier@hotmail.com) ]

Gislaine Camila Lapasini Leal  [ State University of Maringá | [gclleal@uem.br](mailto:gclleal@uem.br) ]

Guilherme Corredato Guerino  [ State University of Paraná | [guilherme.guerino@ies.unespar.edu.br](mailto:guilherme.guerino@ies.unespar.edu.br) ]

Thiago Adriano Coleti  [ State University of Northern Paraná | [thiago.coleti@uenp.edu.br](mailto:thiago.coleti@uenp.edu.br) ]

Renato Balancieri  [ State University of Maringá | [rbalancieri@uem.br](mailto:rbalancieri@uem.br) ]

 Computer Science Collegiate, State University of Paraná, Av. Minas Gerais, 5021 Apucarana, PR, 86813-250, Brazil.

**Received:** 02 July 2024 • **Accepted:** 14 February 2025 • **Published:** 08 April 2025

**Abstract:** The growth of e-commerce has been remarkable in recent years, driven by increasing consumer demand for attention and quick responses. Given the large volume of transactions and complaints accompanying this increase, automating the classification of these complaints can help quickly route them to the appropriate departments. This paper presents a computational approach using the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology to automate the complaints screening process. We categorized 600 real complaints from three e-commerce platforms in Brazil. The learning model was trained progressively, using an initial set of 25 complaints in each category. The classification model obtained an accuracy of 85% and an average of over 80% across all relevant metrics, including precision, recall, and F1-Score. The results confirmed the effectiveness of the developed model for automated complaint classification in e-commerce, providing a computational strategy that improves the customer service process and allows for quicker problem resolution.

**Keywords:** Automatic Classification, E-commerce, CRISP-DM, Machine Learning

## 1 Introduction

E-commerce has experienced a remarkable expansion in recent years. However, this trend accelerated significantly with the outbreak of the coronavirus pandemic in 2020. According to Silva [2022], how financial and economic transactions were conducted was profoundly transformed by pandemic response measures, including social isolation and quarantine. These interventions reconfigured shopping habits, further boosting the growth of e-commerce.

According to the Brazilian Association of Electronic Commerce [Associação Brasileira de Comércio Eletrônico (AB-Comm), 2023], e-commerce revenue reached an impressive mark of 185.7 billion reais in 2023, representing an increase of 9.5% compared to the previous year. This robust growth has fostered an increasingly competitive environment in the sector, as highlighted by Silva [2022].

As competition in the e-commerce market intensifies, consumers have begun to demand to be heard and receive high-quality services [Madanchian, 2024]. According to Santouridis and Veraki [2017], there is a direct connection between the efficiency of customer service and the increase in consumer satisfaction levels. This satisfaction, in turn, is intrinsically linked to customer loyalty. Loyal customers tend to make frequent purchases, spend more, and be less sensitive to prices, as pointed out by de Leaniz and del Bosque Rodríguez [2016].

However, given the vast scale reached by e-commerce today, maintaining high-quality customer service can take time and effort, leading to occasional problems. In such circumstances, to preserve customer loyalty, it becomes essential that such issues are promptly directed to the competent de-

partment and resolved as quickly as possible.

In this context, our work addresses the automatic categorization of complaints on e-commerce platforms, a solution to tackle the challenge of effectively managing and resolving an increasing volume of complaints. The research employs advanced techniques in Machine Learning (ML), Natural Language Processing (NLP), and Artificial Neural Networks (ANN), aiming to categorize complaints automatically. This approach not only optimizes the complaint-handling process but also improves customer satisfaction and, consequently, the perception of quality regarding the product/service.

Our study explores categories encompassing a wide range of issues, including product quality, payment complications, and delivery delays. Through the automatic categorization of these complaints, companies can accelerate the problem-solving process, allowing for more efficient allocation of resources to enhance operational processes. This strategy improves effectiveness in resolving issues and contributes to the continuous optimization of the customer experience. In this context, this study aims to answer the research question: "How can complaint classification in e-commerce be automated to improve customer experience?"

Therefore, the central objective of our study is to develop a computational approach capable of automatically classifying customer complaints on e-commerce platforms into predefined categories. For this purpose, the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology will be employed. This method provides a structured framework that facilitates the analysis and understanding of data [Wirth and Hipp, 2000], allowing for more efficient and accurate categorization of complaints. CRISP-DM is structured in phases, each with clearly delineated tasks

and outcomes, offering flexibility in executing these phases [Bokrantza *et al.*, 2024].

We have organized the paper as follows: Section 2 provides the theoretical foundation, addressing concepts about Machine Learning, Natural Language Processing, and Artificial Neural Networks. Section 3 details the steps developed for conducting this research, including the methodology definition, business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The results obtained and the analyses of the automatic complaint categorization model are presented and discussed in Section 4, emphasizing the performance evaluation of the model. Finally, Section 6 presents the conclusions of this study and suggests directions for future research.

## 2 Theoretical Background

This section theoretically grounds the study, exploring fields of Artificial Intelligence: Machine Learning (ML), where algorithms learn and make predictions from data; Natural Language Processing (NLP), which allows computers to understand, interpret, and manipulate human language; and Artificial Neural Networks (ANN), systems inspired by the functioning of the human brain that simulate the ability to learn. These interconnected domains constitute the theoretical foundation of this investigation, playing a vital role in driving the advancement of innovative approaches. This section will also include a Related Work section to review prior studies and methodologies relevant to this research.

### 2.1 Machine Learning

Machine learning (ML), a subfield of Artificial Intelligence, is dedicated to developing computational techniques that enable systems to acquire knowledge automatically [Monard and Baranauskas, 2003]. The central idea is that, through the data provided and its respective categorizations, the machine or system can "learn" and create a memory base, facilitating the identification of patterns that lead to specific outcomes, as explained by Mitchell [1997]. In practice, the "experience" for these systems or machines is represented by data. By feeding the learning algorithm with this data, a model is created that can make predictions or recognize new situations, highlights Zhou [2021].

ML techniques have evolved significantly with the explosion in data generation, becoming more sophisticated. This evolution has enabled the adoption of various algorithms to solve problems, including supervised and unsupervised learning algorithms, as described by Osisanwo *et al.* [2017]. An indication of the advancement in these techniques is their wide application in practical, real-world situations. For example, in computer vision, systems based on ML range from facial recognition to the automatic classification of microscopic cells. Other notable applications include speech recognition and robotic control, among other areas, demonstrating the versatility and impact of this discipline, as highlighted by Mitchell [2006].

Within the ML field, four main learning categories differ in how information is presented to the system for train-

ing. Supervised learning involves training the model with a dataset that contains both the inputs and the desired outputs, allowing the model to make accurate predictions or classifications with new data [Sanches, 2003]. On the other hand, unsupervised learning works with data that do not have predefined labels, seeking to identify intrinsic patterns or groupings in the data [Rodríguez *et al.*, 2019]. Semi-supervised learning combines aspects of the first two, using labeled and unlabeled data, which is particularly useful when obtaining labels for all data is impractical [van Engelen and Hoos, 2020]. Finally, reinforcement learning focuses on teaching the model how to act in a given environment to maximize cumulative reward through trial and error and direct feedback on actions [Yanai *et al.*, 2020]. Each type of learning is suitable for different kinds of problems and datasets, making the field of ML extraordinarily versatile and powerful.

### 2.2 Natural Language Processing

Natural Language Processing (NLP) is a branch of Artificial Intelligence dedicated to understanding, analyzing, and producing natural language in a way that enables interactions with machines similar to human interactions [Jackson and Moulinier, 2007] [Olujimi and Ade-Ibijola, 2023]. This field aims to facilitate natural and intuitive communication between humans and machines, eliminating the need for programming languages or specific commands to interact with technology [Just, 2024].

Linguistic ambiguity represents one of the greatest obstacles to the advancement of NLP [Jackson and Moulinier, 2007]. This challenge stems from the multifaceted nature of human language, in which a single expression can carry multiple meanings. In communicative contexts, humans have the ability to unravel these multiple meanings through contextual clues, personal experiences, and cultural knowledge, allowing for a fluid and accurate understanding of language. However, machines face significant difficulties in replicating this capacity for interpretation, as they lack the ability to effectively integrate and apply these contextual and cultural elements in language analysis [Pinto, 2015].

The most commonly used techniques in supporting NLP include:

- **Lemmatization and Stemming:** Lemmatization refers to the process of reducing a word to its base form or lemma, facilitating the identification of related words. For example, words like "gostei" and "gostar" share the same root but differ in tense and form; "boas" and "bom" are variations of the same adjective, differing by gender and number; and "mesas" and "mesa" refer to the same noun but differ in singular and plural forms. On the other hand, stemming seeks to extract the stem of a word, simplifying it to its essence. Examples include reducing "gostar" to the root "gost," simplifying "esperto" and "espertinho" to "espert," and shortening "encantar" to "encant." This process involves identifying and extracting the common root or stem of related words, often by removing suffixes or endings, to standardize variations for analysis [Pinto, 2015].
- **Token Generation:** Essential in text data analysis and

NLP, tokenization divides the text into smaller units, known as tokens. This process is crucial for preparing the data for subsequent analyses, facilitating the understanding of the text's structure and content [Rodríguez and Bezerra, 2020].

- **Part-of-Speech Identification:** Identifying the grammatical class to which a word belongs is crucial for unraveling its function in the sentence and, by extension, clarifying its semantic meaning. This technique significantly reduces linguistic ambiguity, allowing for a more accurate interpretation of the text [Pinto, 2015].

In NLP, ML algorithms play a crucial role, especially those designed to handle inputs and outputs of a fixed and well-defined length. The heterogeneous and unstructured nature of textual data in its raw form presents significant challenges, given its lack of uniform standardization. Therefore, converting texts into a numerical and vector representation is essential. This conversion process facilitates the handling and analysis of data by ML algorithms, allowing them to extract meaningful insights and perform complex NLP tasks more efficiently and accurately.

## 2.3 Artificial Neural Networks

An Artificial Neural Network (ANN) is a computational structure designed to simulate how the human brain processes information when performing specific tasks [Haykin, 2001]. The configuration and type of network used are often determined by the specifics of the problem to be solved, including the constraints that define the applicable learning algorithms [Rauber, 2005]. ANNs are particularly useful for addressing complex issues where the behavior of multiple variables cannot be completely predetermined [Fleck *et al.*, 2016]. Through training with exemplary datasets, these networks can generalize the knowledge acquired and apply it to previously unknown datasets, demonstrating their robustness and adaptability in various application scenarios [Gorgens *et al.*, 2009].

## 2.4 Related Works

This paper reviewed and analyzed previous research, contextualizing how it addressed similar classification problems in e-commerce environments and highlighting the significant contributions of these studies in the field.

In the study by Gonçalves [2016], the feasibility of using Sentiment Analysis on the Reclame Aqui<sup>1</sup> website is explored. Employing NLP and ML techniques, the author identified patterns in the satisfaction and dissatisfaction ratings expressed in consumer complaints. This work demonstrates how these technologies can be applied to understand consumer perceptions better and improve feedback management.

The study conducted by Rabbi *et al.* [2018] established a data standard for implementation in the databases of the

Brazilian PROCON<sup>2</sup>. For this purpose, the Knowledge Discovery in Database (KDD) process was used as the methodology to facilitate data mining. This mapping identified valuable information, including the average duration of the most problematic complaints. This discovery was essential for better understanding the dynamics and challenges consumers and regulators face in the context of Brazilian commerce.

Peixoto's study (Peixoto [2021]) explores the use of Text Mining and Machine Learning techniques to enhance the complaint screening process in financial services, aiming to reduce the likelihood of unresolved complaints. The proposed methodology was evaluated using three architectures of increasing complexity: the base model employs a Naive-Bayes SVM; the intermediate model uses FastText embeddings with a Multi-layer Perceptron classifier; and the advanced model adopts DistilBERT, a natural language processing technique. The results of Peixoto [2021] indicate that these approaches can add significant value to operational decision support systems, improving customer service, increasing consumer satisfaction, and mitigating risks to the company's reputation.

In the study of Itsari and Budi [2022] presents a solution to automate the categorization of complaints at Bukalapak. The company has experienced significant growth, increased complaints, system instability, and difficulty in reclamation categorization, making it difficult to find solutions. This has led to a reduction in user satisfaction. Were used in this research Logistic Regression, k Nearest Neighbor, and Support Vector Machine. Being that the Logistic Regression presented the best performance.

In Ayanoğlu *et al.* [2023] a system for detecting and classifying customer reviews containing complaints is presented, developed to determine which product or service characteristic the complaint refers to. The study was conducted in two stages, the first to identify whether the comment was positive or negative. In the second stage, the negative comments were categorized using Word2Vec and BERT. The best performance was obtained with the Word2Vec method.

Vinayak and Jyotsna [2023] classified consumers' complaints which is in form of text, into 6 classes using deep learning models and embedding techniques. In this study the classes represent departments where complaints are routed. Deep learning models like LSTM, BiLSTM, GRU and 1D CNN are used, along with word embedding techniques like Word2Vec, Fasttext, Bert and Distilbert to represent text. The better results were obtained with DistilBert and CNN.

This Related Works section explores various complaint categorization approaches across different platforms and in multiple languages, employing techniques ranging from sentiment analysis and text mining to complex deep learning models. Notable contributions include models like SVM, Naive Bayes, convolutional neural networks, Word2Vec, and BERT, which assist in processing and categorizing complaints. Unlike these studies, our research focuses on the automated categorization of complaints on e-commerce platforms using ML, NLP, and ANN to optimize the complaint-

<sup>1</sup>Reclame Aqui is a Brazilian Website where consumers can post complaints about companies regarding customer service, purchases, sales, products, and services. Available at: <https://www.reclameaqui.com.br/>

<sup>2</sup>Procon is a public agency present in all Brazilian states as well as in various cities, dedicated to consumer protection, performs several essential functions, among which extrajudicial mediation of conflicts between consumers, companies, and service providers stands out.

handling process and improve customer satisfaction. Our methodology is structured around CRISP-DM, which provides rigor in analysis and facilitates more efficient and accurate categorization.

### 3 Research Method

The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology was adopted for this study. Developed in 1996 by a consortium of companies, this model is recognized for its structured and systematic approach, which guides the data mining process from the initial understanding of the business to the effective implementation of the results [Chapman *et al.*, 2000].

The CRISP-DM methodology is structured into six main phases, as illustrated in Figure 1: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Each phase has specific tasks that help transform large volumes of raw data into useful information and actionable insights [Shearer, 2000]. The focus on understanding both the business context and the data ensures that the results of data mining projects are relevant and aligned with the organization's strategic needs.

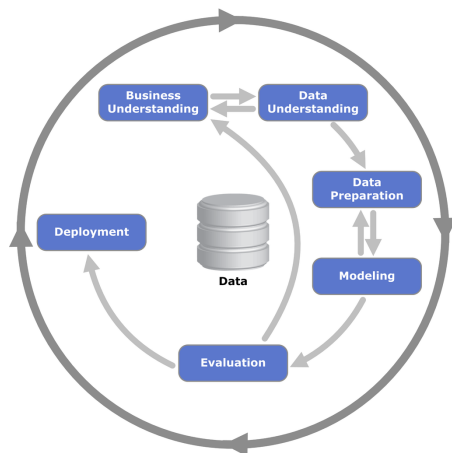


Figure 1. CRISP-DM Methodology (Santos *et al.*, 2019)

Next, each of the phases of the CRISP-DM Methodology is detailed.

#### 3.1 Business Understanding

This initial phase is dedicated to deeply understanding the project's objectives and requirements from a business perspective. It was observed that complaints could be categorized in various ways, including complaints about high prices, quality of service, or dissatisfaction with a purchase. The study focused on classifying the most frequent complaint categories on the Reclame Aqui Platform, explicitly addressing the following issues:

- **Product:** situations where the customer receives an incorrect or defective product;
- **Payment:** includes cases of improper charges or failures in receiving payment through the website;

- **Delivery:** refers to issues related to the transportation of the product and meeting the established delivery deadlines.

Due to the lack of a clear and absolute standard for classifying these topics, employing a supervised learning model became essential. This model was trained based on previous complaints to identify patterns and the most commonly used words, thereby allowing for the precise and standardized classification of future complaints.

#### 3.2 Data Understanding

This stage involves collecting, exploring, and validating data quality to ensure it is suitable for further analysis. We needed a comprehensive dataset of real complaints on the selected topics to enhance the model's efficiency. The dataset for this analysis was sourced from the Reclame Aqui Platform, which maintains a standardized record of complaints concerning various companies. For this study, we extracted a total of 600 complaints from three major e-commerce companies operating in Brazil: Amazon<sup>3</sup>, Magazine Luiza<sup>4</sup>, and Submarino<sup>5</sup>. These complaints were evenly divided across three primary categories: product, payment, and delivery issues, with 200 complaints in each category. The comments within the data are in Portuguese (Brazil).

In this study, precautions were taken to ensure the privacy and confidentiality of the complaint data analyzed. The complaints extracted from the Reclame Aqui Platform did not include any personal identification of the complainants, thus preserving the privacy of individuals. Additionally, data usage was restricted to the specific purpose of automated analysis and categorization aimed at improving internal complaint handling processes. The data were aggregated and anonymized, ensuring that no sensitive or identifiable information was accessed or disclosed. This approach respects ethical guidelines related to data privacy and the integrity of the individuals involved without requiring additional user consent due to the public and anonymized nature of the information.

The authors manually validated and labeled each complaint to ensure accuracy and consistency in the classification process. Labels were assigned based on predefined topics to facilitate the categorization and subsequent training of the model. This careful labeling process was essential to maintain the dataset's quality and ensure the model could effectively learn from the categorized data. The dataset obtained is available on the open data platform Zenodo<sup>6</sup> at address <https://zenodo.org/records/14056151>.

#### 3.3 Data Preparation

In this stage, data preprocessing occurs, which aims to enhance the data quality used. Thus, it was necessary to perform an initial spell check to standardize the use of words and improve the analysis of the frequency and relevance of each

<sup>3</sup><https://www.amazon.com.br/>

<sup>4</sup><https://www.magazineluiza.com.br/>

<sup>5</sup><https://www.submarino.com.br/>

<sup>6</sup><https://zenodo.org/>

term used in the complaints. The analysis of the complaints revealed a significant number of irregularities; many of them contained abbreviations, such as "vcs", (*the Portuguese abbreviation for "you"*), instead of the complete form "vocês" (*"you"*), and were also riddled with numerous spelling errors. The standardization of the vocabulary, essential for the reliability of subsequent analyses, was achieved through spell correction using Python's Enchant library.

After the spell correction, several NLP techniques were implemented to refine the data. Initially, token generation was carried out, as described in subsection 2.2. This process aims to facilitate the application of subsequent textual analysis techniques. Figure 2 illustrates an example of token generation from a sentence.

Portuguese
<b>Input:</b> As minhas compras não chegaram no prazo estipulado no site.
<b>Output:</b> 'As', 'minhas', 'compras', 'não', 'chegaram', 'no', 'prazo', 'estipulado', 'no', 'site'.
English
<b>Input:</b> My purchases did not arrive within the deadline specified on the website.
<b>Output:</b> 'My', 'purchases', 'did', 'not', 'arrive', 'with', 'the', 'deadline', 'specified', 'on', 'the', 'website'.

Figure 2. Token generation in a sentence

After segmenting the text into tokens, the elimination of so-called stopwords was carried out. According to Zuin *et al.* [2016], stopwords are words, generally of a functional nature, considered to be of little relevance, which does not significantly contribute to the semantic value of the analyzed text, such as 'my', 'did', 'not'.

Figure 3 displays an example of removing stopwords in a sentence.

Portuguese
<b>Input:</b> As minhas compras não chegaram no prazo estipulado no site.
<b>Output:</b> 'compras', 'chegaram', 'prazo', 'estipulado', 'site'.
English
<b>Input:</b> My purchases did not arrive within the deadline specified on the website.
<b>Output:</b> 'purchases', 'arrive', 'deadline', 'specified', 'website'.

Figure 3. Removal of stopwords in a sentence

After removing stopwords, the subsequent techniques applied were lemmatization and stemming. Both techniques are detailed in subsection 2.2 and were employed to standardize the text further and unify each word's meaning about the complaint. Figure 4 demonstrates the results of each stage.

During the preparation of the text to feed the machine

Portuguese
<b>Input:</b> As minhas compras não chegaram no prazo estipulado no site.
<b>Lemmatization Output:</b> 'compra', 'chegar', 'prazo', 'estipular', 'site'.
<b>Stemming Output:</b> 'compr', 'cheg', 'praz', 'estipul', 'sit'.
English
<b>Input:</b> My purchases did not arrive within the deadline specified on the website.
<b>Lemmatization Output:</b> 'purchase', 'arrive', 'deadline', 'specify', 'website'.
<b>Stemming Output:</b> 'purchas', 'arriv', 'deadlin', 'specif', 'websit'.

Figure 4. Lemmatization and Stemming in a sentence

learning model, the TF-IDF (Term Frequency - Inverse Document Frequency) algorithm<sup>7</sup> was employed to process the complaints grouped by similar topics. This technique facilitates the model's training by highlighting the most relevant words in each thematic group.

We chose the TF-IDF algorithm because it effectively identifies the importance of terms within each complaint, balancing term frequency with the uniqueness of terms across the entire dataset. This approach was preferred over other techniques, such as word embeddings or one-hot encoding because it offers a straightforward, interpretable method to capture the distinctiveness of words within specific complaint topics without requiring extensive computational resources. TF-IDF also avoids issues of dimensionality that can arise with one-hot encoding and does not require large datasets for practical training, unlike some embedding-based approaches, making it well-suited to our dataset's size and scope.

### 3.4 Modeling

This phase involves selecting and applying various modeling techniques and optimizing their parameters. Therefore, the choice was made to use the Multilayer Perceptron ANN, a practical algorithm for solving nonlinear problems and multiclass classification [Falcão *et al.*, 2013]. The ANN was chosen due to its effectiveness in handling nonlinear problems and multiclass classification tasks, inherent in the complaint dataset analyzed. This neural network is particularly suitable for complex relationships among variables, allowing for deep learning of underlying patterns. Additionally, the ANN provides flexibility for parameter tuning and performance optimization, which was essential in the incremental training and evaluation process conducted in the study.

The model's training began with an initial set of 300 complaints evenly distributed among the three selected topics, all manually classified by the authors. The training was incrementally increased to evaluate the model's effectiveness

<sup>7</sup>The resulting value from the TF-IDF algorithm is a statistical measure that aims to represent the importance of a word within a document, in comparison to a complete collection of documents (Filho *et al.*, 2023).



and determine the necessary number of complaints for optimal performance. With each test round, 25 new complaints per topic were added. The complaints were processed using the resulting value from the TF-IDF algorithm, allowing the model to identify and learn the most relevant words from each topic to classify future complaints effectively.

The programming language selected for developing the model was Python 3.10, chosen for its vast library availability that supports machine learning development. The most used libraries were:

- **Matplotlib:** used for data visualization through graphs;
- **NTLK:** employed for natural language processing;
- **Scikit-learn:** used for building, training, evaluating, and obtaining performance metrics of the model.

### 3.5 Evaluation

In this phase, the evaluation is carried out to check whether the objectives set in the initial phase have been achieved. For this purpose, tests were conducted using a set of 300 previously labeled complaints extracted from the Reclame Aqui Platform and manually classified by the authors. The classification generated by the model was then compared with the expected results.

A confusion matrix was implemented to extract statistical metrics from the obtained results. The confusion matrix, as described by Grandini *et al.* [2020], is an essential tool for comparing the obtained results with the expected ones, facilitating the extraction of performance metrics. The confusion matrix allows for visualizing other derived metrics, such as accuracy, precision, recall, and the F1-Score. These metrics are crucial for comprehensively evaluating the model's performance.

The choice of Accuracy, Precision, Recall, and F1-Score metrics to evaluate the model's performance was based on the relevance of these metrics to the context of multi-category complaint classification. Precision and Recall were essential to ensure the model correctly identified each category, minimizing false positives and negatives—critical aspects in handling customer complaints. The F1-Score was chosen to balance these factors and provide a robust overall view of performance in situations with potential class imbalances. Although metrics such as AUC and Matthews Correlation Coefficient (MCC) also offer valuable insights, we opted for metrics focusing on direct counts of true and false positives and negatives, which are more interpretable for the applied context.

### 3.6 Deployment

In this phase, the developed model and the knowledge gained throughout the process were organized so that the categorization of complaints enables faster and more efficient distribution. This system facilitates directing complaints to the responsible departments, optimizing response time and resolving identified issues. The results of this approach will be presented and discussed in the next section.

## 4 Results and Discussions

In this experiment, we used a database of 600 authentic complaints extracted from the Reclame Aqui platform, evenly distributed among the identified topics (product, payment, and delivery) during the context understanding phase. Each topic included 200 complaints, which the study's authors manually classified. The total of 600 complaints was determined by the platform's limitation, which did not allow for extracting a larger dataset. For the training and evaluation of the classification model, these 600 complaints were organized into two sets of 300, each containing 100 complaints from each analyzed topic.

To evaluate the progression of the model, the complaints designated for training were presented gradually, in batches of 25 complaints from each topic. After each batch, the model was tested by classifying the 300 complaints from the test set, thus allowing the extraction of metrics to assess the evolution of the learning. The metrics extracted from the confusion matrix to evaluate the model's effectiveness include Accuracy, Precision, Recall, and F1-Score.

### 4.1 Model Training

The model was fed gradually, with the results being analyzed after each insertion of a group containing 25 complaints per topic. Considering that the database has 100 complaints for each topic and each group includes 25 complaints, the study was divided into four testing phases. Each phase has its specific results, and the relevant metrics were extracted for evaluation.

### 4.2 First Round

Figure 5 presents the confusion matrix resulting from the first test round. Concurrently, Table 1 shows the metrics obtained in this round. The model was fed with only 25 complaints per topic during this initial stage.

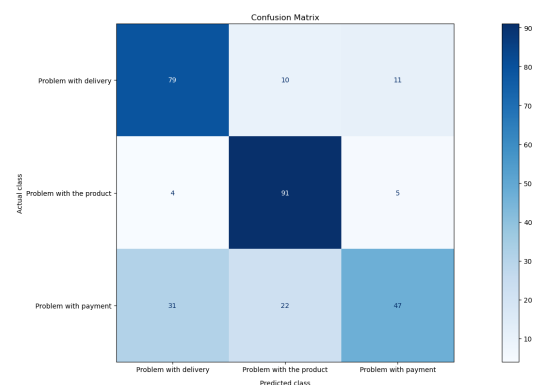


Figure 5. Confusion matrix from the first round of tests

	Precision	Recall	F1-Score
Problem with delivery	69%	79%	74%
Problem with product	74%	91%	82%
Problem with payment	75%	47%	58%
Accuracy	72%		

Table 1. Metrics extracted from the first round of tests

It can be observed in Table 1 that the model exhibited reasonable performance in classifying complaints related to product issues, achieving a recall of 91%. However, the performance was less satisfactory for complaints about payment issues, which recorded a recall of only 47%. Often, the model erroneously classified these as belonging to other problem categories.

### 4.3 Second Round

The model was fed with 50 complaints from each topic in the second round of tests. The results of this stage can be observed in the confusion matrix, presented in Figure 6, and the detailed metrics in Table 2.

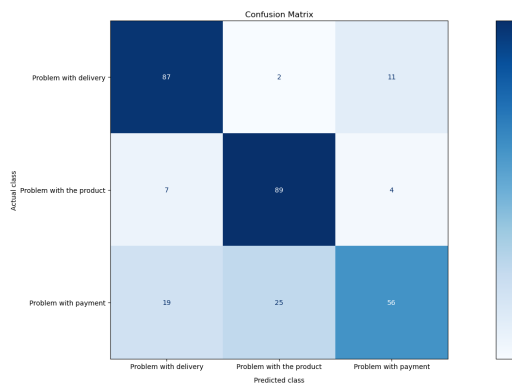


Figure 6. Confusion matrix from the second round of tests

	Precision	Recall	F1-Score
Problem with delivery	77% (+8%)	87% (+8%)	82% (+8%)
Problem with product	77% (+3%)	89% (-2%)	82%
Problem with payment	79% (+4%)	56% (+9%)	65% (+7%)
Accuracy	77% (+5%)		

Table 2. Metrics extracted from the second round of tests

The analysis of the results presented in Figure 6 and Table 2 reveals an improvement in the model's performance. Specifically, there was a 9% increase in recall for payment-related problems and an 8% increase in the F1-Score for delivery issues. However, there was a slight reduction in recall for product-related problems. The model's overall accuracy also improved, moving from 72% to 77%, representing a gain of 5%.

In this round of testing, the model displayed a significant improvement, particularly evident in its enhanced ability to classify issues related to payment, compared to the previous round.

### 4.4 Third Round

In the third round of testing, after being fed with 75 complaints, the model's performance is continuously evolving, as demonstrated by Figure 7 and Table 3. A particular improvement is noted in the ability to classify payment-related problems, indicating that the increase in data volume positively contributes to the model's effectiveness.

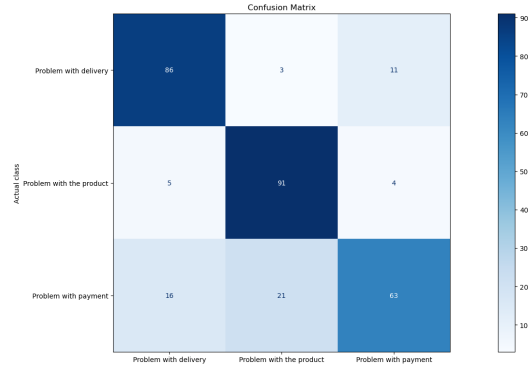


Figure 7. Confusion matrix from the third round of tests

	Precision	Recall	F1-Score
Problem with delivery	80% (+3%)	86% (-1%)	83% (+1%)
Problem with product	79% (+2%)	91% (+2%)	82%
Problem with payment	81% (+2%)	63% (+7%)	71% (+6%)
Accuracy	80% (+3%)		

Table 3. Metrics extracted from the third round of tests

### 4.5 Fourth Round

In the fourth and final round of testing, the model was fed with 100 complaints from each topic. The results obtained can be viewed in the confusion matrix, presented in Figure 8, and the corresponding metrics are detailed in Table 4.

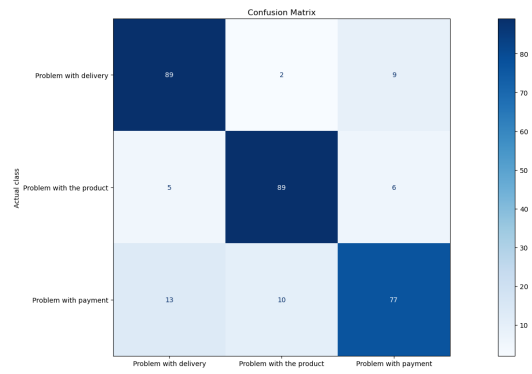


Figure 8. Confusion matrix from the fourth round of tests

	Precision	Recall	F1-Score
Problem with delivery	83% (+3%)	89% (+3%)	86% (+3%)
Problem with product	88% (+9%)	89% (-2%)	89% (+7%)
Problem with payment	84% (+3%)	77% (+14%)	80% (+9%)
Accuracy	85% (+5%)		

Table 4. Metrics extracted from the fourth round of tests

In this round, the overall metrics of the model showed significant improvement, particularly in classifying payment-related problems. The recall for this category increased markedly by 14%, while the F1-Score improved by 9%, reaching 80%, a highly satisfactory result. With the improvements in this final phase, the model achieved an accuracy of 85%, and all topics registered an F1-Score of 80% or more. These results indicate that the model has significantly evolved throughout the testing phases.

## 4.6 Discussion of the Results

In the first round, the results were moderate for issues associated with delivery and product. However, performance was significantly lower for payment-related problems, with the model achieving only 47% recall and 58% F1-Score. Payment issues often needed clarification, erroneously classified 31 times as delivery problems and 22 times as product issues.

The low values observed in this first round were expected and can largely be attributed to using the TF-IDF algorithm for the textual representation of the complaints. This algorithm calculates the importance of a term within a document relative to a set of documents based on the frequency of the term compared to the total number of words in the corpus. Initially, this methodology can induce confusion in the model, especially when the number of samples is limited, as in the first round, making it difficult to define the most relevant words precisely. This issue is particularly critical in the category of payment problems, where a greater diversity of words is needed to describe the issues adequately. Thus, with limited samples, the model needs help identifying the most relevant keywords for this category.

The TF-IDF algorithm, while helpful in calculating term relevance based on word frequency, has limitations, especially in scenarios with small sample sizes, as seen in the initial rounds of our study. TF-IDF's reliance on frequency can lead to challenges in identifying nuanced relationships between terms, especially in categories like payment problems, which require a richer vocabulary to represent diverse issues accurately. In contrast, alternative text representation methods, such as word embeddings (e.g., Word2Vec) or transformer-based models (e.g., BERT), capture semantic relationships between words, making them more effective in recognizing contextual meaning even with limited data. These methods might enhance the model's ability to distinguish between subtle differences in complaint categories, thereby improving accuracy. This comparison highlights how embedding-based approaches could address TF-IDF's limitations by leveraging a deeper understanding of language nuances. This may yield better performance in categorization tasks with complex language requirements.

On the other hand, the other two categories were less affected because they frequently repeated keywords and, therefore, quickly became relevant to that type of document in the TF-IDF algorithm calculation. However, this also proves problematic, as it leads the model to classify more false positives when it encounters these words in other categories. This phenomenon is particularly evident in the categories of delivery and product issues. Despite showing a higher recall score in the first phase than the payment issues category, they registered a lower precision due to the high number of false positives.

A deeper analysis of the model's errors reveals that false positives are concentrated in complaints where common keywords across categories, such as "delivery" and "product," lead the model to misclassify the complaint. For instance, in some payment-related complaints, terms like "product" or "delivery" appeared contextually, causing the model to classify them under the delivery or product issue categories. This type of confusion occurs due to TF-IDF's limitation in cap-

turing semantic context, as it relies solely on term frequency. Identifying these occurrences allows us to propose future improvements, such as adopting embedding-based models or transformers, which better capture context and can reduce false positives by distinguishing subtle differences in language across categories.

From the second round onwards, a significant improvement in the overall score of the model was observed, with increases of up to 14% from one round to the next. This advancement is a direct consequence of the factors previously mentioned. As the number of samples increased, the precision in defining the most relevant words of each topic also increased, thereby improving the model's ability to differentiate each category based on these words. By the end of the fourth round, the model achieved an average of over 80% in all metrics. A highlight was the 30% improvement in recall for the category of payment-related problems. This means that, if 1000 complaints related to payment issues were presented in the first round of testing, the model would identify only 470; however, by the end of the rounds, it would be capable of recognizing 770, representing a substantial improvement. All these improvements can be detailed in Table 5.

	Precision		Recall		F1-Score	
	1st	4th	1st	4th	1st	4th
PD	69%	83% (+14%)	74%	86% (+12%)	79%	89% (+10%)
PP	74%	88% (+14%)	82%	89% (+7%)	91%	89% (-2%)
PPY	75%	84% (+11%)	58%	80% (+22%)	47%	77% (+30%)
Accuracy	1st: 72%		4th: 85% (+13%)			

**Table 5.** Comparison between the first and fourth rounds of testing  
Caption: PD: Problem with delivery; PP: Problem with product; PPY: Problem with payment

The analysis of Table 5 reveals that the developed model successfully achieved the proposed objectives. Throughout the various phases of training and evaluation, there was a significant improvement in the model's effectiveness in categorizing complaints related to delivery, product, and payment issues. This improvement demonstrates the model's enhanced ability to appropriately respond to different categories of complaints.

## 5 Threats to validity

The threats to the validity of this study were classified following the definitions established by Runeson and Höst [2009]:

- **Internal Validity:** the accuracy of the complaint classification can be influenced by various internal factors, such as the quality and representativeness of the dataset used, as well as the risk of model overfitting. To mitigate potential threats to internal validity, the model's performance was regularly evaluated using an independent test dataset that had not been used during training. This practice ensures that the model maintains good performance under real usage conditions. Additionally, we implemented a process of continuous monitoring, allowing for systematic verification of the model's performance over time. This strategy is crucial for identifying and correcting any declines or inconsistencies in the model's performance, ensuring its effectiveness and ongoing reliability.



- **External Validity:** Ensuring the model's generalizability to other e-commerce platforms and contexts is a key consideration in this study. To address this, we increased the diversity and volume of data collected, incorporating data from three major Brazilian e-commerce stores that span various product categories and complaint types. This strategy enhances representativeness and reduces potential training bias in the model. Additionally, we acknowledge that applying the model to international datasets may introduce challenges, such as linguistic nuances and cultural differences in complaint expression. Addressing these nuances in future work could improve the model's adaptability across various linguistic and cultural contexts. Finally, we consider the model's practical applicability in real-world customer service platforms, where its integration could improve response times and customer satisfaction, demonstrating its potential utility across diverse operational settings.
- **Construct Validity:** the appropriateness of the predefined complaint categories and the accuracy with which they are applied to customer complaints are essential to ensure the practical relevance and utility of the proposed model. Ambiguous definitions or poorly specified categories can result in misclassification, thus compromising the effectiveness of the model. To mitigate these risks, we implemented a process of continuous review of the classification criteria. This process is regularly adjusted based on authentic feedback and empirical observations, to ensure that the categories remain accurate and aligned with the real and dynamic needs of the users.

## 6 Conclusion

This paper describes a case study that employs the CRISP-DM methodology for the automatic classification of customer complaints into predefined categories on e-commerce platforms. The model exhibited excellent comprehension and effective categorization of complaints into specific groups, such as issues related to delivery, product, and payment. These results corroborate the effectiveness of the approach used and highlight the model's potential to enhance customer service and optimize business operations in the digital environment.

Initially, the model showed moderate results, particularly in issues related to payment where the recall rate was relatively low. This initial performance was expected, as the textual representation of the complaints was handled using the TF-IDF algorithm. This algorithm can cause confusion in classification when only a limited number of training samples are available.

However, over successive rounds of training, there was a continuous improvement in the model's ability to categorize complaints more accurately. With an increase in the number of samples, the identification of the most relevant words for each category became more accurate. This enhancement in keyword definition allowed the model to classify the different categories of complaints more effectively. As a result,

there was a significant increase in key evaluation metrics, including improvements in recall, precision, and F1-Score.

In the last round of testing, the model achieved an accuracy of 85%, and all categories recorded an F1-Score above 80%, reflecting a high performance in the categorization of complaints. A particular highlight was the notable improvement of 30% in recall for the category of payment-related issues, observed between the first and the last round of testing. This increase demonstrates the significant evolution of the model in accurately identifying these complaints over time.

The main contribution of this study is to present an efficient model, developed from the application of the CRISP-DM methodology, for the automated classification of complaints in e-commerce. Another point that deserves emphasis is the use of real, national data (complaints), which enables a better understanding of the main issues affecting the perception of quality among e-commerce customers in Brazil.

Such automation not only speeds up the response process to complaints, improving customer satisfaction, but also provides insights into recurring patterns of problems, enabling companies to identify and resolve systemic flaws in their products or services. Moreover, efficient and accurate categorization allows customer service resources to be optimized, directing complaints to the appropriate departments swiftly and impacting the response time and customer satisfaction. The integration of this type of computational approach with e-commerce platforms is essential for enhancing customer relationship management, enabling a more positive and efficient user experience, and contributing to long-term customer loyalty and trust.

The development of this work faced several challenges, particularly in obtaining a broad base of real complaints from e-commerce in Brazil, which proved to be a complex task. However, thanks to the collection of complaints carried out through the Reclame Aqui Platform, it was possible to compile a base consisting of 600 authentic complaints.

For future work, we recommend expanding this study by comparing various algorithms to determine which yields the best results, using a more extensive database for training and testing. Additionally, it would be beneficial to include new categories of problems to be classified, enhancing the scope and efficiency of the model for use in large e-commerce operations. Another interesting line of research would involve applying this same model to classify different types of text, beyond complaints, exploring its adaptability and effectiveness in various contexts.

## References

- Associação Brasileira de Comércio Eletrônico (ABComm) (2023). Principais indicadores do e-commerce no Brasil. Relatório online. Disponível em: <https://dados.abcomm.org/>. Acesso em: 01 jul. 2024.
- Ayanoğlu, E., Çolak, Z., Tanyel, T., Sarioğlu, H. Y., and Diri, B. (2023). Detection and classification of customer comments containing complaints. *Avrupa Bilim ve Teknoloji Dergisi*, 2023(52):37–45.
- Bokrantza, J., Subramanian, M., and Skoogh, A. (2024). Realising the promises of artificial intelligence in man-

- ufacturing by enhancing crisp-dm. *PRODUCTION PLANNING CONTROL*, 35(16):2234–2254. DOI: <https://doi.org/10.1080/09537287.2023.2234882>.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide. Technical report, The CRISP-DM consortium.
- de Leaniz, P. M. G. and del Bosque Rodríguez, I. R. (2016). Corporate image and reputation as drivers of customer loyalty. *Corporate Reputation Review*, 19:166–178. DOI: <https://doi.org/10.1057/crr.2016.2>.
- Falcão, H. S., Lovato, A. V., Santos, A., Oliveira, L., Manicoba, R. H. C., Guimarães, M. A., and Santana, M. S. (2013). Classificação de vagas de estacionamento com utilização de rede perceptron multicamadas. *Revista de Sistemas de Informação da FSMA, Visconde de Araújo*, 2013(12):41–48.
- Filho, F. S., Paillard, G., Carmo, R., Lima, E., and Bonfim, M. (2023). Classificação do diálogo freireano em mensagens de fóruns de discussão: Uma análise de desempenho do tf-idf e o bert para sentenças. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 1477–1488, Porto Alegre, RS, Brasil. SBC. DOI: <https://doi.org/10.5753/sbie.2023.235298>.
- Fleck, L., Tavares, M. H. F., Eyng, E., Helmann, A. C., and Andrade, M. A. d. M. (2016). Redes neurais artificiais: Princípios básicos. *Revista Eletrônica Científica Inovação e Tecnologia*, 1(13):47–57.
- Gonçalves, C. d. A. (2016). Análise de sentimentos em reclamações: Uma aplicação no maior site de reclamações do brasil. Dissertação de mestrado, Escola de Matemática Aplicada, Fundação Getulio Vargas, Rio de Janeiro, Brasil.
- Gorgens, E. B., Leite, H. G., Santos, H. d. N., and Gleriani, J. M. (2009). Estimação do volume de árvores utilizando redes neurais artificiais. *Revista Árvore*, 33:1141–1147. DOI: <https://doi.org/10.1590/S0100-67622009000600016>.
- Grandini, M., Bagli, E., and Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv*. DOI: <https://doi.org/10.48550/arXiv.2008.05756>.
- Haykin, S. (2001). *Redes neurais: princípios e prática*. Bookman Editora.
- Itsari, M. Y. I. and Budi, I. (2022). Classification of complaint categories in e-commerce: A case study of pt bukalapak. In *2022 5th International Conference on Information and Communications Technology (ICOIACT)*, pages 317–324. DOI: <https://doi.org/10.1109/ICOIACT55506.2022.9971933>.
- Jackson, P. and Moulinier, I. (2007). *Natural language processing for online applications: Text retrieval, extraction and categorization*, volume 5. John Benjamins Publishing. DOI: <https://doi.org/10.1075/nlp.5>.
- Just, J. (2024). Natural language processing for innovation search – reviewing an emerging non-human innovation intermediary. *Technovation*, 129:102883. DOI: <https://doi.org/10.1016/j.technovation.2023.102883>.
- Madanchian, M. (2024). The impact of artificial intelligence marketing on e-commerce sales. *Systems*, 12:429. DOI: <https://doi.org/10.3390/systems12100429>.
- Mitchell, T. M. (1997). *Machine learning*. McGraw-hill.
- Mitchell, T. M. (2006). The discipline of machine learning. Technical Report CMU-ML-06-108, Carnegie Mellon University.
- Monard, M. C. and Baranauskas, J. A. (2003). Conceitos sobre aprendizado de máquina. In Oliveira Jr., O. N., editor, *Sistemas Inteligentes: Fundamentos e Aplicações*, chapter 4, pages 81–108. Edusp.
- Olujimi, P. A. and Ade-Ibijola, A. (2023). Nlp techniques for automating responses to customer queries: a systematic review. *Technovation*, 3:102883. DOI: <https://doi.org/10.1007/s44163-023-00065-5>.
- Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., and Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, 48(3):128–138.
- Peixoto, L. H. R. (2021). *Aprendizado de Máquina Aplicado no Atendimento de Reclamações de Clientes*. Tese (doutorado), Universidade de São Paulo, São Paulo, Brasil.
- Pinto, S. C. S. (2015). Processamento de linguagem natural e extração de conhecimento. Dissertação de mestrado, Universidade de Coimbra, Coimbra, Portugal.
- Rabbi, B., Klug, D. B., Gonçalves, V. S., Júnior, E. R. G., and Brasil, J. A. (2018). Análise de reclamações sobre produtos e serviços no programa de proteção e defesa do consumidor utilizando mineração de dados. *Engevista*, 20(5):649–660.
- Rauber, T. W. (2005). Redes neurais artificiais. *Universidade Federal do Espírito Santo*, 29:39.
- Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Rodrigues, F. A., and da F. Costa, L. (2019). Clustering algorithms: A comparative approach. *PLOS ONE*, 14(1):e0210236. DOI: <https://doi.org/10.1371/journal.pone.0210236>.
- Rodriguez, M. M. M. S. and Bezerra, B. L. D. (2020). Processamento de linguagem natural para reconhecimento de entidades nomeadas em textos jurídicos de atos administrativos (portarias). *Revista de Engenharia e Pesquisa Aplicada*, 5(1):67–77. DOI: <https://doi.org/10.25286/rep.v5i1.1204>.
- Runeson, P. and Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering*, 14(2):131–164. DOI: <https://doi.org/10.1007/s10664-008-9102-8>.
- Sanches, M. K. (2003). Aprendizado de máquina semi-supervisionado: proposta de um algoritmo para rotular exemplos a partir de poucos exemplos rotulados. Mestrado em ciência da computação e matemática computacional, Universidade de São Paulo, São Carlos.
- Santos, K. B. C. d. et al. (2019). Categorização de textos por aprendizagem de máquina. Mestrado em modelagem computacional de conhecimento, Universidade Federal de Alagoas, Maceió.
- Santouridis, I. and Veraki, A. (2017). Customer relationship management and customer satisfaction: the mediating role of relationship quality. *Total Quality Man-*

- agement & Business Excellence, 28:1122–1133. DOI: <https://doi.org/10.1080/14783363.2017.1303889>.
- Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4):13–22.
- Silva, A. P. C. (2022). E-commerce: Impactos no consumo do segmento de beleza e saúde durante a pandemia covid-19. Graduação em logística, Universidade Federal do Tocantins, Araguaína.
- van Engelen, J. E. and Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine learning*, 109(2):373–440. DOI: <https://doi.org/10.1007/s10994-019-05855-6>.
- Vinayak, V. and Jyotsna, C. (2023). Consumer complaints classification using deep learning & word embedding models. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–5. DOI: <https://doi.org/10.1109/ICCCNT56998.2023.10307286>.
- Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39, Manchester, UK.
- Yanai, F. K. et al. (2020). Detecção de anomalias no funcionamento de software com machine learning. Mestrado em tecnologias da inteligência e design digital, Pontifícia Universidade Católica de São Paulo, São Paulo.
- Zhou, Z.-H. (2021). *Machine learning*. Springer Nature.
- Zuin, G. L., Magalhaes, L. F. G., and Loures, T. C. (2016). Mal-fitt: Myanimelist forum interpreter through text. *XIII Encontro Nacional de Inteligência Artificial e Computacional (SBC ENIAC-2016)*. Recife-PE: SBC, pages 205–216.