Leveraging Advanced Analytics to Prioritize Customer Feedback through Risk Classification in Retail

Inês Daniela Ferreira Faria

Master's Dissertation

Supervisor at FEUP: Professor Maria João Pires
Supervisor at Sonae MC: Ana Freitas

Master in Industrial Engineering and Management

2024-06-27

Abstract

Customer feedback is a crucial asset for retailers, providing insights into consumer behavior and highlighting areas for improvement. However, effectively managing and prioritizing this feedback, particularly in identifying severe-risk cases, poses significant challenges.

This dissertation explores the integration of advanced analytical methodologies, including generative artificial intelligence and machine learning techniques, such as natural language processing, topic modeling, and classification models, to enhance the efficiency and accuracy of handling customer complaints at Sonae MC, a leading retail company in Portugal. Specifically, the focus is on developing a systematic approach to classify and prioritize feedback based on the risk it poses to the company.

One of the main challenges is the absence of a structured dataset with identified risk cases. Thus, the project is structured into three phases, each addressing specific needs and objectives: topic modeling to reduce and refine the dataset; pre-classification of feedback in risk levels using the enterprise ChatGPT-4 API to manage the extensive volume and complexity of the data efficiently; and the development of machine learning models for precise risk classification.

The project begins with the application of Latent Dirichlet Allocation for topic modeling. This technique effectively narrowed down the size of the large, unstructured dataset, making it manageable for subsequent analysis by focusing on the topics where risk is typically more prevalent.

To further address the volume and complexity of the dataset, the ChatGPT-4 API was employed for pre-classification of feedback, significantly accelerating the time required for this task. The accuracy of the classification was ensured through a human-in-the-loop validation approach.

The final phase involved the implementation of machine learning classification models. The combination of Synthetic Minority Over-sampling Technique with random undersampling addressed class imbalance effectively, particularly enhancing tree-based model performance. Term Frequency–Inverse Document Frequency vectorization outperformed other methods, and the Random Forest model emerged as the top performer after hyperparameter tuning, achieving the highest F1-score.

This study highlights the practical benefits of integrating advanced analytical techniques for customer feedback management, providing a scalable and robust framework for prioritizing high-risk cases. Future work should focus on refining these methodologies, exploring more advanced topic modeling techniques, and extending hyperparameter tuning to other models to further improve classification accuracy. By continually enhancing these approaches, businesses can better leverage customer feedback, ensuring timely and effective responses to critical issues, thereby enhancing customer satisfaction and operational efficiency.

Keywords: Customer Feedback, Generative AI, Natural Language Processing, Retail, Risk Prioritization, Text Classification

Resumo

O *feedback* dos clientes é um ativo crucial para os retalhistas, proporcionando conhecimento sobre o comportamento do consumidor e destacando áreas para melhoria. No entanto, gerir e priorizar eficazmente esse *feedback*, particularmente na identificação de casos de alto risco, apresenta desafios significativos.

Esta dissertação explora a integração de metodologias analíticas avançadas, incluindo técnicas de inteligência artificial generativa e aprendizagem automática, como processamento de linguagem natural, modelação de tópicos e modelos de classificação, para melhorar a eficiência e a precisão no tratamento de reclamações de clientes na Sonae MC, uma empresa líder no setor de retalho em Portugal. Especificamente, o foco é desenvolver uma abordagem sistemática para classificar e priorizar o *feedback* com base no risco que representam para a empresa.

Um dos principais desafios é a ausência de um conjunto de dados estruturado com casos de risco identificados. Assim, o projeto está estruturado em três fases, cada uma abordando necessidades e objetivos específicos: modelação de tópicos para reduzir e refinar o conjunto de dados; pré-classificação do *feedback* em níveis de risco, utilizando a API empresarial do ChatGPT-4, para gerir eficientemente o volume e a complexidade dos dados; e o desenvolvimento de modelos de aprendizagem automática para uma classificação precisa de risco.

O projeto inicia com a aplicação da alocação latente de Dirichlet para a modelação de tópicos. Esta técnica reduziu eficazmente o tamanho do conjunto de dados, grande e não estruturado, para análises subsequentes ao focar-se em tópicos onde tipicamente o risco está mais prevalente.

Posteriormente, para ultrapassar o restrição do volume e da complexidade do conjunto de dados, a API do ChatGPT-4 foi utilizada para a pré-classificação do *feedback*, acelerando significativamente o tempo necessário para esta tarefa. A precisão da classificação foi garantida através de uma abordagem de validação humana.

A fase final envolveu a implementação de modelos de classificação de aprendizagem automática. A combinação da técnica de sobreamostragem de minoria sintética com subamostragem aleatória abordou eficazmente o desequilíbrio de classes, melhorando particularmente o desempenho dos modelos baseados em árvores. A vetorização Frequência do Termo-Inverso da Frequência nos Documentos superou outros métodos, e o modelo Random Forest emergiu como o melhor desempenho após a afinação de hiperparâmetros, atingindo o F1-score mais alto.

Este estudo destaca os benefícios práticos da integração de técnicas analíticas avançadas para a gestão do *feedback* dos clientes, proporcionando uma estrutura escalável e robusta para a priorização de casos de alto risco. Trabalhos futuros devem focar-se em refinar estas metodologias, explorando técnicas de modelação de tópicos mais avançadas e estendendo a afinação de hiperparâmetros para outros modelos para melhorar ainda mais a precisão da classificação. Ao melhorar continuamente estas abordagens, as empresas podem aproveitar melhor o *feedback* dos clientes, assegurando respostas oportunas e eficazes para questões críticas, aumentando assim a satisfação do cliente e a eficiência operacional.

Palavras-chave: Feedback de Clientes, Classificação de Texto, IA Generativa, Priorização de Riscos, Processamento de Linguagem Natural, Retalho

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my parents. Your guidance has always shown me that the path to success is made step by step and that with dedication and passion, anything is possible. Your endless encouragement and belief in me have been my greatest strength. You have always put my happiness above everything else, making me believe every day that the future is indeed what we make of it. I am especially grateful to my brother, who continuously shows me how to simplify life and find joy in the simplest moments.

Secondly, I want to express my gratitude to my university friends, who have become lifelong friends. Words cannot express how grateful I am for your love and support. Thank you for being the most attentive and encouraging friends I could ever ask for. I am certain that my journey would not have been as joyful, and I would not have believed so many times that the path is to infinity and beyond.

Thirdly, my heartfelt thanks to my housemates, who listened to me and believed in me every day. It is a blessing to have two places I can call home and a second group of people I can call family. Your belief in me and your daily encouragement has been a constant source of motivation and comfort.

To my supervisor, Professor Maria João Pires, thank you for your insightful guidance, which was instrumental in developing this project. I am grateful for your mentorship and for the time you dedicated to helping me.

I extend my deepest thanks to Sonae MC for giving me the opportunity to develop this project. A special thank you goes to Alexandre, Catarina Almeida, Catarina Pinho, Cláudia, Filipe, Francisco, and Teresa for making me feel at home and for your willingness to share your profound knowledge daily.

In particular, I owe my deepest gratitude to Ana Freitas and Ana Carvalho. Your availability and unwavering support throughout the development of this dissertation were invaluable. Your expertise and encouragement provided me with the confidence and knowledge necessary to overcome challenges and achieve my goals. I am profoundly grateful for your patience, your valuable feedback, and for always pushing me to strive for excellence.

Thank you all from the bottom of my heart for being an integral part of this journey.

"To infinity and beyound."

Buzz Lightyear

Contents

Acronyms and Symbols

AI	Artificial Intelligence
API	Application Programming Interface
BoW	Bag of Words
ChatGPT	Chat Generative Pre-trained Transformer
CRISP-DM	Cross-Industry Standard Process for Data Mining
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GenAI	Generative Artificial Intelligence
GPT	Generative Pre-trained Transformer
LDA	Latent Dirichlet Allocation
LLMs	Large Language Models
LR	Logistic Regression
LSA	Latent Semantic Analysis
MMH	Maximum Margin Hyperplan
NLP	Natural Language Processing
NMF	Non-Negative Matrix Factorization
NPMI	Normalized Pointwise Mutual Information
pLSA	Probabilistic Latent Semantic Analysis
RF	Random Forest
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machines
TF-IDF	Term Frequency–Inverse Document Frequency
TN	True Negative
TP	True Positive
TPR	True Positive Rate
XGBoost	Extreme Gradient Boost

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides an introduction and context for the project, elaborating on its motivations, objectives, and the challenges it aims to address. It further delineates the methodology employed and outlines the structure of the dissertation.

1.1 Project Motivation

Over the past decade, retailers have increasingly recognized the importance of understanding customer preferences and shopping patterns for running a successful business. Studies indicate that one of the most valuable types of knowledge for a business is understanding their customers (Wölbitsch et al., 2020). Therefore, the challenge lies in effectively collecting, managing, and utilizing insights about consumer behavior and feedback.

In the intensely competitive retail landscape, navigating customer feedback is more than just gathering customer opinions and thoughts; it is a unique opportunity to gain actionable insights into the consumer's behavior, make data-driven decisions, and ultimately, lead to sustainable growth (Agag et al., 2023). By meticulously collecting and analyzing customer feedback, companies demonstrate their commitment to customer centricity and responsiveness, thereby strengthening the bond between the brand and its customers. This, in turn, not only fosters loyalty and retention, but also sends a profound message: each customer's voice holds weight, and their input is genuinely valued.

Despite the wealth of feedback data, a significant challenge persists: the effective leveraging of customer feedback to drive strategic improvements and strengthen customer relationships. Although customer feedback often may indicate dissatisfaction with the company's products or services, it also serves as a valuable source of information regarding areas that require improvement and often serves as an early indicator of potential risks to the business (Rahim et al., 2019). For instance, feedback may reveal the potential risk of customers escalating their issues to external entities, which inevitably poses a risk to the company (Carlson et al., 2022).

Therefore, there is a need for systematic methods for prioritizing customer feedback based on its potential impact on the company, as not all feedback carries equal weight in terms of its

implications. Effective prioritization of customer feedback may help manage the workload more effectively and allow for a more efficient allocation of resources to ensure that issues are addressed in a timely manner.

Accordingly, this project is driven by the necessity to fully leverage analytical tools to maximize the potential of customer feedback. By prioritizing feedback effectively, this approach can help enhance consumer satisfaction and minimize potential threats to a company's reputation and overall performance in the competitive retail market.

1.2 Sonae MC and its Loyalty Program

Sonae MC is a subsidiary of the Sonae Group, one of Portugal's largest multinational corporations, and operates a diverse range of business segments, including Continente hypermarkets, Continente Modelo, and Continente Bom Dia convenience supermarkets. Its portfolio also includes health, wellness and beauty products, textile stores for children and adults, products and services for pets, cafeterias and restaurants, and bookshops and stationery stores, thereby catering to a wide spectrum of consumer needs and preferences.

MC has over thirty years of operational experience and has established itself as a leader in the Portuguese food retail business. Each store format serves a different customer base with varying preferences, offering a large range of products to enhance the shopping experience.

Its unwavering focus on customer centricity and innovation is central to MC's success, fostering customer loyalty and providing invaluable insights into individual shopping habits. Accordingly, the company introduced the Cartão Continente on 23 January 2007, which became Portugal's largest loyalty program. By 2023, the Cartão Continente had enrolled over 4 million families, and it is used in more than 2,000 national points of sale. This initiative has transformed the way the company interacts with its customers, enabling it to gain insights into consumer behavior, preferences, and purchasing patterns.

The Cartão Continente serves as a critical touchpoint between the company and its customers, facilitating a holistic view of customer behavior. Consequently, this project is conducted within the Advanced Analytics and Insights team of Cartão Continente department.

1.3 Project Objectives

This project aims to address the challenges identified in the current process of managing customer feedback at Sonae MC, particularly in identifying and prioritizing severe-risk cases.

Currently, the company lacks a structured and systematic approach to classify and prioritize feedback based on the risks inherent to the cases described by customers. This can result in delayed responses to critical issues, potential escalations, and, ultimately, a negative impact on customer satisfaction and the company's reputation.

Therefore, it was identified the need to develop a structured methodology that accurately assesses the level of risk associated with each piece of feedback. This system is designed to enable

quicker and more effective responses, prioritizing high-priority feedback to prevent them from developing into major complications and protecting the company's reputation. This project aims not only to meet these immediate needs but also to establish a standard for proactive risk management in customer service within the retail industry.

1.4 Project's Methodologies

To accomplish the project goals, the work will be organized into three distinct phases, each targeting specific needs and goals: 1) topic modeling to refine and reduce the dataset, 2) pre-classification of feedback utilizing generative artificial intelligence (GenAI) to efficiently manage and accelerate the process of handling the extensive volume and complexity of the data, and 3) feedback's risk classification through machine learning models.

The project will begin by utilizing advanced natural language processing (NLP) techniques, such as topic modeling, to identify the main topics present in the customer feedback data. This step is crucial not only for extracting and understanding prevalent themes from the unstructured feedback but primarily for narrowing down the dataset. The cases are not pre-classified in terms of risk, and the dataset is too extensive to be entirely manually pre-classified. Thus, topic modeling will be used to identify topics that are typically less likely to encompass cases with significant risk potential, thereby streamlining the focus of subsequent analysis.

Following, advanced generative artificial intelligence technologies will be implemented to classify the feedback based on their severity and potential impact on the company, including categorizing complaints into 3 risk levels: high, medium, and low. In this phase, these technologies are chosen due to their ability to significantly accelerate the classification process. Given the extensive volume and complexity of the dataset, manual classification would be impractical and time-consuming. To ensure the reliability and relevance of the classifications, the generative AI's output will be validated by the Customer Service team, thereby maintaining high standards of accuracy and precision in the classification process. Specifically, an overview will be conducted for low-risk cases, while high-risk cases will undergo a deep review and detailed analysis.

The final phase involves employing machine learning algorithms to develop a classification model that assesses the feedback in terms of risk. These models will be trained on the categorized feedback to improve their accuracy and reliability.

1.5 Dissertation Structure

The present dissertation is organized into six chapters. Chapter 2 presents a comprehensive review of the literature related to the key themes of the project. It examines the importance of customer feedback in retail, existing methodologies for feedback analysis, and the application of advanced analytics in customer feedback prioritization, such as natural language processing. Additionally, it discusses the application of generative AI for enhancing feedback processing and decision-making. Chapter 3 provides a detailed description of the problem, followed by a presentation of

the initial situation of the project and an overview of the proposed approach. Chapter 4 focuses on the topic modeling phase, detailing the methodologies employed and presenting the key results obtained. Chapter 5 delves into the pre-classification of feedback, detailing the methods and outcomes achieved through generative AI technologies. Chapter 6 is dedicated to the classification modeling, elaborating on the machine learning techniques applied and the respective results. Lastly, Chapter 7 summarizes the key conclusions of the project, identifying some opportunities for future work.

Chapter 2

Literature Review

This chapter provides a comprehensive review of the existing literature relevant to the project. It encompasses the critical analysis of customer feedback, its significance in the retail sector, and the methodologies employed for its analysis and prioritization. Furthermore, it delves into the application of natural language processing and generative AI in enhancing feedback analysis and decision-making. By examining these themes, this chapter establishes a foundational understanding of the current state of research and identifies gaps that this project aims to address.

2.1 Customer Feedback

In today's market, customers expect a high-quality shopping experience, which includes the prompt resolution of any feedback they may have regarding their purchase or any perceived infringements of their rights (Yang et al., 2019). On the other hand, businesses aim to understand how consumers react to their products and services in order to adapt their strategies and make informed business decisions (Yilmaz et al., 2016). Customer feedback serves as a direct representation of user experience and perspectives, making it essential to collect and analyze both structured and unstructured feedback provided by customers (Xin, 2017) . Moreover, feedback is also an important indicator of organizational performance, highlighting issues in internal processes that require prompt resolution (Zairi, 1992). Businesses must recognize that losing customers affects profits and generates negative word-of-mouth, which can further harm the company's reputation (Filip, 2013).

Consumers can be motivated to provide feedback for various reasons, including both positive experiences, when their expectations are met or exceeded, and negative experiences. Some customers may take action with minor issues, particularly if they believe the company has an effective and swift response system, and others may be driven to complain due to situations they consider serious and intolerable (Filip, 2013). In addition, customers may take different actions to register their feedback, including submitting them directly to the company. If the resolution provided by customer service is deemed unsatisfactory, the customer may also escalate their concerns beyond the company, seeking recourse through external entities, pursuing resolution through legal procedures, or even publicly sharing their negative experiences on social media (Yang et al., 2019).

Furthermore, there can be several barriers that may prevent customers from expressing their feedback. Wirtz and Lovelock (2021) identified some disincentives, such as the inconvenience of feedback procedures, the time required to provide it, a lack of confidence in the organization's ability to address and resolve the case, and the fear of feeling embarrassed or reprimanded during interactions with employees. Therefore, it is important for companies to establish efficient and accessible channels for the submission of feedback and to address them timely.

Through these channels, which include platforms established by the businesses as well as more spontaneous feedback like blogs, customer feedback emails, and websites, companies receive a wide range of inputs from their clients (Gamon et al., 2005). The volume of spontaneous feedback has been increasing, and although these sources offer rich information, they are considerably less structured than traditional surveys, as the information is presented in free text rather than in predefined answer sets. Additionally, textual customer feedback data is frequently noisy, containing, for instance, misspellings and word abbreviations. This complexity and lack of structure present significant challenges for businesses attempting to extract valuable insights.

2.1.1 Feedback Analysis and Prioritization

Despite efforts and the development of advanced techniques and methods, such as text mining and key term extraction, challenges remain in learning from customers, particularly in the systematic collection and analysis of data (Fabijan et al., 2015). Analytic technologies to analyze customer feedback data have emerged as powerful tools to address these issues, enabling more accurate and efficient processing of such unstructured data (Xin, 2017). Furthermore, despite its potential, academic research on the challenges of customer feedback analysis remains limited. A lot of the existing work has concentrated on enhancing models to more accurately determine the sentiment expressed in textual feedback. Identifying sentiment is crucial, but more detailed information is contained in textual feedback, such as the components that drive customer sentiments (Anantharaman et al., 2019).

Moreover, the way companies deal with feedback can affect the post-feedback behavior of consumers, including customer loyalty and repeat business (Kumar and Kaur, 2020). Furthermore, the classification of complaint severity can help businesses better understand the nature and urgency of the feedback, facilitating more tailored and effective responses, aligning with the focus on prioritizing feedback based on risk and urgency (Jin and Nikolaos, 2021).

A study conducted by Yang et al. (2019) introduces an approach to identify high-risk complaints that could escalate beyond the company's internal resolution processes. This methodology combines the predictive capabilities of recurrent neural networks with manually-engineered features to effectively flag potential escalations in real time. Blümel and Zaki (2022) aim to automatically prioritize complaints, comparing classical machine learning and deep learning NLP techniques to determine their efficacy. The prioritization of customer complaints is determined using three main criteria: subjectivity, polarity, and frequency of topics. In the healthcare sector, Laliberté et al. (2022) underscores the necessity for a system to prioritize complaints based on urgency. The study supports a two-level priority system where urgent complaints are addressed

immediately, while other complaints are handled on a first-come, first-served basis. The study conducted by Ghodousi et al. (2019) discusses the utilization of data mining techniques, such as K-means and Bees Algorithms, to analyze and prioritize urban needs. Furthermore, Jin and Nikolaos (2021) introduces a detailed classification of complaints into four severity levels: no explicit reproach, disapproval, accusation, and blame, based on the degree of threat the complainer is willing to undertake. They employ transformer-based networks combined with linguistic information to classify the severity of complaints.

2.2 Natural Language Processing

There has been an exponential growth in the volume of consumer generated textual content, and an inevitable need to extract insights or relevant information from such content. Natural Language Processing emerges as an analytical technique to empower the extraction of insights contained in a text, enabling the communication between computers and humans in natural language (Kang et al., 2020). It deals with the automatic processing and analysis of unstructured textual information, including a wide range of tasks in different domains, such as machine translation, speech recognition, and sentiment analysis (Kang et al., 2020).

NLP tasks are varied and considered complex due to the inherent ambiguities in human language. Idioms, metaphors, grammar, and usage exceptions are among the irregularities that make it challenging to determine the intended meaning of text data (Jusoh, 2018). Performing NLP involves several steps, including data cleaning, preprocessing, text representation, model training, and evaluation.

Before diving into the algorithms, taking the necessary cleaning and preprocessing steps is important, including techniques such as tokenization, stemming, lemmatization, and stop-word removal (Hickman et al., 2022). Once the data is preprocessed, researchers must select appropriate methods for representing words in a structured way, to make possible the deployment of the algorithms, such as topic modeling, text classification, or sentiment analysis (Dogra et al., 2022).

In the context of customer feedback analysis, various NLP technologies have been explored to enhance the process. The study conducted by Katsiuba et al. (2024) discusses various NLP technologies like sentiment analysis, content analysis, and text generation to augment the process of responding to online customer feedback. They identified several challenges faced in this context, such as the imperative need for personalization, the maintenance of response quality, and logistical obstacles to timely responses. Moreover, topic modeling, a NLP technique, has been applied in various domains such as health, hospitality, education, social networks, and finance. Its utility extends to both academic and industrial purposes, with numerous studies concentrating on applying specific topic modeling techniques within particular domains (Krishnan, 2023). Such advancements underscore the evolving landscape of NLP and its capacity to address complex analytical challenges in customer feedback analysis.

2.2.1 Topic Modeling

In the current age of information, the amount of written material encountered each day makes it beyond the capacity of an individual to process large collections of unstructured bodies of text. In the realm of text analysis, one common goal is to discover patterns of words, which involves determining the topics or concepts discussed within a document. Techniques such as topic modeling are directly related to obtaining insights and organizing written data and are essential for achieving this goal (Cai et al., 2021).

Topic modeling is a widely used statistical method for uncovering hidden variables within large datasets (Vayansky and Kumar, 2020). A topic is a group of words that often co-occur, which infers that the topics generated by topic modeling are a set of clusters of similar words (Alghamdi and Alfalqi, 2015). Topic modeling is a versatile method, being applied across various fields, and although it is strongly related to text data, it has also been used to analyze, for instance, images, biological data, and survey information (Kim et al., 2013).

Various applications of topic modeling have been extensively explored in the literature. One notable application is its facilitation of smart literature reviews, enabling researchers to efficiently select and organize relevant topics from research papers, thereby streamlining the review process (Krishnan, 2023). Furthermore, Pyasi et al. (2018) developed a tool for analyzing student feedback using sentiment analysis, topic modeling, and visualization. In this study, topic modeling was effective in extracting multiple topics from single comments (Kastrati et al., 2021).

Further applications of topic modeling include analyzing tweets related to the COVID-19 vaccine, where it was used to extract topics and conduct sentiment polarity analysis (Krishnan, 2023). Furthermore, Kim et al. (2013) introduces a framework that combines probabilistic topic modeling with time-series causal analysis, iteratively refining topics to enhance their semantic coherence and correlation with time-series data. In the context of retail, Cuffie et al. (2020) applied Latent Dirichlet Allocation (LDA), a topic modeling technique, to analyze customer return data, providing detailed insights into the reasons behind product returns. This application of LDA helped identify emerging patterns in return reasons, which are crucial for improving product descriptions and enhancing the overall customer shopping experience. Additionally, Carrasco et al. (2021) demonstrates the powerful application of topic models in retail analytics, particularly in identifying grocery shopping patterns and summarizing customer transactional data.

There are also some inherent challenges in topic modeling, such as the difficulty in choosing the number of topics, optimizing model parameters, and ensuring model stability and interpretability (Kherwa and Bansal, 2020). Moreover, topic modeling is an unsupervised technique. Therefore, there are no labels to rely on during the evaluation. There are no metrics such as F1 or accuracy, and the validation measures may not indicate a clear estimation of a topic model performance. Accordingly, it is important to add human evaluation and domain knowledge (Albanese, 2022).

2.2.1.1 Topic Modeling Approaches

There are several well-known techniques for topic modeling, generally divided into probabilistic and non-probabilistic models (Kastrati et al., 2021). Among the most popular non-probabilistic models are Latent Semantic Analysis (LSA) and Non-Negative Matrix Factorization (NMF). Examples of popular probabilistic models are Probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation. (Kherwa and Bansal, 2020).

Latent Semantis Analysis is a technique that captures the semantic relationships between words, by employing a vector space model to measure text similarity, thereby facilitating the identification of words with similar meanings and organizing them into coherent semantic clusters (Evangelopoulos, 2013).

Non-negative Matrix Factorization is a dimension reduction technique designed to handle datasets containing negative numbers by enforcing non-negativity constraints. Applications of NMF include image processing, segmentation, pattern recognition, and language modeling, which highlights its effectiveness in extracting patterns in complex datasets (Lopes and Ribeiro, 2015). This method outstands with shorter texts and offers a faster computational process. However, NMF is not probabilistic, which can limit its robustness when dealing with varied and noisy datasets, compared to probabilistic methods like LDA (Lind et al., 2022).

Probabilistic Latent Semantic Analysis is described as an improvement over LSA that integrates a probabilistic framework to model each document as a mixture of latent topics, where each topic is characterized by a distribution over words. It operates on the bag of words (BoW) model (Kherwa and Bansal, 2020). This method is advantageous for large corpora as it better handles uncertainty and variability in the data. Nonetheless, PLSA can suffer from overfitting and requires smoothing techniques to enhance its performance, making it less straightforward to implement effectively (Lind et al., 2022).

Furthermore, Latent Dirichlet Allocation is a generative probabilistic model that assumes that documents are composed of multiple topics, and provides a deeper insight into the mixture of topics within a document collection by using a Dirichlet prior over the latent topics (Blei et al., 2003). This method is particularly effective for long, cohesive texts, such as research articles and news reports, due to its robust handling of extensive and structured data (Lind et al., 2022).

2.2.1.2 Topic Modeling Evaluation Measures

Evaluating topic models presents a significant challenge, as there is no universally accepted approach (Krishnan, 2023). In the literature, perplexity, log-likelihood, and coherence are commonly employed for evaluating the quality of topic models (Kherwa and Bansal, 2020).

Perplexity assesses how well a probability model predicts a sample, with a lower perplexity score indicating better predictive performance (Tijare and Rani, 2020). Conversely, log-likelihood measures the likelihood that the model that generated the topics would generate the observed data. A higher log-likelihood value indicates a model better fitting to the data (Tijare and Rani,

2020). These methods are typically associated with probabilistic models that utilize expectation maximization algorithms for topic learning (Kherwa and Bansal, 2020).

Another evaluation measure for topic modeling is topic coherence, which measures how meaningful the topics generated by the model are, checking whether the high-probability words in each topic frequently appear together in the same documents in the corpus (Tijare and Rani, 2020). Coherence effectively assesses how closely related the words within a topic are, factoring in their statistical independence, making it a suitable metric for evaluating topic models regardless of their underlying computational approach (Kherwa and Bansal, 2020). Various researchers, such Krishnan (2023) and Abdelrazek et al. (2023), consider this metric the most suitable when human users interpret the output. Among the various coherence measures, CV, UCI, UMass, and Normalized Pointwise Mutual Information (NPMI) are the most commonly used (Hadiat, 2022).

2.3 Machine Learning for Text Classification

The application of machine learning classifiers in the field of text classification is a common practice among researchers (Kadhim, 2019). Text classification plays an important role in organizing documents into predefined categories and it has been used for various tasks such as finding related documents, categorizing subjects by documents, and organizing documents into different subjects (Kadhim, 2019). Nowadays, with the vast amounts of text documents generated daily globally, the need for text classification is more critical than ever (Hassan et al., 2022).

These methods can be broadly categorized into supervised and unsupervised learning. Supervised algorithms can be compared and evaluated based on metrics such as F1 score, recall, and precision. On the other hand, unsupervised algorithms typically rely on metrics that may be challenging to analyze (Kang et al., 2020).

Supervised machine learning is a well-studied approach for text classification and information retrieval, involving automated categorization of documents into predefined classes based on their content (Kadhim, 2019). Further, text classification can be divided into two main types: single-label and multi-label classification. In single-label classification, each document is assigned only one predefined class, while in multi-label classification, a document can be assigned multiple classes (Kadhim, 2019). Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes, and Decision Tree are particularly popular supervised learning algorithms (Kang et al., 2020).

2.3.1 Logistic Regression

Logistic Regression is a widely utilized linear model for binary classification tasks, offering a robust foundation for various machine learning applications. It operates by modeling the probability that a given input belongs to a particular class, typically represented as either "0" or "1". The model is grounded in the principles of supervised learning, where it leverages labeled training data to learn the relationship between the input features and the binary outcome (Hassan et al., 2022).

Logistic Regression is known for its simplicity and interpretability, making it a good baseline model for binary classification problems. The model estimates the probability of the dependent variable as a function of the independent variables using the logistic function, also known as the sigmoid function (Occhipinti et al., 2022).

The model can be expressed by Equation 2.1, where $P(Y = 1|X)$ is the probability of the dependent variable Y being 1 given the independent variables $X_1, X_2, \ldots, X_n$. The coefficients $\beta_0, \beta_1, \ldots, \beta_n$ are estimated from the training data using the maximum likelihood estimation method (Occhipinti et al., 2022).

$$\text{P(Y=1|X)} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}} \tag{2.1}$$

Feature scaling is further crucial for Logistic Regression, as the algorithm computes distances and assumes that features are on a similar scale. Without scaling, features with larger ranges can disproportionately influence the model, leading to suboptimal performance.

Logistic Regression is considered highly effective for binary classification problems. One advantage is its interpretability, which can be simpler than other machine learning algorithms. Previous studies have demonstrated its capacity, for instance, in spam detention (Occhipinti et al., 2022).

2.3.2 Support Vector Machines

Support Vector Machines are a powerful tool primarily used for classification problems, although they can also be applied to regression tasks (Hassan et al., 2022). The fundamental objective of SVM is to classify data by finding the optimal hyperplane that maximizes the margin between different classes. This optimal hyperplane is referred to as the Maximum Margin Hyperplane (MMH), and it is determined by the support vectors, which are the data points closest to the hyperplane (Occhipinti et al., 2022).

SVM is versatile in handling both linear and non-linear classification problems. For linear data, SVM directly identifies the MMH in the original feature space. However, real-world data often exhibit non-linear patterns, necessitating a more sophisticated approach (Hartmann et al., 2019). In such cases, SVM employs kernel functions to transform the original feature space into a higher-dimensional space. This transformation allows the SVM to construct linear hyperplanes in this new space, which correspond to non-linear decision boundaries in the original feature space (Hassan et al., 2022).

To this model, feature scaling is important since the algorithm's optimization process involves calculating distances between data points. If features are not on a similar scale, those with larger ranges can dominate the distance calculations, leading to biased model performance.

Support Vector Machines are often utilized in text classification due to their ability to achieve high precision, meaning they accurately identify relevant instances (Ikonomakis et al., 2005).

2.3.3 Random Forest

Random Forest is a robust ensemble learning method primarily used for classification and regression tasks. It is a powerful and versatile machine learning algorithm known for its high accuracy, resistance to overfitting, and capability to provide valuable insights into feature importance (Occhipinti et al., 2022).

The Random Forest algorithm initiates by creating multiple subsets of the original dataset through bootstrap sampling. This involves random sampling with replacement, resulting in some data points appearing multiple times within a subset while others may be excluded. For each subset, a decision tree is constructed. In contrast to traditional decision trees, each node in a Random Forest tree is split based on the best feature selected from a random subset of features (random feature selection) (Chen et al., 2020).

RF typically employs the decrease of Gini impurity criterion for split selection, where Gini impurity measures the likelihood of misclassification at a specific node in a decision tree based on the distribution of labels in the subset (Occhipinti et al., 2022). The Gini impurity for a binary classification problem is shown in Equation 2.2, where p_i is the probability of class i at a particular node.

$$\text{Gini} = 1 - \sum_{i=1}^{n} p_i^2 \tag{2.2}$$

Random Forest is recognized for its high predictive performance, attributable to the ensemble approach that averages out biases and diminishes variance. The randomization introduced via bootstrap sampling and random feature selection confers robustness against overfitting, particularly when employing a large number of trees (Hartmann et al., 2019). RF can furthermore assess the importance of each feature in predicting the target variable (Jaiswal and Samikannu, 2017).

2.3.4 Naive Bayes

Naive Bayes classifiers are based on Bayes' theorem with the assumption of independence between features. Despite its simplicity, Naive Bayes is highly efficient and performs well with high-dimensional data. (Ikonomakis et al., 2005) This classifier is well-suited for text classification tasks where the assumption of feature independence is often reasonably met (Baharudin et al., 2010).

The foundation of Naive Bayes is Bayes' theorem, outlined in Equation 2.3, where $P(C|X)$ is the posterior probability of class C given the features X, $P(X|C)$ is the likelihood of features X given class C, $P(C)$ is the prior probability of class C, and $P(X)$ is the prior probability of features X.

$$\text{P(C|X)} = \frac{P(X|C) \cdot P(C)}{P(X)} \tag{2.3}$$

In text classification, Naive Bayes demonstrates notable effectiveness by leveraging the probabilistic relationships between words and document classes (Hartmann et al., 2019). The classifier estimates the likelihood of a document belonging to a particular class based on the presence of

specific words, assuming that the occurrence of each word is conditionally independent of the others given the class.

2.3.5 Extreme Gradient Boosting

XGBoost is an optimized gradient boosting algorithm designed to be highly efficient, and flexible. The algorithm works by iteratively combining weak "learners" into a stronger "learner" by constructing new models that aim to correct the errors of the previous ones (Occhipinti et al., 2022).

At each stage of the gradient boosting iteration, XGBoost improves upon the current model by using the residuals of the previous model to construct a new one. This iterative process helps to refine the model and improve its predictive accuracy. XGBoost supports various objective functions, making it versatile for a range of tasks, including regression, classification, and ranking. It is optimized for sparse input data and has demonstrated superior performance compared to other machine learning methods in diverse applications. This performance superiority is attributable to its advanced optimization techniques, including regularization, which mitigates overfitting, and its sophisticated handling of missing data, which ensures robustness in real-world applications (Dhieb et al., 2019).

The objective function in XGBoost consists of a loss function and a regularization term. This is outlined in Equation 2.4, where $l(y_i, \hat{y}_i)$ is the loss function that measures the difference between the predicted value $\hat{y}_i$ and the actual value y_i, and $\Omega(f_k)$ is the regularization term that penalizes the complexity of the model. The component n is the number of instances, K is the number of trees, and ϕ represents the parameters of the model (Dhieb et al., 2019).

$$\mathscr{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.4}$$

XGBoost leverages its gradient boosting framework to handle high-dimensional data and can efficiently process text data by converting it into numerical features. The algorithm then iteratively improves the classification model by focusing on the misclassified instances from previous iterations, ultimately yielding a highly accurate text classifier (Dhieb et al., 2019).

XGBoost is a highly efficient and effective algorithm designed for parallel tree-boosting, optimizing both speed and performance. It has demonstrated superior performance compared to other machine learning methods in a variety of applications (Occhipinti et al., 2022).

2.3.6 Feature Selection

The primary goal of feature selection is to choose a subset of features from the original documents that provide a better understanding and improved performance for the learning process (Baharudin et al., 2010). Feature selection is crucial for simplifying models by reducing the number of parameters, decreasing training time, enhancing generalization, and avoiding the curse of dimensionality.

This process is especially beneficial in high-dimensional datasets where many features are present (Jaiswal and Samikannu, 2017).

Random Forest has emerged as an effective algorithm for handling feature selection, even with a large number of variables. It provides a robust mechanism to identify and eliminate unimportant features, thereby improving model accuracy and performance (Chen et al., 2020). The algorithm generates an unbiased estimate of the generalization error during the forest construction process, making it reliable for future data additions. The Gini index is used as a criterion to measure the accuracy of decision trees, with higher improvements in Gini decrease indicating higher importance of the feature (Jaiswal and Samikannu, 2017).

For instance, a study conducted by Chen et al. (2020) utilized various feature selection methods alongside Random Forest, such as varImp(), Boruta, and Recursive Feature Elimination (RFE). The results consistently highlighted that Random Forest-based feature selection methods provided better accuracy.

2.3.7 Class Balancing

Imbalanced datasets pose a critical problem for machine learning algorithms since they can lead to insufficient training of the minority class, undermining the predictive accuracy of machine learning models. Under-representation of minority class samples is a common problem that needs addressing to improve model performance (Sha et al., 2022).

In class imbalance scenarios, the focus on overall accuracy can lead to models that perform well on the majority class but fail to adequately capture the characteristics of the minority class. This failure is critical because the minority class often represents the more important or rare events that need to be detected accurately (Ali et al., 2015).

To address the issue of imbalanced datasets, various class balancing techniques have been developed. These techniques aim to balance the majority and minority class samples, thereby enhancing the model's ability to learn effectively from all classes.

Undersampling techniques reduce the number of majority class samples to balance the dataset. The simplest form, random undersampling, involves randomly removing samples from the majority class (Sha et al., 2022).

Oversampling techniques increase the number of minority class samples to achieve a balanced dataset. The most widely used oversampling method is the Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic samples by interpolating between existing minority class samples (Ali et al., 2015).

Furthermore, combining undersampling and oversampling techniques can leverage the strengths of both methods to achieve better performance (Sha et al., 2022).

2.3.8 Evaluation Metrics

In the context of imbalanced datasets, metrics such as precision, recall, and F1-score become particularly important as they provide a more nuanced view of model performance. Traditional

metrics such as accuracy, although important, can be misleading, since they may be dominated by the majority class (Seliya et al., 2009).

- **Accuracy** (Equation 2.5): Measures the overall correctness of the classification model. It represents the ratio of correctly classified instances (both positive and negative) to the total number of instances in the dataset (Hassan et al., 2022).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.5}$$

- **Precision** (Equation 2.6): Measures the proportion of positive results that were correctly classified. It is defined as the number of correct positive results (true positive, TP) divided by the number of all the positive results (total of TP and false positive, FP) (Hassan et al., 2022).

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2.6}$$

- **Recall** (Equation 2.7): Measures the proportion of actual positive results that were correctly classified by the algorithm. It is defined as the number of correct positive results (TP) divided by the number of positive results that should have been returned (total of TP and false negative, FN) (Hassan et al., 2022).

$$\text{Recall} = \frac{TP}{TP+FN} \tag{2.7}$$

- **F1-score** (Equation 2.8): The F1-score measures the accuracy of the test. A weighted average of precision and recall, providing a single metric to assess the balance between them, and it is defined as follows (Hassan et al., 2022).

$$\text{F1-score} = \frac{2TP}{2TP+FP+FN} \tag{2.8}$$

- **ROC Curve** (Equation 2.9): The data is represented by plotting the false positive rate (FPR) on the x-axis and the true positive rate (TPR) on the y-axis, where a steeper curve towards the y-axis is desired (Occhipinti et al., 2022).

$$\text{False Positive Rate} = \frac{FP}{FP+TN} \tag{2.9}$$

$$\text{True Positive Rate} = \frac{TP}{TP+FN} = \text{Recall} \tag{2.10}$$

2.4 Generative AI

In recent times, significant global attention and interest has been captured by Generative AI, notably driven by the release of Chat Generative Pre-trained Transformer (ChatGPT) in 2022 (Nah et al., 2023). Unlike traditional Artificial Intelligence (AI) models, generally utilized for prediction

and classification, Generative AI encompasses a variety of machine learning techniques capable of creating new content by identifying and learning from patterns found in existing data (Brynjolfsson et al., 2023). It can generate content in diverse forms, including image, audio, text, and video, empowering an ample scope of applications.

Generative AI technology is anticipated to significantly transform various aspects of society, including our working methods, learning, and communication (Nah et al., 2023). It is being disruptive globally across virtually every industry, and researchers suggest that industries requiring less critical thinking, fewer personal and effective relationships, and less creativity may be particularly susceptible to its impact (Nah et al., 2023). To maximize the opportunities and navigate the challenges of Generative AI in diverse domains, collaborative efforts are essential to integrate and enhance the capabilities of this technology (Nah et al., 2023). Some concerns about its effectiveness in real-world settings are often due to potential specific and unfamiliar problems for the model, organizational resistance, the risk of providing misleading information, and the possible spread of bias. Another aspect, directly associated with its recent emergence, is the lack of information about its long-term impacts on organizations (Brynjolfsson et al., 2023).

According to research done by Zhang and Gosline (2023), there are four main principles for human-AI collaboration: AI only, human only, augmented human, and augmented AI. Content creators might use GenAI's output as a reference before finalizing their decision, which is called the 'augmented human' paradigm. Alternatively, the 'augmented AI' paradigm occurs when users input their responses into the model for the AI to edit and finalize.

2.4.1 Chat Generative Pre-trained Transformer - ChatGPT

ChatGPT is a model based on Large Language Models (LLMs) and is tailored for conversational tasks that generate responses perceived as human-like by leveraging its extensive knowledge base. It belongs to the Generative Pre-trained Transformer (GPT) category, a class of LLMs that utilize deep learning techniques for extensive training with vast amounts of data (Nah et al., 2023).

LLMs can produce grammatically correct and relevant text, and their performance and quality depend on the computational resources required for training, the scale of model parameters, and the volume of the dataset (Brynjolfsson et al., 2023). Several advancements in LLMs led to significant improvements in model performance and expanded its capabilities. Examples are the extensive corpus of unlabeled data used to pre-train the models, thereby enabling them to be used in multiple applications, and the refinements of the model to improve its quality and make it more tailored for specific applications (Brynjolfsson et al., 2023). These models possess the ability to generate human-like text, enabling them to contribute significantly to collaborative efforts between humans and AI systems (Chen et al., 2024).

There are already some studies on the performance of ChatGPT compared to traditional machine learning techniques, although the literature is not yet extensive. A study conducted by Chen et al. (2024), for instance, assesses the performance of ChatGPT in classifying health advice sentences and identifying causal relationships in biomedical literature. Furthermore, Loukas et al. (2023) suggests leveraging conversational GPT models for efficient few-shot text classification

within the financial domain. The findings indicate that querying GPT-3.5 and GPT-4 can outperform fine-tuned, non-generative models even with fewer examples.

2.4.2 Prompting Techniques

The selection of appropriate prompts for in-context learning is crucial to the overall effectiveness of Large Language Models in specific tasks. The selection and design of appropriate prompts can tailor the model's responses, enhancing its effectiveness in specific applications. Various research efforts have introduced prompting strategies aimed at constructing more efficient prompts (Peikos et al., 2023). Examples of prompting techniques are prompt customization, prompt expansion, and prompt manipulation.

Prompt customization involves tailoring prompts to specific tasks or domains to guide the model in generating more relevant and accurate responses. Customization can include designing task-specific instructions that align with the particular requirements of a given task (Lester et al., 2021).

Prompt expansion provides additional context and examples to the prompts. Methods like few-shot and zero-shot prompting exemplify this approach. Few-shot prompting involves providing the model with a few input-output examples to induce an understanding of the task, which has been shown to enhance performance on complex tasks (Brown et al., 2020). Chain-of-thought (CoT) prompting expands prompts by including intermediate reasoning steps, guiding the model through a logical process to arrive at the correct answer (Wei et al., 2023). This technique has been particularly effective in improving performance on tasks requiring structured reasoning.

Prompt manipulation focuses on structuring the prompts to convey specific instructions and criteria, thereby influencing the model's outputs. Techniques such as self-consistency and chaining multiple prompts together are notable examples. Self-consistency generates diverse reasoning paths and selects the most consistent answer, enhancing the model's ability to solve complex reasoning tasks (Wang et al., 2022). Chaining prompts involves using the output of one prompt as the input for the next, which helps in breaking down complex problems into simpler sub-tasks, thus improving task outcomes.

The continuous evolution of prompting techniques underscores their critical role in optimizing LLM performance. By leveraging prompt customization, expansion, and manipulation, researchers can enhance the reasoning capabilities, accuracy, and applicability of LLMs across diverse contexts and tasks, thus unlocking their full potential.

2.5 Overview and Contributions

This literature review has provided an analysis of the current state of research in customer feedback analysis, the application of natural language processing, and the use of generative AI for enhancing decision-making processes. The review has highlighted several key themes and methodologies that form the foundation of this field.

Despite the progress in each of these areas, notable gaps remain. Firstly, there is a lack of comprehensive approaches that prioritize feedback based on risk and severity. Secondly, the potential of combining generative AI with traditional machine learning methods to enhance feedback analysis and prioritization remains untapped.

This project contributes to the literature by developing an approach that integrates advanced machine learning techniques with generative AI to prioritize customer feedback based on risk levels. This integrated approach aims to address the identified gaps by providing a more comprehensive system for feedback analysis and prioritization. The findings and methodologies presented in this work have the potential to significantly enhance the operational efficiency of customer service teams and improve overall customer satisfaction.

Chapter 3

Sonae MC Case Study

This chapter delves into the current state of customer feedback management at Sonae MC, highlighting the key challenges and areas for improvement. It is structured into three main sections. Section 2.1 outlines the customer feedback analysis process at Sonae MC. Section 2.2 defines the problem, pointing out the significant challenges in the existing process and establishing the criteria for what constitutes a "Severe-Risk Case". Finally, Section 2.3 presents the proposed approach to tackle these challenges, introducing methodologies for identifying and classifying severe-risk cases.

3.1 Problem Statement

3.1.1 Customer Feedback Analysis at MC

Sonae MC receives customer feedback through various communication channels, and the process of handling this feedback is divided into three main stages: 1) To Define, 2) To Act, and 3) To Learn. This dissertation focuses mainly on the first phase. These processes are managed by the Customer Service team, ensuring a structured and efficient approach to feedback management.

3.1.1.1 To Define

The first stage, "To Define", involves the initial processing and classification of customer feedback. Customers have multiple channels through which they can submit feedback, including web forms, emails, telephone calls, in-store complaint books, emails via company websites and partner portals, electronic complaint platforms, emails to the call center, face-to-face interactions, social media, letters to entities, mobile apps like Cartão Continente and Quico, Google reviews, letters, webchat, voicemails to the call center, tweets, and compliments books.

These channels vary in terms of the format of feedback submission, with some allowing only written text, and others involving phone calls or more structured forms. For instance, channels such as web forms and the Cartão Continente app require customers to fill out a series of predefined fields, which assists in typifying the complaint and provides a more structured feedback format.

When customers submit feedback, it is automatically categorized using a predefined classification tree built on machine learning algorithms. This classification tree is pre-established and tailored to the business needs. Examples of classifications include "Quality, product, and service" and "Store environment and customer service."

Furthermore, sentiment analysis is conducted to classify the sentiment expressed in the messages received from customers. Assessing the emotional tone of the feedback also assists the operator in determining the most appropriate approach to addressing the situation.

3.1.1.2 To Act and To Learn

The second stage, "To Act", aims to streamline the response process, and involves resolving customer issues based on the classified and analyzed feedback from the first phase. Each case is resolved differently and the response is customized. It can involve, for example, the exchange of emails, a refund, or even the intervention of external entities, depending on the nature of the case.

Finally, the third stage, "To Learn", focuses on extracting insights from the feedback data, which is also an extension of this dissertation focus.

3.1.2 Challenges in the Current Process

Despite the structured stages implemented, and the company's dedication to customer satisfaction, the current process of collecting, analyzing, and solving customer feedback at Sonae MC lacks a process to identify and classify risks inherent in the feedback data.

Currently, customer feedback is addressed on a first-come, first-served basis. As a result, feedback associated with severe-risk issues, which should be addressed more promptly, may have a response time that exceeds what is required by the nature and urgency of the case. This can result in an ineffective allocation of resources and inadequate attention to urgent issues that require immediate resolution, which may have a potential impact on the company's operations, reputation, and financial performance.

Another challenge encountered is that, although there are indeed cases of risk and they are recognized as such, their classification is not performed. This means that once analyzed, cases are not documented as being severe-risk or not. Consequently, this lack of classification leads to a situation where the severity of issues is not systematically recorded or prioritized. Therefore, there is a critical need to establish a pre-classification process.

Although sentiment analysis is performed to categorize feedback as positive, negative, or neutral based on customer expressions, which could give a sense of priority, focusing on this classification alone may not be optimal. This is because, for instance, feedback describing a situation in a neutral manner could contain a case equally or more severe than someone expressing their sentiments negatively. Sentiment analysis lacks the contextual nuance needed to accurately interpret feedback and overlooks its specific nature and implications, as a highly negative sentiment does not necessarily correlate with a case that is considered high risk. Relying solely on sentiment for prioritization may ignore critical risk factors, exposing the organization to potential harm.

3.1.3 Definition of 'Severe-Risk Case'

As severe-risk cases are already recognized as such, it is a priori evident that there should be specific criteria for a case to be considered severe-risk. This definition is crucial because it forms the foundation of the developed work and is closely linked to the business, specifically the Customer Service team, that typically does this recognition. It is then essential to have a detailed, clear, and objective description of what constitutes a severe-risk case that should be prioritized.

In collaboration with the Customer Service team, the following definition was established:

"A feedback case associated with a severe-risk case presents significant threats to the company's reputation, customer loyalty, and/or safety. These cases have the potential to cause substantial damage to the company's image or result in significant financial losses. They demand immediate action and thorough investigation to identify and, if necessary, resolve underlying causes. Such problems are considered priorities and require urgent responses to mitigate any further damage. Examples include data security incidents, dangerous or defective products posing health or safety risks to customers, and situations that significantly harm the company's reputation."

To summarize, the definition of a critical case rests on the severity of the following factors: reputation, customer loyalty, customer safety, financial stability, and any type of discrimination. These factors underscore the importance of prioritizing cases that pose substantial risks to the company in these critical areas.

3.2 Proposed Approach

To address the challenges in the current process, this dissertation proposes the development of methodologies for clear identification, and classification of feedback related to severe-risk cases that need prioritization. Although feedback is currently categorized using a typification tree, this classification only considers the nature of the complaint's problem, and identical typifications can represent complaints with varying degrees of associated risk.

Furthermore, there is a considerable volume of customer feedback, yet it is estimated that the number of cases representing significant risk is not substantial. Additionally, there is no current documentation of risk-associated cases, and the dataset is too extensive to be entirely manually pre-classified. As this pre-classification is essential to employ classification models, there's a paired challenge: 1) to identify and pre-classify cases based on the presence and degree of severe risk from historical data, and 2) to classify and prioritize risks in customer feedback on a real-time basis, based on the findings from the previous phase.

The first challenge encompasses topic identification through topic modeling and pre-classification of feedback. Topic modeling, specifically Latent Dirichlet Allocation model, was applied to the entire dataset which resulted in the identification of 15 distinct topics. LDA was employed to identify the main topics present in the feedback and reduce the dataset by removing topics typically less associated with severe-risk cases. Upon further analysis, 2 out of the 15 topics were excluded, thereby narrowing the focus to the remaining topics.

Despite this reduction, the dataset remained too extensive to be manually classified alone. Thus, the project leveraged generative AI methods, specifically the enterprise ChatGPT-4 API, for the initial identification of cases. While the primary interest is to identify whether cases are severe-risk or not, it was considered advantageous to have the output of this classification in three risk levels: high, medium, and low. The classifications obtained were then reviewed and verified by the Customer Service team. Cases classified as medium or low risk will not be considered severe cases and will therefore be analyzed on a first-come, first-served basis, while high-risk cases will be prioritized to ensure timely and appropriate responses.

Subsequently, for the second challenge, a supervised analysis was conducted to create a systematic tool for the classification and prioritization of cases. This involved training machine learning models on the pre-classified data to accurately predict the risk level of new customer feedback. The models were evaluated using metrics such as precision, recall, F1-score, and accuracy to ensure their effectiveness in identifying severe-risk cases. Several machine learning algorithms were explored, including logistic regression, naive bayes, random forest, support vector machines, and gradient-boosting classifiers. Each model was tested with different configurations of text vectorization (TF-IDF and CountVectorizer) and class balancing techniques (SMOTE and SMOTE with undersampling). The performance of each model was assessed based on the F1-score, given its ability to balance precision and recall.

3.2.1 Methodology

This dissertation adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, which provides a robust and well-defined framework for tackling data mining projects, ensuring a structured approach to understanding and solving business problems through data analysis.

The CRISP-DM methodology consists of six major phases, which are iterative and adaptive to the specific needs of the project.

The first phase, "Business Understanding", focuses on comprehending the business objectives and requirements. It involved understanding the specific challenges faced by Sonae MC in handling customer feedback, particularly in identifying and prioritizing severe-risk cases. Meetings and workshops with the Customer Support team were conducted to gather detailed insights and define what constitutes a 'Severe-Risk Case,' as well as to understand the types of risks most pertinent to the company.

In the second phase, "Data Understanding", the collected feedback data from various existing channels was examined to understand its structure, quality, and content. This included initial data exploration, assessment of data quality, and identifying the key features relevant to the risk classification.

The "Data Preparation" phase involved cleaning and preprocessing the feedback data to make it suitable for analysis. This included removing irrelevant information and transforming unstructured text data into a format that could be used for modeling. Techniques such as tokenization, stemming, and stop-word removal were applied to prepare the text data.

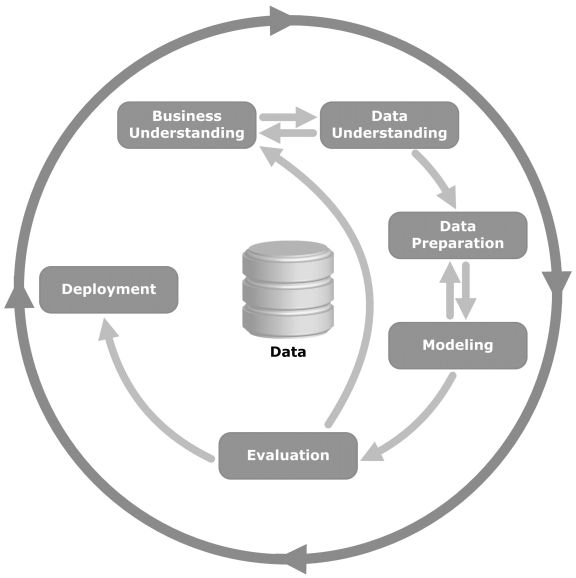

Figure 3.1: Phases of CRISP-DM process model

In the fourth phase, "Modeling", various data mining techniques were applied to the prepared data. Initially, as mentioned, unsupervised learning methods, particularly topic modeling, were used to identify the main topics present in the complaints and reduce the dataset. Subsequently, generative AI methods were used for the initial identification of severe-risk cases. Finally, supervised learning models were developed to systematically classify and prioritize feedback based on the risk level.

The "Evaluation" phase involved assessing the performance of the developed models to ensure they meet the business objectives and requirements. Various metrics, such as precision, recall, and F1-score, were used to evaluate the models' effectiveness in identifying and prioritizing severe-risk cases.

The final phase of the CRISP-DM process, "Deployment", involves deploying the developed models into the business environment. This phase ensures that the models are integrated into Sonae MC's customer feedback handling process, enabling the automated classification and prioritization of feedback.

Chapter 4

Topic Modeling

This chapter delves into the topic modeling phase of the project. Topic modeling is a powerful technique that automatically identifies and groups feedback into coherent topics, providing valuable insights into the underlying themes within the data.

In this project, the motivation for using topic modeling goes beyond identifying themes within customer feedback. It serves a strategic purpose in identifying topics that are typically less correlated with severe-risk cases, thus enabling the efficient reduction of the dataset from the beginning, given its extensive nature. Hence, this approach aims to support the subsequent analysis and assist in understanding the nuances of customer feedback.

This chapter outlines the methodologies employed in the topic modeling process and presents the key results obtained. The following sections will detail the data preprocessing steps, the selection and training of the topic modeling algorithm, and the interpretation and filtering of topics.

4.1 Methodologies

Sonae MC maintains documentation regarding customer feedback received through various communication channels, including web forms, emails, and in-store complaint books. The dataset of customer feedback compiled by the Customer Service team for this project consisted of 18 columns containing both structured and unstructured data for each case, including case number, subject, description of the feedback, and respective categorizations. The dataset is summarized in Table 4.1.

Given the extensive nature of customer feedback data and the estimated low frequency of severe-risk cases, the project narrows its focus to a subset of the feedback: customer complaints. This decision is based on the collective experience of the Customer Service team, which indicates that past instances of severe risk are typically more associated with customers who submit complaints. The dataset contained 98,064 customer complaints spanning the years 2020 to 2024.

Table 4.1: Variables in the customer feedback dataset

Column	Description
Número do Caso	Unique identifier for each feedback case
Tipo	Type of feedback case
Data de abertura	Opening date of the feedback case
Assunto	Brief description of the feedback
Descrição	Detailed description of the feedback
Tipificação 1	Primary categorization of feedback type
Tipificação 2	Secondary categorization of feedback type
Tipificação 3	Tertiary categorization of feedback type
Tipificação 4	Quaternary categorization of feedback type
Origem do caso	Source of the feedback case
Meio	Channel through which the feedback was received
Loja de Ocorrência ou Transação	Store where the incident or transaction occurred
Insígnia de Ocorrência	Brand associated with the incident
Motivo do caso	Reason for the feedback case
Motivo Cliente	Customer's reason for the feedback
Entidade Externa	External entity involved, if any
Nº caso na entidade Externa	Case number at the external entity
Solicitada indeminização/comparticipação	Request for compensation or reimbursement

4.1.1 Data Preprocessing

After carefully analyzing the dataset, it was concluded that the '*assunto*' (subject) and '*descrição*' (description) columns were particularly relevant for analyzing customer complaints since both consist of free-text entries, allowing customers to express their concerns in their own words. The '*assunto*' column briefly describes the complaint's subject, while the '*descrição*' column provides a detailed explanation of the case. These columns were selected for topic modeling analysis because they are the only free-text fields in the dataset that are directly filled by the customers, thus capturing the most authentic and detailed representation of the customers' experiences and concerns. As such, the topic modeling analysis will focus on extracting topics from these columns. Additionally, only cases related to the '*Insígnia de Ocorrência*' *Continente* and *Continente Online*, which represent 95% of the dataset, were retained for the analysis.

To ensure the quality of the dataset, 7,797 cases with empty descriptions were removed, as they were deemed incomplete for detailed analysis and thus excluded from the dataset. Subsequently, cases with fewer than five words in the description were removed from the dataset. This decision was based on the understanding that short descriptions often lack the necessary context and detail to provide meaningful information and are very unlikely to represent severe-risk cases . For example, phrases like "bad service" or "too expensive" do not provide enough specific details, making it difficult to comprehend the nuances of the complaint or to extract relevant topics effectively. Short descriptions can then introduce noise into the topic modeling technique, resulting in poor-quality topics that do not accurately reflect the themes present in the customer feedback data.

Following these steps, the dataset was reduced to 65,853 cases.

Finally, the '*assunto*' and '*descrição*' columns were concatenated to enhance the richness of contextual data, as the result of these columns capture both the concise context and the detailed case, increasing the amount of information available.

4.1.1.1 Text Preprocessing

Once the dataset has been prepared, the next step is to preprocess the text to clean and prepare the concatenated column for analysis. The following preprocessing steps were applied to the customer feedback data to enhance its quality and relevance for subsequent analysis. The preprocessing was implemented using libraries such as 'unicode', 'snowballstemmer' and 'stanza'.

First, newline characters ("\n") were removed from the text, which appeared when paragraphs were present in the feedback. Further, all the text was converted to lowercase to ensure uniformity and avoid duplicating words that differ only in case. Non-alphabetic text, numbers, URLs, email addresses, accents, and punctuation were removed to clean the text of extra elements that do not contribute to meaningful analysis. Common greetings, such as "*boa tarde*" (good afternoon) and "*boa noite*" (good night) were also removed to eliminate text that is not relevant to the content of the feedback.

In concordance, stopwords, common words that do not have significant meaning, were removed to reduce noise in the data. Both Portuguese and English stopwords were removed and the stopword dictionary was updated to include specific words such as "*continente*", "*obrigado*" (thank you) and "*cumprimentos*" (regards). The word "*cliente*" (customer) and its variations were also removed from the dataset, to ensure the focus on more informative terms. This business specific word was highly frequent, appearing in 40% of the cases.

Stemming and lemmatization techniques were applied to help standardize words to their base or root form. Stemming involves trimming words to their root form (e.g., "arguing" becomes "argu"), while lemmatization involves using a vocabulary and morphological analysis to return the base or dictionary form of a word (e.g., "studies" becomes "study"). These processes are important to help to reduce the number of unique words in the dataset and to improve the consistency of the data (Hickman et al., 2022). Both processes were applied separately to determine which yielded better results for the analysis. The effectiveness of each method was evaluated by running the model and comparing the coherence scores and the top words within each topic. The results indicated that the coherence scores and the top words for each topic were very similar for both stemming and lemmatization. Since the model output was essentially the same for both methods in terms of topic quality and interpretability, stemming was chosen as it is computationally faster and still efficient.

By completing these preprocessing steps, the text data was standardized and cleaned, making it suitable for the subsequent topic modeling analysis. An example of the text before and after preprocessing is present in Table 4.2.

Table 4.2: Example of text before and after preprocessing

Original Text in '*Assunto*'	**Original Text in '*Descrição*'**	**Processed Text**
Falta de qualidade	*Comprei um sumo da vossa marca tropical cenoura a data 27/10/21 que consta na tampa da garrafa diz que já está fora de prazo, quero explicações da vossa parte como isto pode acontecer. Obrigado*	*falta qualidade comprei sumo vossa marca tropical cenoura data consta tampa garrafa diz prazo quero explicacoes vossa parte pode acontecer*

4.1.2 Latent Dirichlet Allocation

The technique chosen for topic modeling in this project was Latent Dirichlet Allocation. This choice was driven by its proven effectiveness, interpretability, and robustness. It provides interpretable results by assigning a probability distribution over words for each topic, allowing for the identification of key terms and themes associated with each topic. This interpretability is essential for understanding and categorizing the diverse range of customer complaints present in the dataset. Additionally, LDA accommodates the possibility that each case may contain multiple topics, reflecting the complexity and multifaceted nature of customer feedback (Alghamdi and Alfalqi, 2015).

Latent Dirichlet Allocation is one of the earliest and the most frequently utilized Topic Modeling methods, with a wide range of applications across a vast number of objectives. LDA is a probabilistic model based on statistical (Bayesian) topic models and its output is the topics and their related words (Alghamdi and Alfalqi, 2015).

LDA considers that each corpus is a composition of a predefined number of topics and each document within the corpus mixes these topics in varying proportions. It also assumes that each topic is a probability distribution of words that specifies the likelihood of a given word occurring within a specific topic. The process involves learning the distribution of topics, which is treated as a hidden variable that needs to be inferred through statistical methods (Blei et al., 2003).

The model presumes that the number of topics is fixed and known, and the way to best determine that value is not well defined for the major part of Topic Modeling methods, inclusive for LDA. Researchers often generate topic models for the same data using a different number of topics and then analyze certain metrics, such as coherence and perplexity (Blei et al., 2003).

Another assumption of this method is that documents are independent, which is inherent to the Dirichlet probability, and that words within documents are also independent, which is related to the BoW assumption. This means that this generative model does not consider the sequential order of words in the generation of documents. Figure 4.1 illustrates the graphic representation of the model.

In this model, each document D is a composition of K latent topics, and each describes a distribution over W word vocabulary. The variable θ represents the topic distribution for a document

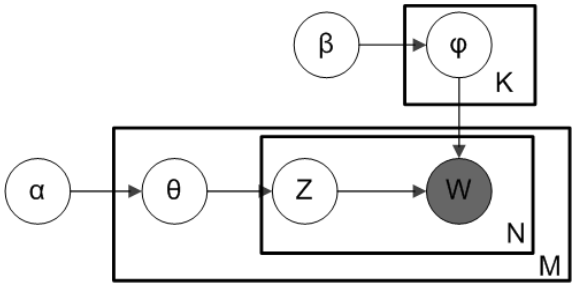

Figure 4.1: Graphical representation of LDA

m, α is the parameter of the Dirichlet prior on the per-document topic distributions, β represents the Dirichlet prior parameter of the word distribution in a topic, z_{mn} refers to the topic for the n^{th} word in document m, w_{mn} refers to the specific word, and ϕ refers to the word distribution for topic k (Blei et al., 2003).

LDA is the most used topic modeling approach since it effectively generates topics by disambiguating words and precisely aligning keywords to topics, closely reflecting the original text collection. This capability makes it particularly suitable for analyzing large volumes of textual data, such as customer feedback, where capturing the underlying themes and issues is crucial for strategic decision-making.

4.1.3 Implementation

4.1.3.1 Training the model

The process began with tokenizing the text data, breaking down each document into smaller units or tokens. This step is essential for enabling the model to accurately analyze and learn from the text data. The tokenized text data is then converted into a dictionary, mapping each unique token to a specific identifier. This dictionary is crucial for converting the text data into a numerical format that the LDA model can process.

Using the dictionary, a "Bag of Words" representation was generated for each document, which lists the frequency of each token within the document. This simplified representation helps the model understand the importance of each word based on its frequency of occurrence.

An important step in the implementation of LDA is choosing the number of topics for the model. Selecting too many topics can lead to overfitting while selecting too few topics can lead to underfitting. To ensure the model was optimized for this project's dataset, a grid search was conducted, specifically varying the number of topics and the number of passes. The number of topics ranged from 5 to 25, incrementing by 2, and the number of passes was set to 10, 15, and 20. The number of topics determines how many distinct topics the model should identify, while the number of passes refers to how many times the model iterates over the entire corpus during training. This approach allows for a systematic exploration of different configurations to find the most effective setup. The model was trained using the 'gensim' library.

4.1.3.2 Evaluation

For each combination of parameters, the LDA model was trained and its performance was evaluated using coherence and perplexity scores. Coherence scores measure the interpretability of the topics generated, indicating how well the words within each topic cohere semantically. Higher coherence scores indicate more interpretable and coherent topics, which are easier to understand and analyze. Perplexity scores, on the other hand, measure the model's ability to predict the data, indicating how well the model fits the data. Lower perplexity indicates better generalization performance.

While perplexity is a common metric for evaluating topic models, it has been observed that perplexity does not always correlate with human interpretability. Coherence, on the other hand, directly addresses the interpretability of the topics, which is crucial for practical applications such as understanding and categorizing customer feedback (Krishnan, 2023).

Given the project's goal of identifying and prioritizing severe-risk cases in customer feedback, it was crucial to ensure that the generated topics were statistically robust but also meaningful and interpretable. Therefore, although both coherence and perplexity were measured, it was decided to prioritize coherence as the primary criterion for model selection.

4.1.3.3 Visualization

Effective visualization is crucial for interpreting and understanding the topics generated by the LDA model. Visualizing the topics helps in identifying key themes, patterns, and trends within the customer feedback data, facilitating actionable insights.

PyLDAvis is an interactive LDA visualization tool that combines several visual elements to provide a comprehensive view of the topic model. The model includes intertopic distance maps, topic-specific term frequency and term saliency and relevance. In other words, it visually maps the distances between topics, displays the most relevant terms along with their frequency and relevance and shows the most informative terms for each topic, balancing both frequency and exclusivity. With this framework it is possible to explore the topics, zoom in on specific terms, and adjust parameters to better understand the underlying structure of the customer feedback data. It not only improves the understanding of the data but also allows for a more efficient and focused analysis, highlighting important areas for further investigation. The ability to adjust term relevance and visualize relationships between topics provides an analytical depth that is essential for robust and detailed qualitative analysis.

4.2 Results and Discussion

A critical step in obtaining effective results with LDA is determining the optimal number of topics. As a result of the grid search conducted to identify the ideal number of topics, the model with the highest coherence score was selected. Coherence scores are crucial as they measure the semantic similarity between words within a topic, thereby ensuring the interpretability and meaningfulness

of the topics generated. Figure 4.2 illustrates the variation in coherence scores across the different number of topics tested. The graph demonstrates how coherence scores change with the number of topics, highlighting the optimal point where the interpretability of the results is maximized.

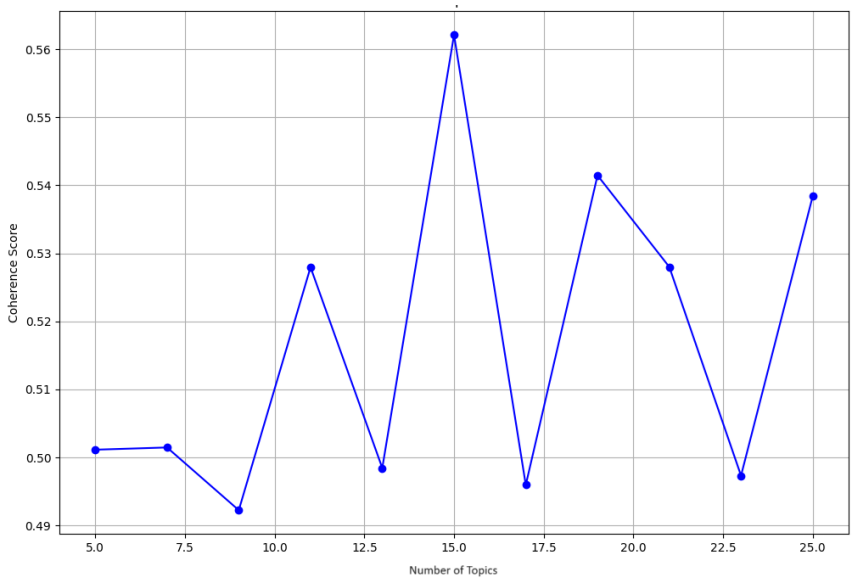

Figure 4.2: Variation in coherence scores

From this graph, it can be inferred that the optimal number of topics is 15, which maximizes the interpretability of the results. The maximum coherence score obtained is 0.562, indicating the degree of semantic consistency within the topics.

Once the optimal number of topics was determined, the LDA model was implemented using these parameters. Table 4.3 illustrates the top 10 words associated with each topic, along with their corresponding coherence scores. These coherence scores were derived from an evaluation of how well the words within each topic cohere semantically, providing insight into the quality and interpretability of the topics generated by the LDA model.

Topics with higher coherence scores, such as topics 2, 4, and 6, are likely to represent more coherent and interpretable themes within the customer feedback data.

Topics 1, 3, 7, 9, 10, 12, 13, and 14 have moderate coherence scores ranging from 0.4299 to 0.5908. These scores suggest that while the topics are relatively coherent, they may include a broader range of terms or themes that are slightly less distinct. For example, topic 1 includes terms like "*produt*" (product), "*emb*" (packaging), and "*encomend*" (order), indicating themes related to product orders and packaging issues.

Topics 5, 8, and 15 have the lowest coherence scores, with topic 5 at 0.2667 being the lowest. These low coherence scores suggest that the words within these topics are less semantically related, making the topics less interpretable. Topic 5 includes terms such as "*maquin*" (machine), "*cupo*" (coupon), "*caf*" (coffee), and "*capsul*" (capsule), which might relate to issues with specific products like coffee machines and capsules.

Table 4.3: Top 10 words per topic with coherence scores

Topic	Top 10 Words	Coherence Score
1	*un, produt, emb, gr, encomend, kg, batat, iogurt, leit, envi*	0.5094
2	*reclamaca, receb, nov, dad, reserv, queix, mor, direit, marc, nom*	0.8273
3	*situaca, cas, respost, seguiment, reclamaca, poss, resolv, envi, process, pt*	0.5208
4	*prec, compr, descont, valor, sel, campanh, dia, carta, fiz, loj*	0.8215
5	*maquin, cupo, caf, capsul, receb, reclam, col, agu, rebent, beb*	0.2667
6	*reclamaca, livr, reclamaco, praz, dias, equip, dev, envi, submet, term*	0.9087
7	*voss, parqu, estacion, apresentaca, tod, carrinh, melhor, arrumaca, carr, empres*	0.4299
8	*produt, voss, loj, compr, vez, gost, pod, dest, faz, marc*	0.5285
9	*devoluca, qualidad, motiv, artig, funcion, part, falt, reclam, pec, deix*	0.5288
10	*encomend, entreg, reclamaca, dia, receb, contact, onlin, inform, artig, indic*	0.5748
11	*carta, compr, cupa, valor, dia, sald, loj, descont, credit, tala*	0.4755
12	*fatur, via, fs, valor, solicit, envi, cobr, nif, factur, marcaca*	0.5908
13	*contact, envi, favor, pod, email, ajud, joa, consegu, catarin, mar*	0.5420
14	*atend, caix, loj, reclam, colabor, funcion, esper, faz, ped, reclamaca*	0.5871
15	*compr, sabor, embalag, bolor, artig, produt, cheir, alter, reclam, alteraca*	0.4197

Furthermore, understanding the distribution of complaints across these themes also provides important insights. Figure 4.3 shows the distribution of complaints across its dominant themes.

Knowing that a significant proportion of complaints revolves around a specific issue indicates a substantial customer pain point in this area. Notably, topics 9, 15, and 14 have the highest frequencies, indicating these themes are the most common among complaints. These topics generally indicate issues involving product returns, quality concerns, service wait times, and employee responsiveness.

However, it is crucial to differentiate between the frequency of complaints and the severity of issues. For instance, frequent complaints about minor service inconveniences, such as delays in promotional offers or minor discrepancies in card balances, may be indicative of systemic issues that affect a large number of customers but do not pose significant risks to the organization's reputation or financial stability. Therefore, while it is important to analyze the most prevalent issues, there is not a direct correlation between the frequency of complaints and their severity or risk level.

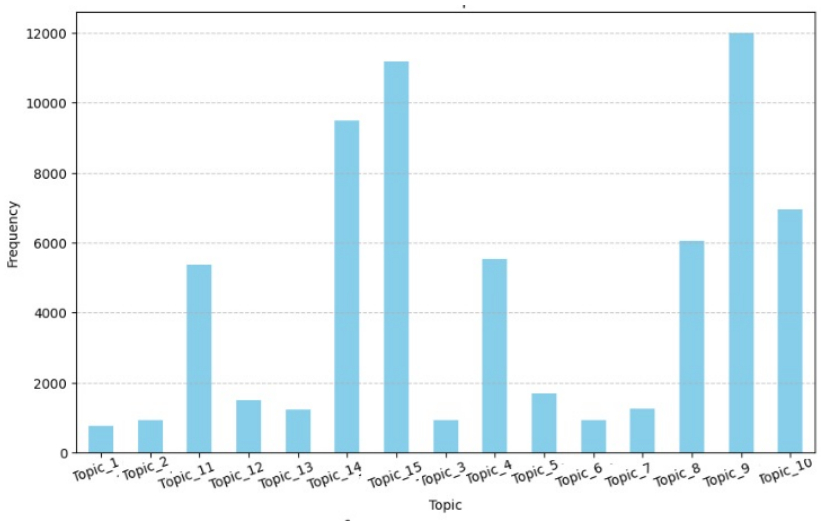

Figure 4.3: Distribution of complaints across themes

4.2.1 Visualization

The pyLDAvis tool, which is used to visualize the topics generated by the Latent Dirichlet Allocation model was useful for qualitative data analysis, allowing a clear and intuitive visualization of the topics generated. This specific visualization helps in understanding the structure of topics, their prevalence, and the most salient terms associated with each topic. Each bubble represents a topic and its size is proportional to the prevalence of the topic across the corpus. The distance between the centers of the bubbles indicates how similar or dissimilar the topics are. The concept of similarity or dissimilarity in this context refers to the degree of overlap or disparity in the word distributions between topics. Figure 4.4 shows a screenshot of the PyLDAvis interface.

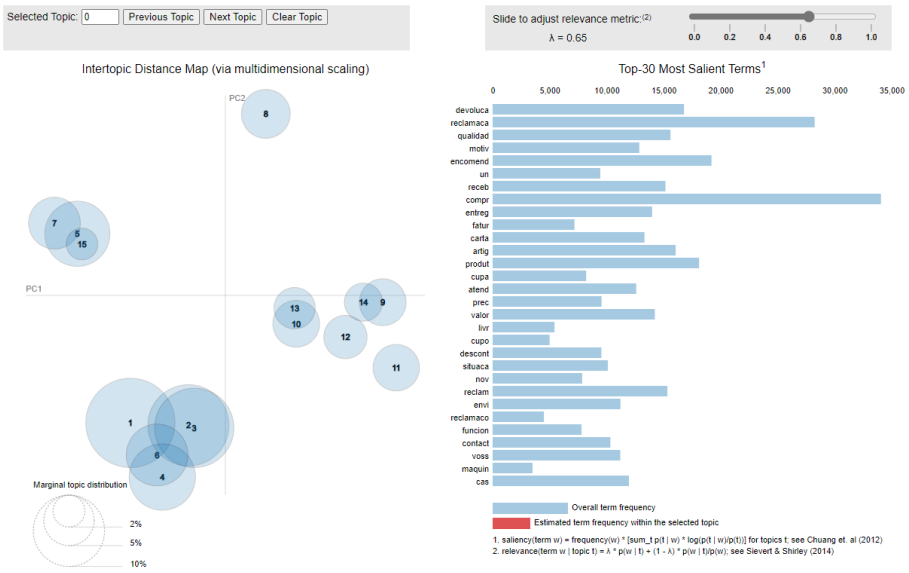

Figure 4.4: Interactive visualization with PyLDAvis

Topics that are closer together, such as topics 1, 2, 3, 4, and 6 show a significant degree of

similarity in their word distributions. This clustering suggests that these topics share common themes or issues and may overlap in the concerns they represent.

Topic 8 is isolated in the top-right quadrant, indicating it is quite distinct from the other topics in terms of its word distribution. This uniqueness suggests that topic 8 addresses a specific set of issues or concerns that are not significantly related to other topics.

The right side of the visualization shows the top 30 most salient terms across all topics. The most salient terms in the corpus include "*devoluca*" (return), "*reclamaca*" (complaint), "*qualidad*" (quality), "*motiv*" (reason), and "*encomend*" (order), among others. These terms appear frequently across multiple topics and are also quite distinctive to certain topics.

4.2.2 Topics Selection

A single case can often relate to multiple themes, reflecting the multifaceted nature of customer complaints. Thus, the probability of each case being associated with various topics was determined. For each piece of feedback, the LDA model assigns probabilities that indicate how likely it is for the feedback to belong to each of the identified topics.

For further analysis, it was considered only the topic with the highest probability for each case. This ensures the understanding of the primary concern of each complaint from the customer's perspective. Focusing on multiple themes for each piece of feedback could dilute the focus on the most critical issue while focusing on the topic with the highest probability helps to prioritize the most significant issue highlighted by the customer. Moreover, considering multiple topics for each case could lead to overlapping categories, which might obscure important insights.

The identified topics were subjected to meticulous analysis and discussion, involving close collaboration with the customer service department. Through this collaborative effort, it was recognized that, despite the specificity of certain topics, there remained a possibility for severe cases to emerge across all topic categories. Consequently, a more defensive and conservative approach was adopted to ensure that the analysis remained robust and comprehensive in identifying potential risks, even within topics that may not intuitively suggest high severity. As a result, two topics, namely topic 9 and topic 11, were deemed less pertinent and subsequently removed from further consideration. It is essential to note that these issues are indeed important, but they do not typically necessitate the same level of urgent intervention as more severe risk-related cases.

The decision to exclude topic 9, centered on product returns and quality issues, was based on several factors. Firstly, while issues regarding product returns and quality are undoubtedly important, they often do not represent immediate or severe-risk concerns that justify prioritization over other critical issues. Additionally, these types of complaints are typically handled through existing quality control and product management processes rather than requiring urgent intervention from risk prioritization methodologies. It is also important to note that while severe risks related to product quality can occur, in such cases, the dominant topic would likely not be this one. Instead, these severe-risk cases would be more prominently associated with topics specifically related to critical concerns.

Similarly, Topic 11, centered around coupons and card balance inquiries, was deemed less critical in the context of risk identification and prioritization. While these topics may generate a significant volume of feedback, they are often routine inquiries related to loyalty programs or promotional activities. Standardized processes are already in place to address these issues effectively, ensuring that customer concerns are promptly and appropriately handled. Consequently, these inquiries do not typically indicate potential risks to the company's reputation or financial stability.

The decision to exclude topics 9 and 11 was part of a broader strategy aimed at refining the scope of the analysis to focus on topics that could have a higher propensity for presenting severe cases. This refinement was necessitated by the significant limitation of not being able to manually validate all cases due to the extensive volume of data. While the remaining topics exhibited varying degrees of relevance to customer feedback, they were also subject to considerations regarding their potential correlation with high-severity cases.

As a result of this refinement, the dataset was reduced to 48,471 cases, representing a 26,4% reduction from the original cases.

Chapter 5

Pre-Classification of Feedback

This chapter focuses on the development and implementation of an automated pre-classification system using the enterprise ChatGPT-4 API. This chapter covers the methodologies, including the objectives and constraints, data selection, prompt design, and implementation, and the results and discussion of the classification accuracy and effectiveness.

5.1 Methodologies

5.1.1 Objectives and Constraints

After reducing the dataset through topic modeling, it was ideally envisioned that the remaining data would be manually classified by the specialized customer service team to inform the classification model's implementation. However, this manual classification was not feasible due to the extensive dataset volume and the project's time constraints. Consequently, the enterprise ChatGPT API was utilized to significantly reduce the time required for pre-classification.

The primary objective was to input the resulting cases from the topic modeling into the ChatGPT API along with the prompt and obtain the cases in order of severity. This approach would allow to define from which point onward a case would no longer be considered severe from the company's perspective. Ideally, all cases would be sent to ensure a relative classification.

However, there were inherent limitations to this practice. The ChatGPT API has a token limit of 128 thousand tokens, which represents the maximum number of tokens that can be processed in a single interaction, including both the input (prompt and/or conversation history) and the output (model response). This limitation significantly affects the planning of interactions with the model, especially in cases of long texts. Consequently, it becomes impractical to analyze extensive documents without either dividing them into smaller segments or potentially losing content. For example, when attempting to input approximately 400 cases, the API is unable to respond and return the desired classification.

5.1.2 Data Selection

Considering the time required for the verification and validation steps, it was decided to reduce the number of cases analyzed in the subsequent stages to accelerate the process and ensure feasibility within the project timeline. Consequently, three out of the thirteen topics resulting from the topic modeling were carefully selected, with the support of the customer service team, as they were deemed to have a higher propensity to include cases of severe risk due to the nature of the topic and the cases contained within it. Topics 3, 8, and 13 were selected, and including human expertise was crucial to ensure the accuracy and reliability of the selection. The reduced number of topics allows the team to further review and correct any potential errors, thereby maintaining high standards of quality and precision in the classification process. Table 5.1 illustrates the top 10 words associated with each selected topic.

Table 5.1: Top 10 words of the selected topics

Topic	**Top 10 Words**
3	situaca, cas, respost, seguiment, reclamaca, poss, resolv, envi, process, pt
8	produt, voss, loj, compr, vez, gost, pod, dest, faz, marc
13	contact, envi, favor, pod, email, ajud, joa, consegu, mar

It is important to note that the selection of topics 3, 8, and 13 for this phase does not imply that only these topics will be analyzed and prioritized in the future. The current focus on these topics was a strategic decision to streamline the process and validate the approach. As the system evolves, the analysis can be expanded to include additional topics, ensuring a comprehensive approach to risk assessment and feedback management.

The selection of these topics was based on several important considerations. These topics had moderate coherence scores, meaning that they were neither too specific nor too broad. Moderate coherence scores suggest that the topics include a broad range of terms and themes, making them less likely to be overly narrow or overly dispersed. This balance ensures that the topics are comprehensive enough to capture a variety of issues while still being interpretable and relevant. Moreover, the nature of the complaints within these topics typically involved issues with significant potential impacts on customer satisfaction and company reputation.

Historical data and the experience of the customer service team played a crucial role in the selection process. This context ensured that the selected topics were not only relevant but also aligned with the types of issues that had previously posed significant risks.

As a result, 8,618 cases were pre-classified at this stage, which corresponds to 17,78% of the current dataset.

5.1.2.1 Prompting

Prompting refers to the input of text provided to the ChatGPT API to generate a response, which in this case will be the classification. In the context of this specific stage, the purpose of prompting is to provide the API with the cases of customer feedback to be classified, the definition of severe-risk cases, and the request for classification.

The design of the prompt was critical to ensure that the model received sufficient context for accurate classification without exceeding the token limit. Balancing the complexity and length of the prompt with the need to provide detailed instructions to ChatGPT required careful consideration and iterative refinement. The prompting used in this model is shown in Figure 5.1.

```
prompt = """
Analise os casos de feedback de clientes e determine se cada caso é considerado de risco severo elevado, médio ou baixo. Use as definições de risco e os exemplos
temáticos fornecidos abaixo para fazer essa determinação.

Definição Geral de Risco Severo Elevado:
"Risco Severo elevado: Este tipo de reclamação representa ameaças significativas à reputação da empresa ou à lealdade e/ou segurança do cliente e têm o potencial
de causar danos significativos à empresa ou resultar em perdas financeiras substanciais. Exigem uma ação imediata e uma investigação aprofundada para esclarecer o
sucedido e, se necessário, resolver as causas subjacentes. São problemas considerados prioritários e requerem uma resposta urgente para mitigar quaisquer danos
adicionais."

Exemplos de Temáticas que Podem Indicar Casos de Risco Severo elevado:
"Exemplos incluem incidentes de segurança de dados, produtos perigosos ou defeituosos que representam um risco para a saúde ou segurança dos clientes, e situações
que resultam em danos significativos à reputação da empresa, tais como assédio, xenofobia e homofobia. Nem todos os casos com temáticas presentes nos exemplos são
de risco severo, devendo ser prioritizados os que representam dano efetivo e não apenas dano potencial."


"Risco Baixo: As reclamações ou casos com baixo risco têm um potencial mínimo para prejudicar a reputação da empresa, a lealdade do cliente ou a segurança. Estes
casos podem ter um impacto limitado na imagem da empresa ou em perdas financeiras. Podem não exigir ação imediata ou investigação intensiva. Exemplos de casos de
baixo risco podem incluir reclamações menores de clientes sobre problemas de produto que podem ser facilmente resolvidos sem repercussões significativas.

Risco Médio: Os casos de risco médio representam um nível moderado de ameaça à reputação da empresa, à lealdade do cliente ou à segurança. Embora possam não ter um
impacto imediato e severo, ainda têm o potencial de causar danos perceptíveis se não forem abordados. Estes casos requerem algum nível de ação e investigação para
identificar as causas subjacentes e mitigar os riscos. Exemplos de casos de risco médio podem incluir recalls de produtos devido a problemas de qualidade, ou
reclamações de clientes sobre a qualidade do serviço.

Risco Elevado: Os casos de elevado risco representam ameaças significativas à reputação da empresa, à lealdade do cliente ou à segurança. Têm o potencial de causar
danos substanciais à imagem da empresa e resultar em perdas financeiras significativas se não forem abordados prontamente e eficazmente. Os casos de elevado risco
exigem ação imediata, investigação minuciosa e respostas urgentes para mitigar quaisquer danos adicionais. Exemplos de casos de elevado risco incluem incidentes
graves de segurança de dados, produtos perigosos ou defeituosos que representam riscos significativos para a saúde ou segurança dos clientes, ou situações que
poderiam prejudicar significativamente a reputação da empresa, como publicidade negativa generalizada ou questões legais."


A resposta deve estar no formato JSON, com os campos 'nr_caso' (com o respectivo número) e 'risco_severo' (com 'elevado' ou 'médio' ou 'baixo')."""
```

Figure 5.1: Model prompting

The prompt provides detailed instructions and examples for the task at hand, which is to analyze customer feedback cases and determine their level of severity (high, medium, or low risk). The prompt includes definitions of each risk level, thematic examples, and clear guidelines for how to make the determination. The prompt emphasizes the need for a JSON response format with the fields '*nr_caso*' (case number) and '*risco_severo*' (severity level: 'high', 'medium', or 'low'). This prompt aims to provide comprehensive guidance to the model, ensuring that it understands the task and context well enough to generate accurate responses. It combines elements of prompt customization (tailoring the method to the specific task), prompt expansion (providing additional context and examples), and prompt manipulation (structuring the prompt to convey specific instructions and criteria).

The refinement of the prompt in the pre-classification of customer feedback using the ChatGPT API involved an iterative approach. Initially, the prompt was crafted based on general guidelines and examples. However, as the model's responses were analyzed, several iterations were made to

improve clarity and specificity. For instance, the definitions of risk levels were refined to include more detailed examples and contextual information. Additionally, instructions on how to format the response in JSON were emphasized to ensure consistent output.

5.1.3 Implementation

The ChatGPT-4 API was used for classification purposes, as it can handle large volumes of data and understand the context and nuance of textual information. Both ChatGPT-3.5 and ChatGPT-4 were available for use; however, ChatGPT-4 was chosen due to its significant improvements in model performance and contextual understanding. These enhancements are crucial for accurately identifying and classifying cases based on severity levels. This version of the model has been trained on a more extensive and diverse dataset, enhancing its ability to generate relevant and accurate responses in various contexts.

Before applying the model to the entire dataset, a proof of concept was conducted. Twenty cases were manually classified and entered into the API. This step was carried out to validate the accuracy of the model given the prompting, before applying it to the full dataset. These cases were entered into the API 15 times to evaluate the consistency and reliability of the results. It allowed for the verification of possible variations or discrepancies in the classification generated by the model. Overall, repeating the process aims to increase confidence in the decision-making process.

Given the token limitation, it was impractical to input large batches of cases in a single interaction. Next to the proof of concept, and to overcome the limitations, all the complaints were processed in smaller batches of 50 cases chosen randomly, to ensure diversity and that each group stayed within the token limit.

Initially, the cases were introduced to the model as unstructured text strings. However, this approach proved disadvantageous since the model struggled to accurately interpret the unstructured text data. The primary issue was the model's difficulty in identifying the boundaries between individual cases, leading to inaccuracies in classification. Consequently, the cases were reformatted to JSON, with the fields '*nr_caso*' (case number) and '*descrição*' (description). The structured format ensures that the data is organized and clearly delineated, making it easier for the model to accurately process and classify each case. An example of a case in JSON format is shown below:

Listing 5.1: Example of Case in JSON Format

```
{
    "nr_caso": "123456",
    "descricao": "bom dia venho remeter foto de produto com bolor quando
        abri esta unidade detetei esta situacao pelo que optei nao consumir
         o mesmo nem os restantes que compoe a embalagem o que podera estar
         na origem desta situacao  desde que o adquiri o mesmo foi
        conservado corretamente no frigorifico referente a vossa fatura
        aguardo esclarecimentos"
}
```

These JSON-formatted batches were submitted to the ChatGPT-4 API along with the prompt, enabling the retrieval of responses. The resulting responses were subsequently saved to a text file, thereby automating the interaction process with the model and streamlining the collection of responses for further analysis and processing.

Since the data was processed in batches, a relative ranking of all cases by severity was not possible, as the model could not access all cases simultaneously to make a comprehensive comparison. Therefore, it was decided to request the classification of cases into three risk levels: '*baixo*' (low), '*médio*' (medium), and '*elevado*' (high). This approach assists in maintaining a standardized classification system, ensuring consistency. Furthermore, it can simplify the validation process, as it is more advantageous for the customer service team to review and confirm classifications when they are grouped into a limited number of categories.

To further refine the classification accuracy and ensure that no high-risk cases were overlooked, the feedback initially classified as medium risk was reintroduced into the ChatGPT-4 model for re-evaluation. This step aimed to identify if any of these medium risk cases could be reclassified as high risk upon a second review by the model. The same prompting method was used for this step, reinforcing consistency in the model's decision-making process.

While the primary risk levels of 'high,' 'medium,' and 'low' remain the core classifications, the 'medium-high' sublevel provides a nuanced approach to ensure no critical cases are overlooked. This approach helps to catch any borderline cases that might have been underestimated in terms of severity during the initial classification, thereby enhancing the reliability and robustness of the classification system.

The reclassified cases were then reviewed by the customer service team, integrating a "human-in-the-loop" approach. This ensures that the AI model's decisions are validated by human expertise, adding an additional layer of exploration and accuracy.

5.2 Results and Discussion

The pre-classification process using the ChatGPT-4 API resulted in the categorization of 8,618 customer complaint cases into four risk levels: '*elevado*' (high), '*médio*' (medium), '*médio-alto*' (meidum-high) and '*baixo*' (low).

To assess the accuracy of the ChatGPT-4 model, a proof of concept was conducted, as described in the Methodology section. Twenty cases, spanning all risk levels, were manually classified and input into the API as a proof of concept. This process was repeated 15 times to assess the consistency and reliability of the model's classifications. The model's classifications were compared against the manual classifications to calculate the accuracy rate. The results are summarized in Table 5.2.

The mean accuracy of the classifications is calculated to be 86.20% with half of the cases having an accuracy of at least 86%. This shows a consistent performance across the majority of the cases. The data demonstrates that the ChatGPT-4 API model performs reliably in pre-classifying

Table 5.2: Accuracy of model classifications

Case Number	**Accuracy (%)**
Case 1	100%
Case 2	100%
Case 3	100%
Case 4	73%
Case 5	100%
Case 6	100%
Case 7	100%
Case 8	100%
Case 9	86%
Case 10	80%
Case 11	100%
Case 12	100%
Case 13	86%
Case 14	73%
Case 15	100%
Case 16	46%
Case 17	73%
Case 18	53%
Case 19	80%
Case 20	86%

customer feedback into risk levels. The high mean and median accuracy values indicate that the model is effective in most scenarios.

Following this initial accuracy assessment, all 8,618 cases were classified using the model. The results, as illustrated in Figure 5.2 reveal that a substantial portion of the feedback was classified as low risk. This predominance of low-risk cases aligns with the expectations set by the business insights, as it is generally anticipated that the majority of customer feedback will not involve severe issues. From the company's experience, high-risk cases are expected to be fewer in number, and the results obtained in this study confirm these expectations.

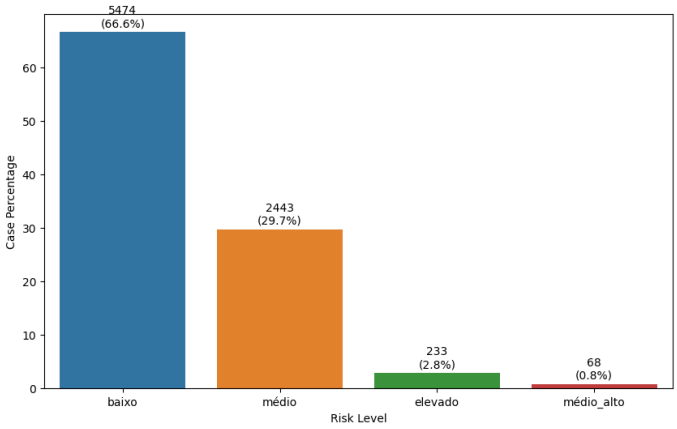

Figure 5.2: Case distribution per risk

The customer service team reviewed the classifications generated by the ChatGPT-4 model.

Their feedback was integral in validating the model's performance. The team found that the classifications generally aligned with their expectations and provided valuable insights into the model's accuracy. Cases classified as 'high' risk were reviewed with particular attention, and the team confirmed that these cases indeed warranted immediate action. During the review, it was noted that certain cases initially classified as 'medium-high' risk needed reclassification to 'high' risk, and vice versa, to ensure accurate prioritization. Specifically, 8 cases initially classified as 'medium-high' were reclassified as 'high' risk, while 6 cases initially classified as 'high' risk were adjusted to 'medium-high'. This iterative review process ensured a more precise identification of severe-risk cases, confirming the robustness of the AI model's classifications.

Chapter 6

Feedback's Classification

In this section, the focus shifts towards the development and implementation of machine learning classification models aimed at identifying and prioritizing severe-risk cases within the customer feedback dataset. The classification process involves training models on labeled data to predict the risk level associated with each feedback instance.

6.1 Methodologies

The primary goal was to optimize the performance of the machine learning models in accurately classifying and prioritizing severe-risk cases within the customer feedback dataset. An iterative process of experimentation was conducted, involving various text vectorization techniques, and class balancing methods, to identify the most effective model configurations for each algorithm.

Several machine learning algorithms can be explored for the binary classification task. This project explores logistic regression, naive bayes, random forests, support vector machines, and gradient-boosting classifiers.

The performance of the methods is evaluated using accuracy, precision, recall, and F1-score. ROC curves were also utilized to measure the accuracy of the classifiers.

6.1.1 Data Preparation

Before model training, the data undergoes preprocessing steps such as tokenization, text normalization, and feature engineering. The text preprocessing steps are consistent with those outlined in the topic modeling stage, ensuring coherence and uniformity across the analysis pipeline.

For classification, the feedback cases were categorized into binary classes: high-risk cases were labeled as 1, while all other cases were labeled as 0. This binary classification approach simplifies the model's task of distinguishing between severe-risk and non-severe-risk feedback instances, ensuring a focused prioritization process.

The dataset was divided into training and testing sets, with 80% of the data used for training and 20% reserved for testing. This split ensures that the models are trained on a substantial portion of the data while retaining a separate set for evaluating model performance.

The scikit-learn library was utilized for implementing machine learning algorithms, text vectorization, and class balancing techniques.

6.1.1.1 Class Balancing Techniques

The dataset was highly imbalanced, with severe-risk cases constituting only 2.9% of the total feedback. Although this imbalance was expected, it necessitated the use of specific techniques to ensure effective model training.

To address class imbalance two primary methods were tested: Synthetic Minority Over-sampling Technique and a combination of SMOTE with random undersampling. The choice of these techniques was driven by the need to balance the dataset effectively while retaining as much valuable information as possible.

SMOTE generates synthetic samples for the minority class (i.e., high-risk cases) by interpolating between existing minority class examples. This method helps to balance the class distribution without duplicating existing data points, thus providing more robust training samples.

SMOTE with Random Undersampling approach first applies SMOTE to increase the number of minority class samples, followed by random undersampling of the majority class to further balance the dataset. Specifically, the strategy involved oversampling the minority class to 30% of the majority class size and then undersampling the majority class to 70% of its original size. This dual strategy aims to prevent overfitting that can occur with excessive oversampling and ensures a more balanced class distribution by reducing the majority class size.

The decision to implement these techniques was based on the nature of the customer feedback dataset. Given that severe-risk cases are relatively rare, solely using undersampling could result in the loss of critical information. Therefore, oversampling through SMOTE was chosen to generate synthetic, yet realistic, samples of high-risk cases, enhancing the model's ability to recognize and prioritize such instances. The combination with random undersampling helps to manage the majority class size, preventing it from dominating the training process and ensuring a balanced dataset for more accurate model training.

6.1.1.2 Text Vectorization

For text vectorization, both CountVectorizer and TF-IDF methods were tested to transform textual data into numerical features suitable for machine learning models.

CountVectorizer converts text data into a BoW model, where the frequency of occurrence of each word in the document is counted. This method provides a straightforward representation of text data, capturing the presence and frequency of words.

TF-IDF vectorization considers not only the frequency of words within a document but also their importance relative to the entire corpus. This method gives higher weight to words that are unique to a particular document and less weight to common words that appear across many documents. By doing so, TF-IDF helps to highlight terms that may be more indicative of the specific content and context of the feedback.

6.1.1.3 Scaling

To ensure that numerical features are on a comparable scale, scaling is performed for models that require this preprocessing step, such as Support Vector Machines or Logistic Regression. Features with larger ranges could disproportionately influence the model, compromising the integrity of all features in contributing to the predictive power of the model.

For this purpose, the StandardScaler technique is employed. This technique standardizes the features by removing the mean and scaling each feature to unit variance. This process transforms the data to have a mean of zero and a standard deviation of one.

6.1.1.4 Feature Selection

Feature selection is a crucial step in the machine learning pipeline, aimed at identifying the most relevant features that contribute to the model's predictive performance. By selecting the most informative features, the complexity of the model is reduced, leading to improved performance and interpretability.

The initial feature set consisted of the numerical representations of the text data obtained through vectorization methods (CountVectorizer and TF-IDF). Additionally, several features were computed to capture further information about each feedback instance. These features aimed to enrich the dataset with linguistic and structural characteristics of the feedback text.

- **Number of Adjectives:** Adjectives from the feedback text were extracted using linguistic processing techniques, specifically utilizing the Stanza library, and calculated the total count of adjectives present in each feedback instance.
- **Number of Words:** The total number of words in the feedback text provides insight into the length and complexity of each feedback instance.
- **Number of Letters:** Similarly, the total number of letters in the feedback text offers a measure of the textual complexity and verbosity of each feedback instance.
- **Number of Vowels:** It was also calculated the number of vowels in the feedback text to capture additional linguistic characteristics.
- **Number of Unique Words:** The count of unique words in the feedback text indicates the diversity and richness of vocabulary used in each feedback instance.
- **Percentage of Unique Words:** This feature calculates the percentage of unique words in the feedback text, providing further insight into vocabulary richness.
- **Percentage of Adjectives:** The percentage of adjectives relative to the total number of words in the feedback text, offering another dimension to understand the descriptive nature of the feedback.

Although these features were all computed, not all may contribute equally to the predictive power of the models.

A correlation matrix was used to identify and remove highly correlated features, as they can introduce redundancy and multicollinearity into the model, which can negatively impact its performance. This involved computing the correlation coefficients between all non-text features and the target variables to identify redundant or highly correlated features.

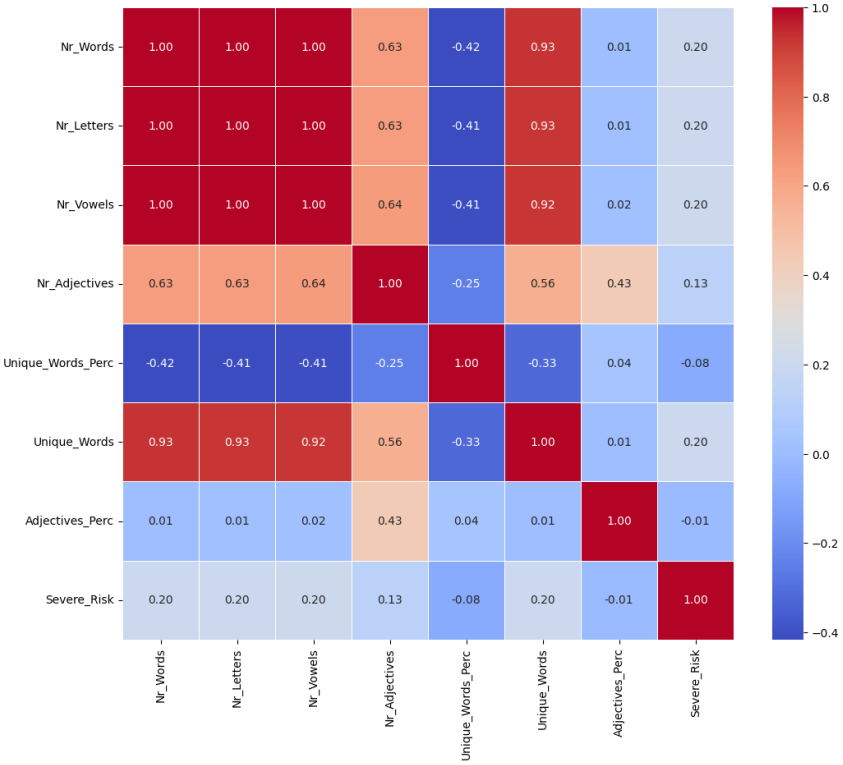

Figure 6.1: Features correlation matrix

The correlation matrix was used to identify highly correlated features. It was observed that variables such as Number of Words, Number of Letters, Number of Vowels, and Number of Unique Words had correlation coefficients close to 1. This high correlation indicates redundancy, as these features measure similar aspects of the feedback text length and complexity. To mitigate redundancy, the Number of Vowels, the Number of Letters and the Number of Unique Words were removed from subsequent analysis.

To further refine the feature selection process, feature importance scores were obtained using a Random Forest classifier. The Random Forest model ranks features based on their contribution to the model's predictive performance, providing insight into which features are most valuable. This method computes the average impurity reduction across all forest trees due to each feature, effectively ranking them by their importance.This ranking allows for the selection of the top features that significantly contribute to the model's predictive performance.

The model was initially trained on the full feature set, and the importance of each feature was calculated. Features with higher importance scores were retained for further analysis, while those

with lower scores were discarded.

Given the high-dimensional nature of the dataset (8503 features) and the relatively small training set size (6574 lines before balancing), it was crucial to reduce the number of features to prevent overfitting and improve model efficiency. The high number of features is primarily due to the use of text vectorization techniques, which convert textual data into numerical representations. Each unique word or token in the dataset is treated as a separate feature, resulting in a large feature space.

To address this issue, a cumulative importance threshold was set at 95%, to ensure that the selected features collectively contributed most to the model's performance and that the model retained most of the critical information while reducing complexity. This process resulted in a selection of 1435 features, which aligns better with the size of the training dataset.

Using the features selected by the Random Forest classifier, other machine learning models were then trained. This approach is advantageous as it leverages the strengths of Random Forest in feature selection to enhance the performance of various models.

6.1.2 Implementation

Following the preparation of data and the selection of classification models, an iterative process of experimentation with class balancing and vectorization methods was conducted. This iterative approach aimed to identify each algorithm's most effective model configurations.

Initialy, the focus was on testing the combinations of the two class balancing techniques (SMOTE and SMOTE with random undersampling) and the two vectorization methods (CountVectorizer and TF-IDF). Each model was thus run four times, once for each combination of class balancing and vectorization techniques. This systematic approach, outlined in Table 6.1, allowed for a thorough evaluation of the impact of these preprocessing steps on the model's performance.

Following this iterative process, the respective performance metrics were collected and analyzed. In the context of this project, high recall would mean that the model detects most of the high-risk cases, ensuring that few critical issues are missed. Alternatively, high precision would mean that when the model predicts a severe-risk case, it is usually correct, which is important for prioritizing responses. Given the specific challenges and goals of this project, the F1-score is chosen as the primary decision metric. The F1-score provides a balance, providing a single metric that captures both false positives and false negatives. The use of the F1-score helps ensure that the model is effective in identifying severe-risk cases that need urgent attention while maintaining a balance that prevents the allocation of excessive resources to false positives.

Once the model with the highest F1-score was identified, hyperparameter tuning was conducted to further optimize its performance. Hyperparameter tuning involves adjusting the parameters of the model, such as learning rate, regularization strength, and the number of trees in ensemble methods, among others. This process is crucial for fine-tuning the model to achieve the best possible performance on the validation set.

While the model with the highest F1-score was selected for further refinement, it is acknowledged that other models could potentially outperform the chosen model if subjected to similar

Table 6.1: Combinations of models with different vectorizers and sampling methods tested

Algorithm	TF-IDF	CountVectorizer	SMOTE	SMOTE + Undersampling	Scaling
XGBoost	X		X		
	X			X	
		X	X		
		X		X	
Random Forest	X		X		
	X			X	
		X	X		
		X		X	
Naive Bayes	X		X		
	X			X	
		X	X		
		X		X	
Logistic Regression	X		X		X
	X			X	X
		X	X		X
		X		X	X
Support Vector Machines	X		X		X
	X			X	X
		X	X		X
		X		X	X

hyperparameter tuning. The decision to focus extensively on one model was made to allocate resources efficiently and achieve a depth of optimization within the project's time constraints.

6.2 Results and Discussion

This section presents the results obtained from the various machine learning models and preprocessing combinations tested during this study. It provides a detailed analysis of the model performance, highlighting the metrics used to evaluate the effectiveness of each model in identifying and prioritizing severe-risk cases within the customer feedback dataset. The discussion includes a comparison of performance across different models and preprocessing techniques, the selection of the best-performing model, and the subsequent hyperparameter tuning to optimize its performance.

6.2.1 Performance Comparison

In evaluating the performance of the machine learning models, several metrics were considered, including precision, recall, and the F1-score. The table 6.2 summarizes the performance metrics for each model and preprocessing combination.

The F1-scores for each model and preprocessing combination are visualized in figure 6.2. These metrics offer insights into how well each model identifies high-risk cases and balances the trade-off between precision (the accuracy of high-risk predictions) and recall (the completeness

Table 6.2: Evaluation metrics for different model configurations

Model	Configuration	Precision	Recall	F1-score	Accuracy
XGBoost	TF-IDF - SMOTE	0.556	0.233	0.328	0.975
	TF-IDF - SMOTE + Undersampling	0.378	0.326	0.350	0.968
	CountVectorizer - SMOTE	0.700	0.163	0.264	0.976
	CountVectorizer - SMOTE + Undersampling	0.500	0.233	0.317	0.974
Random Forest	TF-IDF - SMOTE	0.667	0.186	0.291	0.976
	TF-IDF - SMOTE + Undersampling	0.424	0.326	0.368	0.971
	CountVectorizer - SMOTE	0.750	0.070	0.128	0.975
	CountVectorizer - SMOTE + Undersampling	0.667	0.093	0.163	0.975
Naive Bayes	TF-IDF - SMOTE	0.096	0.535	0.163	0.856
	TF-IDF - SMOTE + Undersampling	0.085	0.558	0.148	0.832
	CountVectorizer - SMOTE	0.142	0.465	0.217	0.912
	CountVectorizer - SMOTE + Undersampling	0.138	0.419	0.208	0.917
LRegression	TF-IDF - SMOTE	0.208	0.488	0.292	0.938
	TF-IDF - SMOTE + Undersampling	0.185	0.512	0.272	0.928
	CountVectorizer - SMOTE	0.236	0.302	0.265	0.956
	CountVectorizer - SMOTE + Undersampling	0.186	0.372	0.248	0.941
SVM	TF-IDF - SMOTE	0.361	0.302	0.329	0.968
	TF-IDF - SMOTE + Undersampling	0.320	0.372	0.344	0.963
	CountVectorizer - SMOTE	0.476	0.233	0.313	0.973
	CountVectorizer - SMOTE + Undersampling	0.379	0.256	0.306	0.970

of high-risk case identification). Additional metrics, including precision, recall, and accuracy for each model and preprocessing combination, and their ROC curves are provided in Appendix A. These comprehensive evaluations enable a deeper understanding of the model performances and their applicability in identifying high-risk cases.

The highest F1-score was achieved by the Random Forest model with SMOTE + undersampling and TF-IDF vectorization.

The combination of SMOTE with random undersampling generally improved the F1-scores for several models, notably for Random Forest and XGBoost. This suggests that balancing the dataset more effectively helped the models better learn to identify high-risk cases. TF-IDF vectorization consistently outperformed CountVectorizer across all models and class balancing techniques, especially for the Random Forest models. This indicates that considering the importance of words in the context of the entire corpus (as done by TF-IDF) provides more informative features. The Naive Bayes models performed relatively poorly across all configurations, highlighting potential limitations in handling the complexity and nuances of the customer feedback dataset. Logistic Regression and SVM models showed moderate performance improvements with TF-IDF and combined balancing techniques but did not outperform the tree-based models (Random Forest and XGBoost).

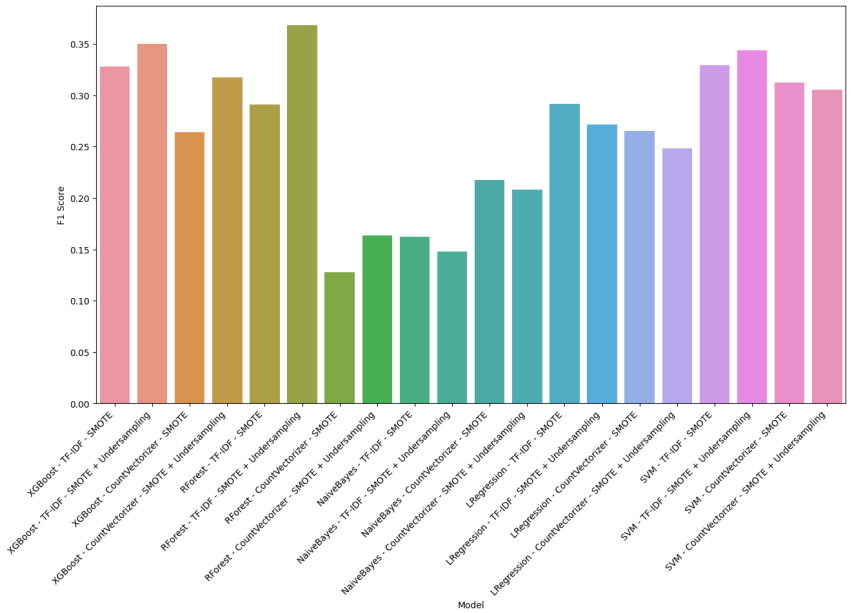

Figure 6.2: F1-score for each model and preprocessing combination

6.2.2 Best Model Results

Given the superior performance of the Random Forest model with TF-IDF and SMOTE with undersampling, hyperparameter tuning was conducted to further optimize its performance. The tuning process resulted in an improvement in the F1-score, further validating the initial selection based on preliminary performance metrics.

The tuning process was performed using grid search with cross-validation, which involved:

- **Number of Trees**: [100, 200, 300]
- **Maximum Depth**: [None, 10, 20, 30]
- **Minimum Samples Leaf**: [1, 2, 4]
- **Minimum Samples Split**: [2, 5, 10]

Grid search with cross-validation allowed for a systematic and exhaustive evaluation of the parameter space to identify the optimal settings for the Random Forest model. By iteratively testing different combinations of parameters, the process aimed to enhance the model's predictive performance on the validation set. This comprehensive approach ensured that the chosen parameters would yield the most robust model. The best model configuration is identified in table 6.3.

The confusion matrix for the best model is shown in Figure 6.3. The matrix illustrates the number of true positives, true negatives, false positives, and false negatives predicted by the model. The top 50 features from the model were also computed (see Appendix B).

From the confusion matrix, it is observed that there are 15 true positive cases (35%) where the model correctly classified instances as severe-risk, 1587 true negative cases (99%) where instances

Table 6.3: Best parameters and evaluation metrics for the optimized random forest model

Best Parameters	**Evaluation Metrics**
Number of Trees: 300	Accuracy: 0.9745
Maximum Depth: None	Precision: 0.5172
Minimum Samples Leaf: 1	Recall: 0.3488
Minimum Samples Split: 2	F1-score: 0.4167

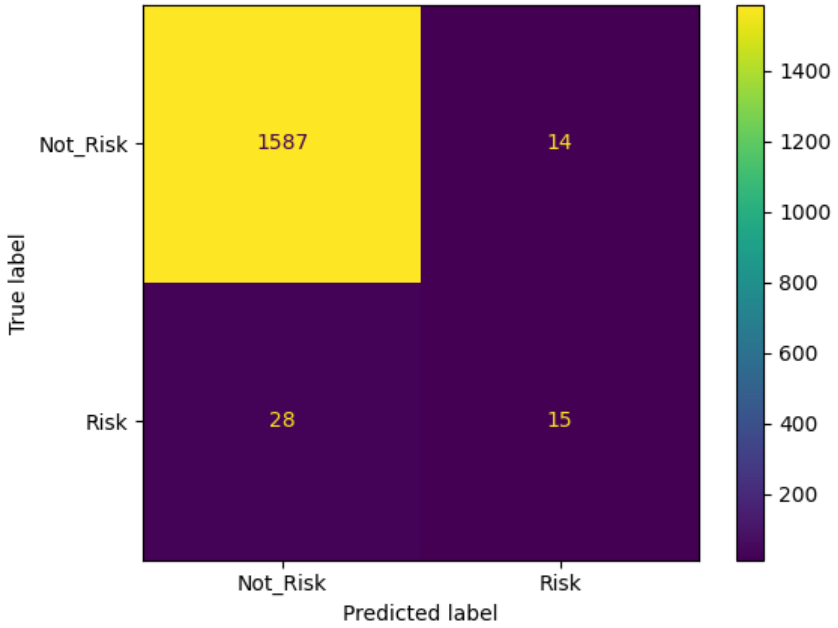

Figure 6.3: Confusion matrix of random forest with TF-IDF and SMOTE + undersampling

were correctly classified as not severe-risk, 14 false positive cases (0.9%) where instances were incorrectly classified as severe-risk when they were not, and 28 false negative cases (65%) where instances were incorrectly classified as not severe-risk when they were actually severe-risk.

The confusion matrix highlights several key aspects of the model's performance. Firstly, the model demonstrates a high true negative rate, correctly identifying 1587 out of 1601 not severe-risk cases. This indicates that the model is effective in identifying cases that do not pose a severe risk, thus preventing unnecessary allocation of resources to non-critical issues. Secondly, the model shows a moderate true positive rate, correctly identifying 15 out of 43 severe-risk cases. These identified cases can help mitigate potential risks and address critical issues proactively, which is vital for maintaining operational stability and customer trust. Nevertheless, there is room for improvement to ensure that more severe-risk cases are detected.

Additionally, there are 14 false positives, where non-severe-risk cases were incorrectly classified as severe-risk. The model's precision of 0.5172 indicates that when it predicts a severe-risk case, it is correct about half of the time. Although this can lead to some misallocation of resources, it is less critical than false negatives, as it ensures that potential severe-risk cases are not overlooked. Moreover, being able to flag these additional cases can be a significant advantage for

the business, to identify underlying issues that might not be immediately obvious. Furthermore, the model produced 28 false negatives, failing to identify some severe-risk cases. This is an area for improvement, as undetected severe-risk cases can pose risks to the organization.

However, it is essential to balance recall with precision to prevent overwhelming the customer service team with an excessive number of false positives, which could dilute focus and reduce efficiency. While a model with higher recall might capture more severe-risk cases, it would significantly increase the workload on the customer service team, necessitating the allocation of additional resources, which may not be feasible or cost-effective.

In this business use case, the chosen model strikes an optimal balance between precision and recall, ensuring that a manageable number of severe-risk cases are flagged for attention. This approach aligns with the strategic objective of optimizing resource allocation while maintaining a focus on high-priority issues. Thus, the model assists the customer service team in concentrating on the most critical cases, thereby enhancing operational efficiency and customer satisfaction.

This balanced approach ensures that the team can effectively address the most pressing concerns without being overwhelmed by less critical issues.

6.3 Implementation and Deployment

The next phase involves deploying the developed model within the company's operational framework. It is recommended the deployment to begin with an extensive evaluation of the model's performance across the remaining topics, as only three topics were utilized during this project. This evaluation will involve applying the model to feedback data from all topics to assess its generalizability and effectiveness.

If the model demonstrates robust performance across these additional topics, validated through rigorous testing and review processes, it will be deployed to classify all customer feedback cases. This comprehensive implementation aims to streamline the feedback management process, ensuring that severe-risk cases are promptly identified and addressed.

However, an alternative approach may be adopted if the model's performance varies significantly across different topics, indicating potential limitations on training a machine learning model on a small set of topics, to classify the entire dataset. In such cases, the framework proposed in this dissertation will be followed for the remaining topics. Specifically, a representative sample from each topic will be processed through the ChatGPT-4 API to classify the risk levels. These classifications will then be used to retrain the model, incorporating the diverse nuances and contexts of the different topics. This iterative retraining process will enhance the model's ability to accurately classify feedback across the entire spectrum of topics.

Chapter 7

Conclusion and Future Work

7.1 Main Conclusions

This dissertation aimed to develop and implement advanced analytical methodologies for prioritizing customer feedback based on risk severity at a leading retail company in Portugal. The application of generative AI and machine learning techniques, including natural language processing, topic modeling, and classification models, has enabled a more structured and systematic approach to managing customer feedback, particularly in identifying and prioritizing rare severe-risk cases.

The use of Latent Dirichlet Allocation for topic modeling proved to be a robust method for refining a large and unstructured dataset into coherent and manageable topics. The use of this method in this context deviated from its more generic application, which typically involves identifying and summarizing broad topics within a corpus. Instead, LDA was utilized with the specific objective of dataset reduction to enhance the manageability and relevance of the data for risk classification. This approach reduced the dataset by 26%, with this non-traditional use of topic modeling proving to be highly effective, demonstrating the flexibility and adaptability of LDA in addressing specific analytical needs.

Leveraging the enterprise ChatGPT-4 API for the pre-classification of feedback provided a practical solution to the challenges posed by the volume and complexity of the dataset and time constraints. This generative AI approach facilitated the categorization of cases into risk levels with a high degree of accuracy, as evidenced by the proof of concept validation. The ChatGPT-4 API significantly reduced the time and effort required to pre-classify a large dataset of customer feedback, enabling more efficient subsequent analysis. Incorporating a human-in-the-loop approach was crucial to the success of this process. This collaboration between AI and human expertise was instrumental in maintaining the accuracy and reliability of the pre-classification results. Notably, only a few cases required adjustments after human review, underscoring the effectiveness of the ChatGPT-4 API in accurately categorizing feedback.

The implementation of machine learning classification models marked the final phase of this project, further enhancing the system's ability to manage and prioritize customer feedback. The

combination of SMOTE with random undersampling proved effective in addressing class imbalance, leading to improved model performance, particularly for tree-based models. TF-IDF vectorization consistently outperformed CountVectorizer, emphasizing the importance of considering the significance of words within the context of the entire corpus. Among the models tested, the Random Forest model, utilizing TF-IDF vectorization and SMOTE with undersampling, achieved the highest F1-score. Hyperparameter tuning of this top-performing model proved advantageous, further enhancing its results. This model achieved a balanced approach, capturing a high number of severe-risk cases while maintaining precision to avoid excessive false positives, which could dilute focus and reduce efficiency.

Overall, the combination of advanced data preprocessing techniques, robust machine learning classification models, and careful evaluation metrics proved valuable in identifying and prioritize rare severe-risk cases within customer feedback.

In conclusion, this dissertation has demonstrated the feasibility and benefits of using advanced analytical techniques for customer feedback management. These advancements aim to translate into practical benefits, including improved customer satisfaction, better resource allocation, and potential cost savings.

7.2 Limitations and Future Work

While this study has achieved significant advancements, there are several areas for future research and development to further enhance the effectiveness of the proposed methodologies.

Future work could explore more topic modeling techniques, such as neural network-based models, to try to improve the quality and interpretability of the topics. One limitation was that some topics included a wide diversity of words and documents, resulting in less reduction of the dataset than initially expected. While the method aimed to filter out less relevant topics, it became apparent that all topics could potentially include cases associated with severe risks, which limited the extent of dataset reduction.

The ChatGPT-4 API's token limit of 128 thousand tokens constrained the number of cases that could be processed in a single interaction, necessitating the division of data into smaller batches. Due to this batch processing approach, a comprehensive relative ranking of all cases by severity was not possible, which could have enhanced the analysis and further accelerated the process as initially anticipated.

Despite this limitation, the model's proven efficacy suggests that it can be employed in future projects to expedite processes. It is further imperative to maintain the human-in-the-loop approach in future work to ensure continued accuracy and reliability. Additionally, the process of prompting the AI model must be carefully tailored and potentially adjusted to suit the specific use cases.

The inherent class imbalance, with severe-risk cases constituting only 2.9% of the total feedback, posed significant challenges in this study. Accurately capturing the nuances and contextual subtleties within the feedback was particularly difficult due to the dataset's complexity and variability. Although the model demonstrated strong performance on the test dataset, its applicability

to new data across different topics and changing contexts can further be tested and validated to ensure the model's robustness and scalability. Future work should further include extending hyperparameter tuning to other models to explore their potential further. This approach could lead to continuous improvements in the classification and prioritization of severe-risk cases.

By addressing the identified limitations and building on the project's successes, future work can further enhance these methodologies, ensuring their robustness, scalability, and applicability in diverse contexts.

Bibliography

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., and Hassan, A. (2023). Topic modeling algorithms and applications: A survey.

Agag, G., Durrani, B., Shehawy, Y., Alharthi, M., Alamoudi, H., El-Halaby, S., Hassanein, A., and Abdelmoety, Z. (2023). Understanding the link between customer feedback metrics and firm performance. *Journal of Retailing and Consumer Services*, 73:103301.

Albanese, N. (2022). Topic modeling with lsa, plsa, lda, nmf, bertopic, top2vec: a comparison a comparison between different topic modeling strategies including practical python examples.

Alghamdi, R. and Alfalqi, K. (2015). A survey of topic modeling in text mining-topic modeling; methods of topic modeling; latent semantic analysis (lsa); probabilistic latent semantic analysis (plsa); latent dirichlet allocation (lda); correlated topic model (ctm); topic evolution modeling.

Ali, A., Shamsuddin, S., and Ralescu, A. (2015). Classification with class imbalance problem: A review. 7:176–204.

Anantharaman, A., Jadiya, A., Siri, C., Bharath, A., and Mohan, B. (2019). *Performance evaluation of topic modeling algorithms for text classification*.

Baharudin, B., Lee, L. H., Khan, K., and Khan, A. (2010). A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology*, 1.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation.

Blümel, J. and Zaki, M. (2022). *Comparative Analysis of Classical and Deep Learning-based Natural Language Processing for Prioritizing Customer Complaints*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Brynjolfsson, E., Li, D., and Raymond, L. R. (2023). Generative ai at work.

Cai, Z., Siebert-Evenstone, A., Eagan, B., and Shaffer, D. W. (2021). Using topic modeling for code discovery in large scale text data. In *Advances in Quantitative Ethnography*. Springer International Publishing.

Carlson, J., Sourdin, T., Armstrong, C., Watts, M., and Carlyle, T. (2022). Return on investment of complaint management: A review and research agenda. *Australasian Marketing Journal*, 31:144135822211048.

Carrasco, M. V., Musolesi, M., O'Sullivan, J., Prior, R., and Manolopoulou, I. (2021). Regional topics in british grocery retail transactions.

Chen, R.-C., Dewi, C., Huang, S.-W., and Caraka, R. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*.

Chen, S., Li, Y., Lu, S., Van, H., Aerts, H., Savova, G., and Bitterman, D. (2024). Evaluating the chatgpt family of models for biomedical reasoning and classification. *Journal of the American Medical Informatics Association*, 31.

Cuffie, H. G., Najar, R. I., and Khasawneh, M. T. (2020). Topic modeling for customer returns retail data.

Dhieb, N., Ghazzai, H., Besbes, H., and Massoud, Y. (2019). Extreme gradient boosting machine learning algorithm for safe auto insurance operations. In *2019 IEEE international conference on vehicular electronics and safety (ICVES)*, pages 1–5. IEEE.

Dogra, V., Verma, Kavita, Chatterjee, P., Shafi, J., Jaeyoung, C., and Ijaz, M. F. (2022). A complete process of text classification system using state-of-the-art nlp models. *Computational Intelligence and Neuroscience*, 2022:1–26.

Evangelopoulos, N. E. (2013). Latent semantic analysis. *WIREs Cognitive Science*, 4(6):683–692.

Fabijan, A., Holmström, H., and Bosch, J. (2015). Customer feedback and data collection techniques in software rd: A literature review. volume 210, pages 139–153. Springer Verlag.

Filip, A. (2013). Complaint management: A customer satisfaction learning process. *Procedia - Social and Behavioral Sciences*, 93:271–275.

Gamon, M., Aue, A., Corston-Oliver, S., and Ringger, E. (2005). Mining customer opinions from free text.

Ghodousi, M., Alesheikh, A. A., Saeidian, B., Pradhan, B., and Lee, C. W. (2019). Evaluating citizen satisfaction and prioritizing their needs based on citizens' complaint data. 11.

Hadiat, A. R. (2022). Topic modeling evaluations: The relationship between coherency and accuracy 2022.

Hartmann, J., Huppertz, J., Schamp, C., and Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing*, 36(1):20–38.

Hassan, S., Ahamed, J., and Ahmad, K. (2022). Analytics of machine learning-based algorithms for text classification. *Sustainable Operations and Computers*, pages 238–248.

Hickman, L., Thapa, S., Tay, L., Cao, M., and Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 25(1):114–146.

Ikonomakis, M., Kotsiantis, S., and Tampakas, V. (2005). Text classification using machine learning techniques.

Jaiswal, J. and Samikannu, R. (2017). Application of random forest algorithm on feature subset selection and classification and regression. pages 65–68.

Jin, M. and Nikolaos, A. (2021). Modeling the severity of complaints in social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online. Association for Computational Linguistics.

Jusoh, S. (2018). A study on nlp applications and ambiguity problems. *Journal of Theoretical and Applied Information Technology*, 96:1486–1499.

Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, 52:273–292.

Kang, Y., Cai, Z., Tan, C., Huang, Q., and Liu, H. (2020). Natural language processing (nlp) in management research: A literature review.

Kastrati, Z., Dalipi, F., Imran, A. S., Nuci, K. P., and Wani, M. A. (2021). Sentiment analysis of students' feedback with nlp and deep learning: A systematic mapping study.

Katsiuba, D., Dolata, M., and Schwabe, G. (2024). Power of language automation: The potential for closing the loop in responding to online customer feedback.

Kherwa, P. and Bansal, P. (2020). Topic modeling: A comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems*, 7:1–16.

Kim, H., Castellanos, M., Hsu, M., Zhai, C., Rietz, T., and Diermeier, D. (2013). Mining causal topics in text data: Iterative topic modeling with time series feedback. pages 885–890.

Krishnan, A. (2023). Exploring the power of topic modeling techniques in analyzing customer reviews: A comparative analysis.

Kumar, A. and Kaur, A. (2020). Complaint management-review and additional insights. *Article in International Journal of Scientific Technology Research*, 9:2.

Laliberté, M., Casgrain, L., and Volesky, K. (2022). Handling complaints: Considerations for prioritizing complaints. *Canadian Journal of Bioethics*, 5:43–47.

Lester, B., Al-Rfou, R., and Constant, N. (2021). The power of scale for parameter-efficient prompt tuning.

Lind, F., Eberl, J.-M., Eisele, O., Galyga, T. H. S., and Boomgaarden, H. G. (2022). Building the bridge: Topic modeling for comparative research. *Communication Methods and Measures*, 16(2):96–114.

Lopes, N. and Ribeiro, B. (2015). *Non-Negative Matrix Factorization (NMF)*. Springer International Publishing.

Loukas, L., Stogiannidis, I., Malakasiotis, P., and Vassos, S. (2023). Breaking the bank with chatgpt: Few-shot text classification for finance.

Nah, F., Zheng, R., Cai, J., Siau, K., and Chen, L. (2023). Generative ai and chatgpt: Applications, challenges, and ai-human collaboration.

Occhipinti, A., Rogers, L., and Angione, C. (2022). A pipeline and comparative study of 12 machine learning models for text classification. *Expert Systems with Applications*, 201.

Peikos, G., Symeonidis, S., Kasela, P., and Pasi, G. (2023). Utilizing chatgpt to enhance clinical trial enrollment.

Pyasi, S., Gottipati, S., and Shankararaman, V. (2018). Sufat - an analytics tool for gaining insights from student feedback comments. pages 1–9.

Rahim, N., Ahmed, E., Sarkawi, M., Jaaffar, A., and Shamsuddin, J. (2019). Operational risk management and customer complaints the role of product complexity as a moderator.

Seliya, N., Khoshgoftaar, T. M., and Hulse, J. V. (2009). A study on the relationships of classifier performance metrics. In *2009 21st IEEE international conference on tools with artificial intelligence*, pages 59–66. IEEE.

Sha, L., Rakovi, M., Das, A., and Chen, G. (2022). Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education. *IEEE Transactions on Learning Technologies*, 15(4):481–492.

Tijare, P. and Rani, P. J. (2020). Exploring popular topic models. volume 1706. IOP Publishing Ltd.

Vayansky, I. and Kumar, S. (2020). A review of topic modeling methods. *Information Systems*, 94.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.

Wirtz, J. and Lovelock, C. (2021). *Services marketing: people, technology, strategy*.

Wölbitsch, M., Hasler, T., Walk, S., and Helic, D. (2020). Mind the gap: Exploring shopping preferences across fashion retail channels. pages 257–265.

Xin, D. (2017). *Big Data Technology and Ethics Considerations in Customer Behavior and Customer Feedback Mining*.

Yang, W., Tan, L., Lu, C., Cui, A., Li, H., Chen, X., Xiong, K., Wang, M., Li, M., Pei, J., and Lin, J. (2019). Detecting customer complaint escalation with recurrent neural networks and manually-engineered features. pages 56–63.

Yilmaz, C., Varnali, K., and Kasnakoglu, B. (2016). How do firms benefit from customer complaints? *Journal of Business Research*, 69(2):944–955.

Zairi, M. (1992). The art of benchmarking: using customer feedback to establish a performance gap. *Total Quality Management*, 3(2):177–188.

Zhang, Y. and Gosline, R. (2023). Human favoritism, not ai aversion: People's perceptions (and bias) toward generative ai, human experts, and human–gai collaboration in persuasive content generation. 18.

Appendix A

Performance Metrics

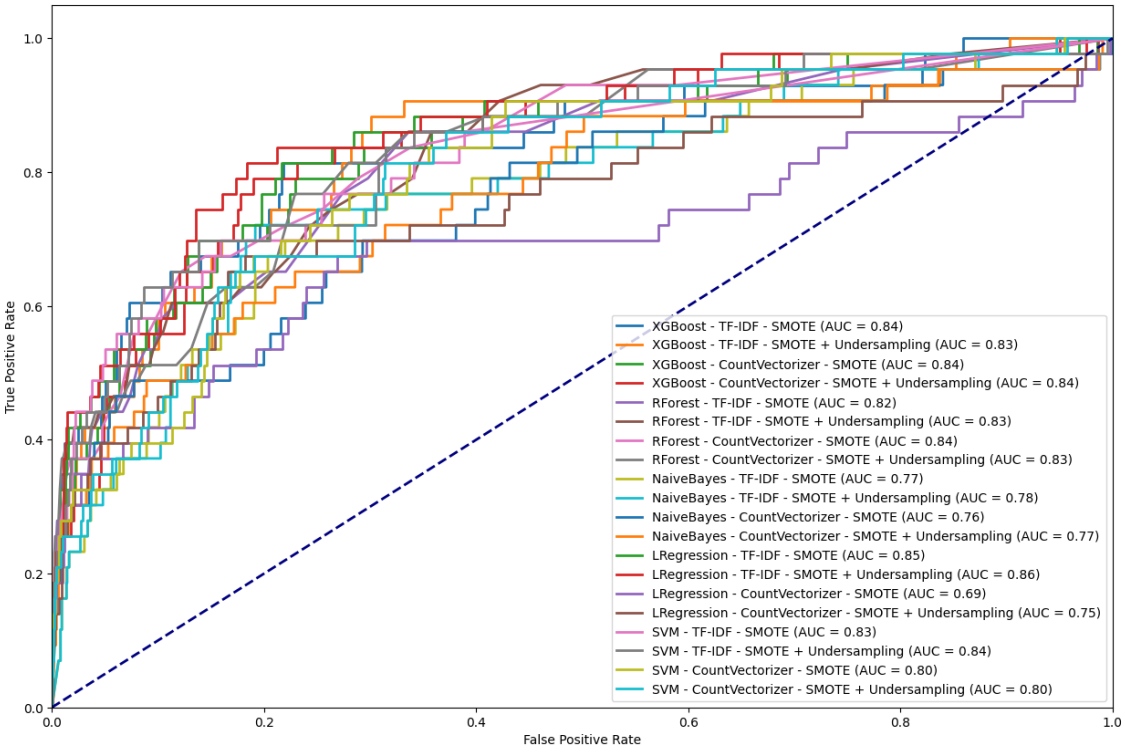

Figure A.1: ROC curves for each model and preprocessing combination

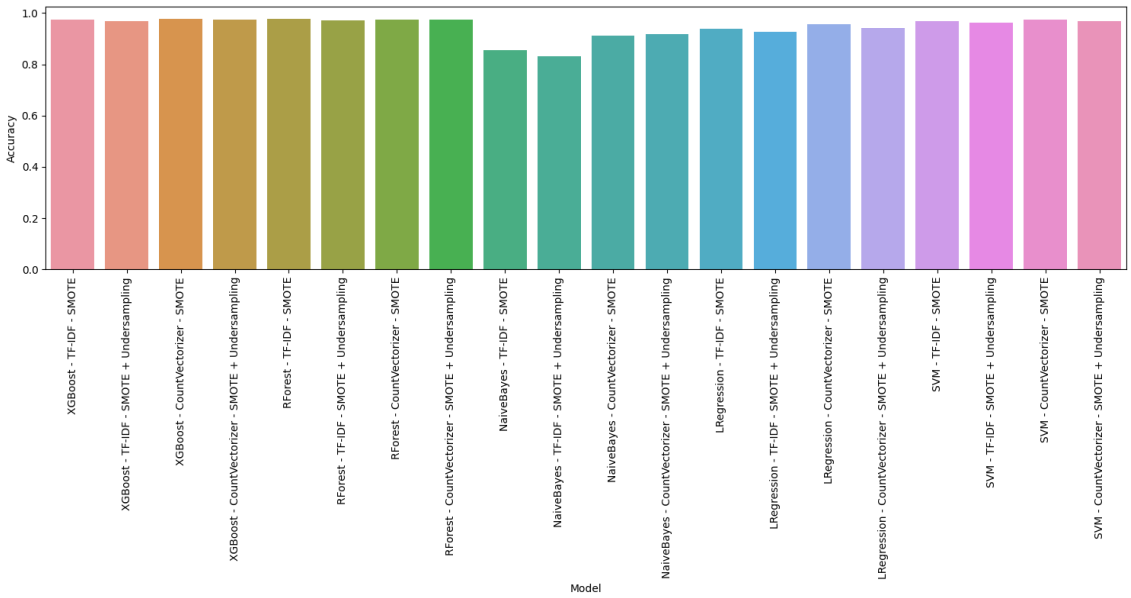

Figure A.2: Accuracy for each model and preprocessing combination

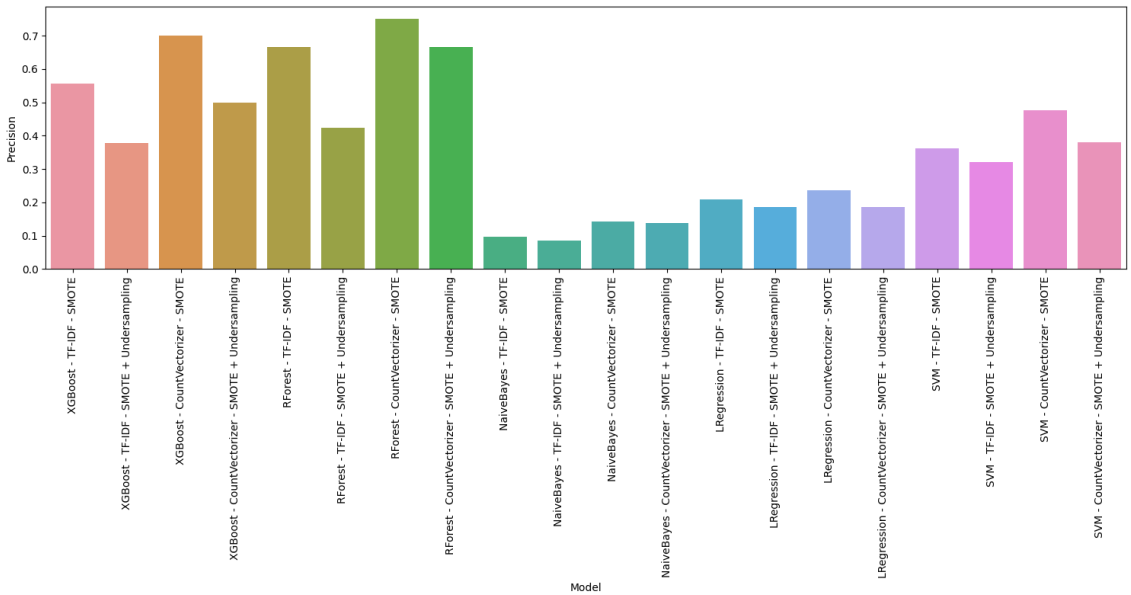

Figure A.3: Precision for each model and preprocessing combination

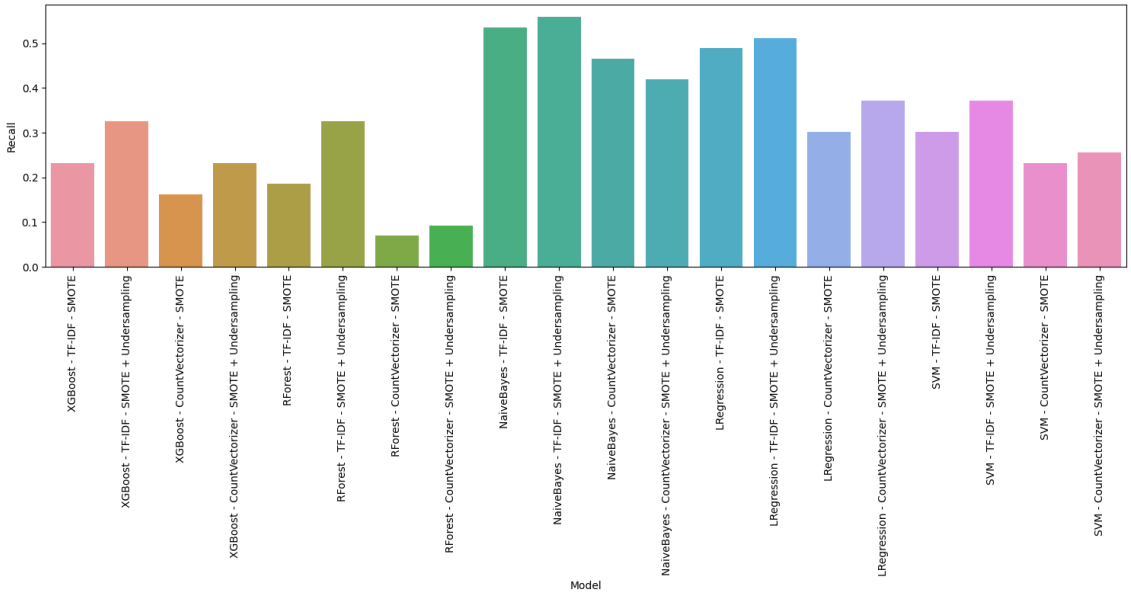

Figure A.4: Recall for each model and preprocessing combination

Appendix B

Feature Analysis

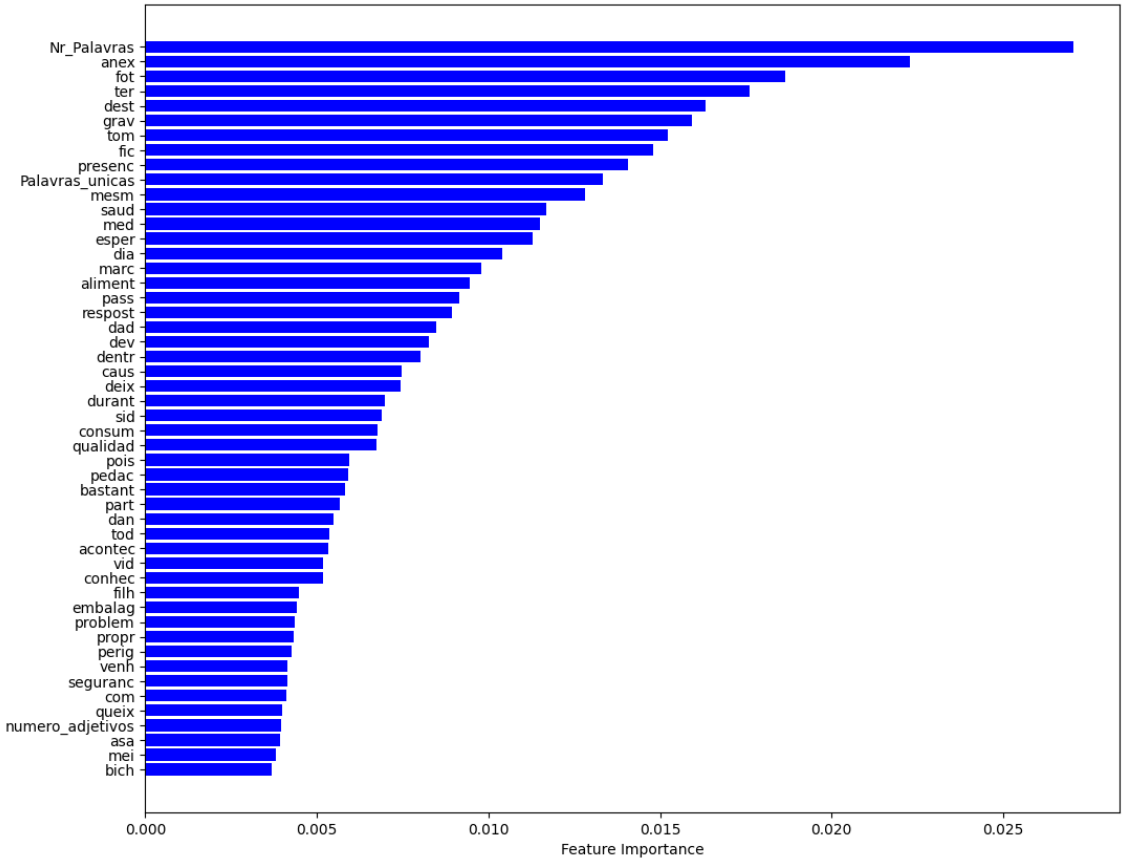

Figure B.1: Top 50 features from random forest model