Automatic Lead Qualification Based on Opinion Mining in CRM Projects: An Experimental Study Using Social Media

Victor Hugo Ferrari Canêdo Radich[a], Tania Basso[b] and Regina Lucia de Oliveira Moraes[c]
University of Campinas - UNICAMP, Limeira, Brazil

Keywords: Lead Qualification, Sentiment Analysis, Opinion Mining, Machine Learning, CRM, Lead Scoring, NLP.

Abstract: Lead qualification is one of the main procedures in Customer Relationship Management (CRM) projects. Its main goal is to identify potential consumers who have the ideal characteristics to establish a profitable and long-term relationship with a certain organization. Social networks can be an important source of data for identifying and qualifying leads, since interest in specific products or services can be identified from the users' expressed feelings of (dis)satisfaction. In this context, this work proposes the use of machine learning techniques and sentiment analysis as an extra step in the lead qualification process in order to improve it. In addition to machine learning models, sentiment analysis, also called opinion mining, can be used to understand the evaluation that the user makes of a particular service, product, or brand. The results indicated that sentiment analysis derived from social media data can serve as an important calibrator for the lead score, representing a significant competitive advantage for companies. By incorporating consumer sentiment insights, it becomes possible to adjust the Lead Score more accurately, enabling more effective segmentation and more targeted conversion strategies.

1 INTRODUCTION

Currently, we live in a world where major changes are taking place in consumption relationships, and the internet and social networks play a key role in these changes. Social networks have helped build a collective identity and create true consumer communities, becoming a major influence on consumption and overcoming marketing communications and even personal preferences. Consumers are continually sharing positive or negative stories about their experiences and preferences. (Kotler et al., 2017) found that spontaneous conversations about brands are more reliable than targeted advertising campaigns.

At the same time, market competition has also been changing, and technology plays an important role in this change. By connecting more than millions of people, businesses, governments, and advertisers, these technologies allow collecting, storing, and processing large amounts of information about the behavior, preferences, interests, ideas, knowledge, and physical and psychological characteristics of Internet users. This information can be a source of significant competitive advantage if they are used, for example, to assess the likelihood that a new contact will become a customer.

In this context, having an efficient process to identify potential customers with the ideal characteristics for a profitable and long-term relationship can be decisive for business continuity. To achieve better customer relationship, companies adopt what is called Customer Relationship Management (CRM), i.e., a software that allows the monitoring of all interactions with current and future customers (Jadli et al., 2022).

In the context of CRM, the management and qualification of leads play a crucial role in improving customer acquisition strategies. A lead is defined as a contact who has shown interest in a company's product or service. To enhance the accuracy of lead qualification, this study integrates traditional lead scoring techniques with sentiment analysis, leveraging data from social networks to adjust lead rankings based on user feedback and historical interaction data. The following section outlines the methodology used to implement and evaluate this enhanced approach. Thus, lead management comprises all the steps taken by a commercial team to track a future client (the lead) from the first contact with the company until the com-

[a] https://orcid.org/0009-0003-8512-2639
[b] https://orcid.org/0000-0003-2467-2437
[c] https://orcid.org/0000-0003-0678-4777

Radich, V. H. F. C., Basso, T. and Moraes, R. L. O.
Automatic Lead Qualification Based on Opinion Mining in CRM Projects: An Experimental Study Using Social Media.
DOI: 10.5220/0013237400003929
In *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025) - Volume 1*, pages 456-466
ISBN: 978-989-758-749-8; ISSN: 2184-4992

pletion of the purchase (Kotler and Keller, 2012). Still, in the context of customer relations, and to help the sales and marketing companies, lead scoring techniques are adopted. These techniques aim to identify leads that are more likely to become customers, qualify them, and then prioritize them (Koschnick, 1995).

With the emergence of big data, it became possible to create data-driven marketing in companies that use collected data for decision-making. However, few companies manage to achieve all the competitive advantages or have found the best way to practice data marketing. Forecasting is a problem for many companies, and most of Salesforce relies on intuition to evaluate each lead, which causes different kind of intuitions, making forecast inaccurate (Kotler et al., 2021).

In this work it is presented an alternative analysis for the mapping of the *persona* (i.e., ideal customer archetype). This analysis comprises the addition of a step in the lead management, which consists of sentiment analysis or opinion mining through social network data. This would provide decision makers with a proposal for improving lead scoring and qualification, and, consequently, assisting in developing more relevant marketing strategies for consumers current and future needs.

After this introduction, Section 2 presents some background, which is essential to understand the work. Related works are presented in Section 3. Section 4 explain the methodology followed by the results and discussion in Section 4.6. Finally, the Section 5 concludes presenting the main challenges for the future, as well as its potential impact on organizations.

2 BACKGROUND

This section brings the fundamental concepts for understanding the proposal. As such, Section 2.1 covers CRM systems and lead qualification, and ML-based lead scoring models are presented in Section 2.2. Section 2.3 addresses natural language processing (NLP) and sentiment analysis.

2.1 CRM Systems and Lead Qualification

A qualified lead is a contact identified by a company's marketing or sales team as a potential customer. Lead management (and interest group segmentation) are practices commonly used in CRM systems to help the company divide the market into groups of customers based on different needs, characteristics, or behaviors that may require a product or a strategy for differentiated marketing.

The main objective of a CRM system is to observe the life cycle and behavior of a consumer. Monitoring this customer-company relationship can also facilitate the creation of actions focused on customer loyalty and satisfaction. Thus, the CRM is a solution that can be used by the marketing and sales teams as well as the after-sales and service teams.

A strategic model widely used by sales and marketing teams to monitor customer relationships is the sales funnel. This tool is a visually structured model separated by stages, where the entire buying journey of a potential customer can be observed (Kotler and Keller, 2012). This model can be adapted according to the realities of each company or type of business. Figure 1 presents an expanded adaptation of the sales funnel of Kotler and Keller (2012).

In Figure 1, the steps 1 (Visitor/Prospect) and 2 (Lead) comprise the *Top of the Funnel*, where knowledge and discovery of the product or service by visitors take place; the marketing team does the prospecting (or attraction) in an attempt to convert the visitors into leads; the collection of basic contact information is performed. The steps 3 (Marketing Qualified Lead) and 4 (Sales Accepted Lead) comprise the *Middle of the Funnel*, where visitors have already interacted with some brand or product content and showed interest by providing some type of personal information, either directly or indirectly; lead qualification takes place and a relationship of trust is established with the future customer, so that he can advance in the sales funnel. Finally, the step 5 (Sales Qualified Lead) comprises the *Bottom of the Funnel*, where leads that have gone through the entire process of getting to know the product or service are identified; the marketing team has classified them as ready to be contacted by a salesperson.

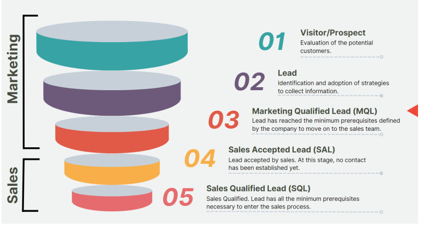

Marketing
Sales
01
Visitor/Prospect
Evaluation of the potential customers.
02
Lead
Identification and adoption of strategies to collect information.
03
Marketing Qualified Lead (MQL)
Lead has reached the minimum prerequisites defined by the company to move on to the sales team.
04
Sales Accepted Lead (SAL)
Lead accepted by sales. At this stage, no contact has been established yet.
05
Sales Qualified Lead (SQL)
Sales Qualified. Lead has all the minimum prerequisites necessary to enter the sales process.

Figure 1: Lead qualification process.(Kotler and Keller, 2012).

The qualification process is typically time-consuming and complex, and it frequently results in loss of temporal aspect of the potential client's need. Many times, when the approach is made, the customer's need has already been solved or no longer exists. For this reason, the need to qualify a lead quickly and assertively becomes increasingly relevant for companies and this is where our approach can help.

2.2 ML Based Lead Scoring Models

Lead Scoring aims to classify leads to determine which are most likely to purchase a particular product or service, and therefore, they should be prioritized within a sales process through a score assigned to lead actions in the funnel stages or by specific characteristics of the potential customer. In traditional lead scoring models, the values assigned to each action or characteristic of the lead are at the discretion of the sales or marketing teams (experts), who, empirically, assign a score to each item that makes up the score. Usually, a sum of these scores is obtained, and the responsible team will indicate whether or not a lead is qualified to proceed with a purchase or service acquisition, considering a final cut-off score previously defined by the company.

In this work, supervised models will be used, whose results, obtained after using classification methods, consist of labels assigned to a sample based on attributes and numbers for methods based on regression. Some ML algorithms were evaluated by Skiena (Skiena, 2017) based on a subjective analysis (considering five dimensions: power, ease of interpretation, ease of use, training speed, and prediction speed) and will be considered in this research for effectiveness analysis and performance in building lead scoring. The idea is to evaluate these different techniques to identify those that present good accuracy after training combined with good prediction speed.

2.3 Natural Language Processing (NLP) and Sentiment Analysis

Natural language processing (NLP) is a field of AI that gives machines the ability to read, understand, and derive meaning from human languages, which encompass both written and spoken language.

Currently, the two most widespread NLP applications are voice-controlled personal assistants and chatbots (which can even understand sentiments) (Kotler et al., 2021).

Sentiment analysis consists of mining texts in order to identify and extract subjective information that may help in understanding and classifying the opinion of the user who wrote the text. The objective of this type of analysis is to label the analyzed text according to the opinion or sentiment contained in it.

Opinion texts are generally informal and contain slang, irony, sarcasm, abbreviations, and emoticons. Due to the complexity of its process of understanding and manipulating language, it is common to use several different techniques to deal with different problems during sentiment analysis. One of these techniques is tokenization, which is used to divide a sentence into several elements (or *tokens*), while discarding some characters, such as punctuation or spaces.

Recent NLP research using artificial neural networks is enabling the creation of pre-trained models. BERT (Bidirectional Encoder Representations from Transformers), launched by Google in 2018 as a new algorithm in its engine, is a recent example of this evolution. This solution consists of a pre-trained NLP model that seeks to improve the user experience with a better understanding of what is being researched, with the objective of presenting increasingly assertive results, learning from the user experience. More recently, in October 2020, a group of researchers presented a version of BERT for Brazilian Portuguese, which was called BERTimbau (Souza et al., 2020). This model is very promising and represents a significant advance for the state of the art in this area of research.

3 RELATED WORK

Regarding lead scoring models, the work presented by (Benhaddou and Leray, 2017) describes a way to build a lead scoring model with a Bayesian network for CRM systems. In training, the model performed well in terms of precision, recall, and accuracy. However, the few available examples and the imbalance presented in the data set indicate that the model still needs to be improved. Custódio et al.(Custódio et al., 2020) proposed the construction of a lead scoring model for companies that operate in the context of public tenders. The authors compared some ML algorithms such as SVM, Random Forests, Neural Networks, and Adaboost, and the SVM presented the best performance for the data set. However, the expected results do not consider the company's expertise, focusing on historical data. Jadli et al.(Jadli et al., 2022) compare the performance of several ML algorithms to predict and drive models using lead scoring. The Random Forest and Decision Tree models presented the highest accuracy scores. This work served as a starting point for selecting the ML algorithms, and the partial results that we obtained corroborate their results regarding the algorithms' applicability tested so far.

Related to sentiment analysis, Feizollah et al.(Feizollah et al., 2019) presented a model that collected Twitter posts about halal tourism and cosmetics in the last ten years. An experiment was carried out to calculate and analyze the sentiment of tweets using deep learning algorithms. They grouped the texts into positive or negative sentiments and quantified them. The authors used extensive data collection to train the algorithms and achieved good accuracy. The work by Nilpao et al.(Nilpao et al., 2022) proposed an application to recommend coffee shops based on Twitter data. Based on the sentiment analysis collected, the application shows the coffee shop mentioned in the texts on the map. The model used for sentiment analysis is the Naive Bayes method and reached 86% of the mean accuracy.

Specific works in Brazilian Portuguese were analyzed. Cardoso and Pereira (Cardoso and Pereira, 2020) presented a supervised method using NLP tools for opinion mining in Portuguese and English languages. In this study, the authors investigated the maturity of the tools for Portuguese in comparison with the already-consolidated tools for the English language. The authors identified that it is not recommended to translate texts from Portuguese to English in order to obtain greater efficiency, as automatic translation introduces losses in the quality of texts.

In the work of Sousa et al. (Souza et al., 2020), BERT (Bidirectional Encoder Representations from Transformers) models for Brazilian Portuguese were trained. The authors identified that the models achieved superior performance for NLP tasks compared to multilingual BERT. This model in the pretrained version, for Brazilian Portuguese (BERTimbau), will be used in the present work. Models based on the Transformers architecture, such as BERT, performed well for our type of analysis (Souza et al., 2020).

Since some works do not consider the company's expertise (Custódio et al., 2020) or dataset characteristics (e.g., unbalanced data) (Benhaddou and Leray, 2017), and using intuition to evaluate each lead makes the forecast inaccurate (Koschnick, 1995), we believe that including sentiment analysis in the lead scoring process would help to improve the results of this process, especially when it comes to sentiment analysis in the Brazilian Portuguese language.

4 THE APPROACH

A quick and efficient automated lead qualification process is a significant competitive advantage for companies and organizations, as response time plays a critical role in converting leads into customers and can greatly influence consumer decision-making. Leveraging machine learning models and sentiment analysis from social media can further enhance this process by accelerating lead qualification and enabling personalized, timely responses, improving overall conversion rates.

So, it is understood that ML models combined with sentiment analysis in social networks can be used as important accelerators. By having these types of monitoring tools available, companies can combine data that is already stored in their knowledge bases with information shared in real time on networks. Still, as in traditional marketing, you can combine all this data in a model with the company's own expertise. Thus, the use of these technologies allows companies to anticipate potential customer recognition and be able to prepare more personalized responses at the most appropriate time.

The proposed approach is a lead qualification process consisting of five steps, as shown in Figure 1. This process is based on the conceptual scheme of the sales funnel, where an adaptation was made to represent the qualification of leads in companies that work with sales of products or services.

The final score will be reached after the execution of all the steps. Figure 2 shows the macro flow of the process, whose steps are:

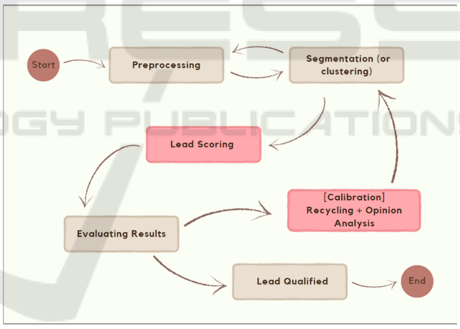

Start
Preprocessing
Segmentation (or clustering)
Lead Scoring
[Calibration] Recycling + Opinion Analysis
Evaluating Results
Lead Qualified
End

Figure 2: Score Calculation Macroflow Process.

- Pre-processing: in this step, the cleaning processes of the collected data will be carried out (removal of duplicate data, normalization, reduction, and transformation of data); the textual elements will also be pre-processed for sentiment analysis;
- Segmentation: in this step, the contact database is segmented based on characteristics and behaviors; converted and non-converted leads will be separated from the historical base; other groups can be established depending on the characteristic to be scored (for example, considering the ideal customer profile for a company, brand, service, or product); a later adjustment in the score can be made according to the definition of the company (expertise);

- Lead scoring: after cleaning the data and segmenting it into interest groups, the score can be defined; available historical data will be used, considering the activity and behavior of each lead and customer (leads already converted and segmented in the Segmentation step), as well as their profiles (as defined by the company), for assigning an initial score. The best ML model will be used to define the score;
- Calibration: a critical step of the entire process, where the contacts will be recycled according to the score obtained in the previous lead scoring stage, and those who have had some change in behavior (online or offline) will receive an appropriate score, in addition to adding the results obtained in the sentiment analysis. In this case, reviews with 1 to 3 stars are classified as negative, while those with 4 or 5 stars are classified as positive.
- Evaluation of results: this step consists of evaluating whether the score obtained is sufficient for the contact to be considered a qualified lead. Partial results will be evaluated according to the cutoff score (minimum score defined for the model); in practice, if the minimum score is reached, it will indicate that the leads in question are ready to be sent to the sales team.

For this study, we selected a public dataset provided by the Kaggle community (Kaggle, 2024), containing 9,240 records and 37 attributes related to lead behavior and the profile of a fictitious education-focused company. This dataset was chosen for being reasonably balanced between users who became customers (3,561) and those who did not convert (5,679), in addition to already containing the leads' behavioral history. These factors enable a more accurate analysis of the experimental results.

Since the calibration step requires sentiment analysis, it is important to define the best strategy to perform it. So, we investigated how the use of machine learning models and the use of an artificial intelligence-based algorithm (ChatGPT) can be applied to identify sentiments in evaluations posted by social network users. The goal is, first, to identify the best machine learning model for this context. Then, investigate whether it is worth using a model trained specifically with the text analysis or if it is better to use the generic ChatGPT model. Details and results of this study are in the subsection 4.4.

4.1 Predictive Model

Initially, the *pre-processing* step was performed (see Figure 2). The first analysis was focused on balancing the classes "converted leads" and "unconverted leads". Although the data were not fully equalized in proportion, we understand that they represented the reality of conversion rate, establishing a proportion of approximately 60% (not converted) to 40% (converted). Then it was necessary to make a cleaning in the base, as it had a lot of blank or null data. For this reason, some variables were removed from the dataset. The criterion adopted was to exclude variables that presented more than 50% of null values. In this first stage the dimension was reduced from 37 to 22 attributes. It is worth mentioning that among the attributes that remained, two represent the company's expertise and were previously evaluated based on the lead's activities. Continuing with a more careful analysis, some outliers were excluded, as they could distort the results of statistical analyses. So, it is important to identify and to treat them appropriately (Mitchell, 1997). With this step, the execution of the Pre-Processing stage of the solution was concluded.

In the *segmentation* stage (Figure 2), the data set was divided into two parts, one referring to converted leads and the other to non-converted leads. Using this approach, it was possible to identify some interesting behaviors and patterns. For example, both converted and non-converted leads come from the same source: Google. Also, in most cases, the last recorded activity of converted leads was sending an SMS, while for non-converters it was sending an email. These are some examples of observed behavior.

After cleaning the data and segmenting it into groups, the score could be defined in the *Lead Scoring* stage of the process. For this end, it was necessary to choose an appropriate Machine Learning (ML) model for the database in use, which was the Logistic Regression. The choice was made due to simplicity of application and the success stories observed in similar situations and reported in the literature (Jadli et al., 2022) (Yadavilli and Seshadri, 2021). The module used to build the model was *Logistic Regression* from the library Scikit-learn (Scikit-learn, 2024b).

For model training, the dataset was divided into two sets: training data (70%) and test data (30%). After evaluating the first training results, it was observed that the excess of variables to be analyzed harmed the results. Therefore, the tool *Recursive Feature Elimination* (RFE) (Scikit-learn, 2024c) was used to assist in the choice of the most important variables for defining the final model, with the 15 best classified by the method being chosen. As a last step, the *p-valor* was analyzed and those attributes that had a *p-valor* > 0.05 were eliminated.

The results obtained when we applied the model to the test data revealed an excellent specificity of

96.38%, meaning that the model is well tuned to correctly identify negative results and avoid false positives. Also, the model reaches a good accuracy of 84.9%, which indicates that it performs well in general in correctly predicting the results, suggesting that the model is well adjusted without signs of overfitting. However, the sensitivity of 66.96% indicates that the model can detect positive cases in a large proportion, but not in an excellent way, that is, the model may still be missing a considerable amount of positives (false negatives).

So, the model presented solid performance with high accuracy and specificity, in addition to maintaining a low false positive rate. However, the sensitivity and the negative predictive value indicate that the model can improve in detecting true positives. For the context of this application, we understand that the calibration stage can adjust the efficiency of the final score.

4.2 Lead Score Calibration

In this stage of the research, a *calibration* (Figure 2) layer was developed for calculation of the Lead Score, aimed at improving accuracy by incorporating sentiment analysis of reviews extracted from Google Play (Google LLC, 2024a) . This layer allows for a more refined adjustment of predictions, integrating user feedback as an additional factor for classifying the probability of lead conversion. Below are the main steps of the training process:

- **Data Collection.** user reviews from Google Play were extracted and analyzed, capturing the sentiments expressed regarding products or services;
- **Sentiment Analysis.** sentiment analysis was applied to the reviews, categorizing them as positive, negative, or neutral. This step utilized Natural Language Processing (NLP) techniques and compared the models with the best performance for this task;
- **Incorporation of Sentiments into the Lead Score.** The sentiment scores from each review were integrated into the Lead Score calculation pipeline. This created a calibration layer that weights the impact of these opinions on conversion predictions;
- **Model Training.** The machine learning model was trained using a dataset that included both traditional lead attributes (such as browsing behavior and previous interactions) and the sentiment variables extracted from the reviews;
- **Adjustment and Refinement.** The calibration of the Lead Score was adjusted based on the results obtained in training, ensuring that user sentiment had the appropriate weight in the final score calculation;
- **Validation and Evaluation.** After training, this layer was specifically added to incorporate the results of sentiment analysis, refining the lead scoring process to better reflect customer sentiment.

With this approach, the Lead Score not only reflects the observable behavior of leads but also considers the subjective perception of users expressed in their reviews. This adds a new dimension to the scoring process, enhancing the ability to predict leads with a higher potential for conversion.

4.3 Data Collection

For the data collection stage, the library *google-play-scraper* (Google LLC, 2024b) was used. This library was chosen because it abstracts the complexity of directly accessing Google Play pages and extracting data, providing a simple and efficient interface for obtaining structured information. Additionally, it offers interesting features such as filtering by language, country, and review rating, allowing for more targeted data collection. This level of flexibility was essential for adapting the extraction process to the specific needs of this study, ensuring that the captured information adequately reflected relevant user interactions and the target audience. The ability to sort reviews by relevance or date was also useful for prioritizing the most representative or recent opinions, respectively.

During the data extraction process (scraping), a total of 357,973 reviews from 94 apps across various categories and types were collected and stored in a local database using Microsoft SQL Server. The collection was systematic, ensuring that reviews were fully and accurately extracted, preserving important metadata such as publication date, star rating, and review content. The decision to store the data in a relational database like SQL Server was strategic to ensure scalability, security, and ease of querying, allowing for efficient and organized analysis.

4.3.1 Implications of Data Protection in Model Implementation

It is important to highlight that this work faced challenges related to the use of personal data, similar to those that many companies encounter when implementing artificial intelligence in their business contexts. The growing concern with regulatory compliance, particularly regarding the protection of personal data, requires a careful balance between innovation and privacy.

The General Data Protection Law (LGPD) in Brazil and the General Data Protection Regulation (GDPR) in Europe establish strict rules on the handling of personal data, including the collection, storage, and processing of sensitive information. In accordance with these regulations, organizations are required to ensure the privacy and security of data, as well as to obtain explicit consent from users for the use of their data for specific purposes, such as training machine learning models.

Due to the requirements imposed by these laws, it was not possible to use a real lead database for the experiments in this study. The use of personal data in an experimental context could infringe upon data protection laws, especially considering that, in many cases, data cannot be effectively anonymized to prevent the identification of individuals.

Thus, to ensure compliance with legal guidelines and mitigate the risks associated with the misuse of personal data, we opted to use a public lead dataset available on the Kaggle platform. This dataset was chosen because it does not contain sensitive or identifiable information, allowing us to perform the necessary simulations for the final lead score calculation without compromising individuals privacy.

Moreover, to ensure data representativeness and validity in the sentiment analysis process, user reviews from Google Play were randomly associated with the leads present in the public dataset. A table named `Lead_x_UserID` was created in the SQL Server database, establishing the relationship between Google Play users and the leads from the dataset. This approach ensured that the study followed compliance guidelines while enabling a robust and realistic analysis of the results obtained.

4.4 Sentiment Analysis Strategies

The prediction accuracy of machine learning models depends on the complexity of natural data and the performance of the learning algorithms (Sarker, 2021). To establish an effective machine learning model and get better prediction accuracy, it is necessary to select a suitable algorithm based on actual problems and then fully improve the model.

To evaluate the performance of different classification algorithms on our dataset, we trained and tested four models: Random Forest Classifier, Support Vector Machine (with Linear Kernel), Multinomial Naive Bayes, and Logistic Regression. These models were chosen because they are frequently cited in the literature as well-suited for text classification tasks (Custódio et al., 2020).

For this analysis, we used cross-validation on the defined models with the scikit-learn library (Scikit-learn, 2024a). Table 1 presents the accuracy rates obtained for the four classification models evaluated. Logistic Regression and Multinomial Naive Bayes achieved the highest performance rates, with accuracies of 0.8970 and 0.8945, respectively.

Table 1: Training set accuracies.

Model	Accuracy
RandomForestClassifier	0.6570
LinearSVC	0.8435
MultinomialNB	0.8945
LogisticRegression	0.8970

Although the difference between these two models is small, the confusion matrix analysis revealed that Logistic Regression has a slightly lower false positive rate, which may be crucial for our application, where accurately identifying positive opinions is a priority. Additionally, Logistic Regression provides a more intuitive interpretation of coefficients, facilitating the analysis of each feature's importance in the classification.

The results obtained reinforce findings from previous literature which highlight the advantages of using Logistic Regression for sentiment analysis and text classification (Mandloi and Patel, 2020; Jadli et al., 2022). Therefore, for this data, Logistic Regression model is the best model among the tested ones.

During the development of this research, ChatGPT was launched, a deep learning-based language model from OpenAI (OpenAI, 2024). The emergence of this new tool represented a promising opportunity for the field of Natural Language Processing, prompting us to consider its inclusion for comparing its performance with the previously chosen model.

Thus, we conducted a performance comparison between ChatGPT and logistic regression in the context of sentiment analysis. Using the ChatGPT API, we implemented automatic classification of the reviews and developed Python code to format the data, ensuring an organized and standardized input. The primary goal of this analysis was to determine which model performs better in categorizing the sentiments of the reviews.

The results obtained are described below:

- **Accuracy:**
  - **ChatGPT.** Achieved an accuracy of **0.8466**, meaning the model correctly classified 84.66% of the cases.
  - **Logistic Regression.** Slightly higher accuracy at **0.8566**, indicating 85.66% correct classifications.

- **Precision:**
  - **ChatGPT.** Precision was **0.8822**, showing that out of all instances predicted as positive, 88.22% were actually positive.
  - **Logistic Regression.** A very similar precision of **0.8821**, meaning 88.21% of the predicted positives were correct.
- **Specificity:**
  - **ChatGPT.** The model's specificity was **0.9301**, which reflects its ability to correctly identify 93.01% of the true negatives (i.e., how well it avoids false positives).
  - **Logistic Regression.** Slightly lower specificity at **0.9274**, indicating it correctly classified 92.74% of the true negatives.
- **Sensitivity (Recall):**
  - **ChatGPT.** Sensitivity (or recall) was **0.7302**, meaning it identified 73.02% of the actual positives.
  - **Logistic Regression.** Higher recall of **0.7579**, detecting 75.79% of the actual positives.
- **F1 Score**:
  - **ChatGPT.** The F1 score was **0.7991**, which is the harmonic mean of precision and recall, reflecting a balance between these two metrics.
  - **Logistic Regression.** A slightly better F1 score at **0.8153**, showing it performs better overall in balancing precision and recall.

In this context, ChatGPT is a slightly better model in terms of precision, indicating that it is a bit more effective at avoiding false positives. However, it has slightly lower sensitivity, which means it may be less effective at identifying all true positives.

Logistic Regression shows an overall advantage in terms of accuracy, sensitivity, and F1-Score. This suggests that, for this dataset, Logistic Regression offers a slightly better balance (better results) between avoiding errors and capturing all positive cases.

Both models exhibit similar performances, as illustrated by the ROC curve in Figure 3. The ROC (Receiver Operating Characteristic) curve visually compares the trade-off between the true positive rate (sensitivity) and the false positive rate (specificity) for different threshold values. However, Logistic Regression might be preferred in scenarios where capturing as many true positives as possible is crucial, due to its higher sensitivity. On the other hand, ChatGPT may be a better choice in situations where precision—minimizing false positives—is more critical, as indicated by its position on the ROC curve.

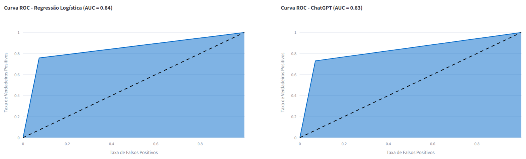


Figure 3: ROC Curve Logistic Regression vs ChatGPT.

Therefore, for the application in this project and with this specific dataset, Logistic Regression is the most suitable model.

4.5 Final Lead Score Calculation with Calibration Step

The calculation of the final lead score combines several aspects of a lead's potential to convert into a customer. This process leverages historical data, predefined corporate criteria, and sentiment analysis to create a comprehensive score. Here's a detailed explanation of each component and how they contribute to the final score:

- **Historical Lead Score** (LS_h)**.** This score is derived from the lead's past interactions with the company. It reflects behaviors such as previous purchases, engagement with marketing campaigns, website visits, and responses to communications.

  If a lead has engaged with emails frequently, visited product pages, and made inquiries, they would have a higher historical lead score (e.g., $LS_h = 80$ out of 100). Conversely, a lead with minimal interaction would have a lower score (e.g., $LS_h = 30$);
- **Corporate Lead Score** (LS_c)**.** This score is determined by the company based on the lead's characteristics and how well they fit the target audience profile. This could include factors like industry, company size, location, or demographic information.

  A lead from a target industry with a high potential for conversion might receive a higher corporate lead score (e.g., $LS_c = 70$). A lead from a non-target industry might receive a lower score (e.g., $LS_c = 40$);
- **Sentiment Lead Score** (LS_s)**.** This score is derived from sentiment analysis of the lead's feedback, reviews, or interactions with the company. It assesses the lead's feelings towards the brand, which can significantly influence their likelihood to convert.

  If sentiment analysis of the lead's interactions indicates a positive sentiment (e.g., through positive comments or high ratings), they would receive a higher sentiment lead score (e.g., $LS_s = 1.2$). In contrast, negative sentiment would lead to a lower score (e.g., $LS_s = 0.8$);

Final Lead Score Calculation. This work proposes the calculation of the final lead score using a combination of key parameters, as described in Equation 1. The calculation considers the lead's historical interactions with the company (LS_h), the alignment of the lead's profile with corporate criteria (LS_c), and the sentiment expressed by the lead towards the company (LS_s). The formula integrates these elements to provide a comprehensive score:

$$LS_{\text{final}} = (LS_h + LS_c) \times LS_s \tag{1}$$

Calculation Example 1. Let's consider a hypothetical lead with the following scores:

- Historical Lead Score (LS_h): 80
- Corporate Lead Score (LS_c): 70
- Sentiment Lead Score (LS_s): 1.2 (positive sentiment)

Using the Equation 1:

$$\text{LS}_{\text{final}} = (80 + 70) \times 1.2 = 150 \times 1.2 = 180 \tag{2}$$

In this case, the final lead score would be 180. This score indicates a strong potential for conversion, helping the sales team prioritize this lead over others with lower scores.

Calculation - Example 2. Now, let's consider a different lead:

- Historical Lead Score (LS_h): 30
- Corporate Lead Score (LS_c): 40
- Sentiment Lead Score (LS_s): 0.8 (negative sentiment)

Calculating the final lead score:

$$\text{LS}_{\text{final}} = (30 + 40) \times 0.8 = 70 \times 0.8 = 56 \tag{3}$$

In this example, the final lead score is 56, suggesting a lower likelihood of conversion, which would prompt the sales team to focus on leads with higher scores.

4.6 Final Training and Results

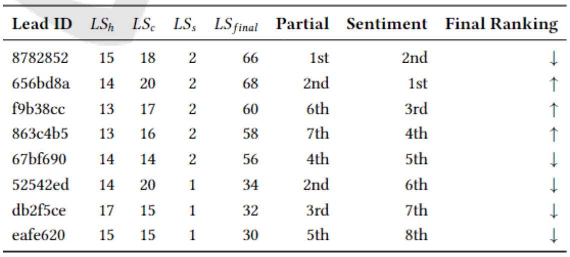
For the final training, the Logistic Regression model was used for sentiment analysis on the textual reviews, with the data split into 70% for training and 30% for testing and validation. A new column called ”sentiment” was created based on the star ratings, categorizing them as ”positive” (4-5 stars) or ”negative” (1-3 stars) through Python functions. After this, the tokenization process was carried out using BERTimbau, and the tokens were converted into NumPy arrays for use in scikit-learn.

To enhance the analysis, the TF-IDF technique was applied. It calculates the product of Term Frequency and Inverse Document Frequency, normalizing word counts and weighing the relevance of terms in each review. With the model adjusted, the training was executed, yielding the following results:

- Accuracy: 89.23%
- Precision: 90.86%
- Specificity: 95.52%
- Sensitivity: 79.06%
- F1 Score: 83.98%

These metrics demonstrate the efficiency of the model in predicting sentiment from textual reviews.

Figure 4 illustrates an example of the final lead score calculation, based on Equation 1, which was simplified through the assignment of weights in sentiment analysis. In this example, positive sentiments are assigned a weight of 2, while negative sentiments receive a weight of 1. The columns ”Partial” and ”Sentiment” display the partial ranking and the relative position of each lead.



Lead ID	LS_h	LS_c	LS_s	LS_{final}	Partial	Sentiment	Final Ranking
8782852	15	18	2	66	1st	2nd	↓
656bd8a	14	20	2	68	2nd	1st	↑
f9b38cc	13	17	2	60	6th	3rd	↑
863c4b5	13	16	2	58	7th	4th	↑
67bf690	14	14	2	56	4th	5th	↓
52542ed	14	20	1	34	2nd	6th	↓
db2f5ce	17	15	1	32	3rd	7th	↓
eafe620	15	15	1	30	5th	8th	↓

Figure 4: Comparative Lead Score Calculation.

In this example, the lead identified as 656bd8a has a final lead score of 68, initially ranking 2nd in the partial classification. However, due to a stronger sentiment score, where it is ranked 1st in sentiment, it moved up to 1st position in the Final Ranking.

Conversely, the lead 8782852, which initially held the 1st position in the partial ranking, fell to 2nd place in the Final Ranking due to a less favorable sentiment score, remaining 2nd in this metric.

In this context, it is possible to highlight how different factors, such as history, behavior, and sentiment, directly influence the final ranking of each lead, allowing a more detailed analysis for strategic prioritization decisions.

5 CONCLUSIONS

This study demonstrated that by incorporating user feedback alongside historical data and company-defined scores, the proposed model provided a more comprehensive and accurate prediction of lead conversion potential.

The results indicate that sentiment analysis substantially improved the calibration of the final score, capturing not only the objective behavior of leads but also users' subjective perception of the brand, product, or service. This approach enables companies to prioritize leads with a higher likelihood of conversion more effectively, representing a competitive advantage in an increasingly data-driven marketing environment.

However, challenges and opportunities for future research remain. A relevant improvement would be to expand the dataset to include more extensive and diverse sources of social media reviews, which could enhance the model's generalization across different sectors. Additionally, testing other deep learning architectures may further increase the accuracy of sentiment analysis and lead scoring. Moreover, enhancing real-time sentiment analysis and integrating it more seamlessly with CRM systems could provide actionable insights for sales and marketing teams, enabling quicker responses and more personalized engagement with potential customers.

In summary, the incorporation of sentiment analysis into lead scoring models represents a significant advancement in optimizing lead management processes. By refining the way companies assess and prioritize potential customers, this approach has the potential to increase conversion rates and support more targeted and effective marketing strategies.

ACKNOWLEDGEMENTS

This work was financed by CAPES - Coord. de Aperfeiçoamento de Pessoal de Nível Superior - Brasil, finance code 001. Also, it was supported by the PPGT/FT - UNICAMP.

REFERENCES

Benhaddou, Y. and Leray, P. (2017). Customer relationship management and small data — application of bayesian network elicitation techniques for building a lead scoring model. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*.

Cardoso, B. and Pereira, D. (2020). Evaluating an aspect extraction method for opinion mining in the portuguese language. In *Symposium on Knowledge Discovery, Mining and Learning (KDMILE)*.

Custódio, J., Costa, C. J., and Carvalho, J. P. (2020). Success prediction of leads – a machine learning approach. In *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*.

Feizollah, A., Ainin, S., Anuar, N. B., Abdullah, N. A. B., and Hazim, M. (2019). Halal products on twitter: Data extraction and sentiment analysis using stack of deep learning algorithms. *IEEE Access*, 7:83354–83362.

Google LLC (2024a). Google play. Acesso em: 12 out. 2024.

Google LLC (2024b). Google play. Acesso em: 12 out. 2024.

Jadli, A., Hamim, M., Hain, M., and Hasbaoui, A. (2022). Toward a smart lead scoring system using machine learning. *Indian Journal of Computer Science and Engineering (IJCSE)*, 13(2):433–443.

Kaggle (2024). Dataset. Acesso em: 12 out. 2024.

Koschnick, W. (1995). *Dictionary of Marketing*. Gower Pub Co.

Kotler, P., Kartajaya, H., and Setiawan, I. (2017). *Marketing 4.0—Moving from Traditional to Digital*. John Wiley and Sons.

Kotler, P., Kartajaya, H., and Setiawan, I. (2021). *Marketing 5.0: Technology for Humanity*. John Wiley and Sons.

Kotler, P. and Keller, K. (2012). *Marketing Management - 14th Edition*. Pearson Education Inc., Prentice Hall.

Mandloi, L. and Patel, R. (2020). Twitter sentiments analysis using machine learning methods. In *2020 International Conference for Emerging Technology (INCET)*.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill, Alemanha.

Nilpao, P., Nanta, N., Suetrong, N., and Promsuk, N. (2022). Development of the recommended coffee shops application based twitter sentiment analysis. In *2022 37th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*.

OpenAI (2024). Ask chatgpt anything. Acesso em: 12 out. 2024.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Comput. Sci.*, 2(3).

Scikit-learn (2024a). Cross-validation: evaluating estimator performance. Acesso em: 12 out. 2024.

Scikit-learn (2024b). Machine learning in python. Acesso em: 12 out. 2024.

Scikit-learn (2024c). Rfe. Acesso em: 12 out. 2024.

Skiena, S. S. (2017). *The Data Science Design Manual*. Springer, Suiça.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *In: Cerri, R., Prati, R.C. (eds) Intelligent Systems. BRACIS 2020. Lecture Notes in Computer Science, vol 12319*.

Yadavilli, S. and Seshadri, K. (2021). A framework for predicting item ratings based on aspect level sentiment analysis. In *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 327–332.