Identification of fake reviews in Portuguese using machine learning

Lucas Percisi [**Universidade Federal da Fronteira Sul** | ***lucas.percisi@estudante.uffs.edu.br***]
Guilherme Dal Bianco [**Universidade Federal do Rio Grande do Sul** | ***guilherme.dalbianco@uffs.edu.br***]

Universidade Federal da Fronteira Sul, SC-484, Km 02, s/n, Chapecó, SC, 89815-899, Brasil.

Received: 30 Novembro 2023 • **Accepted:** 11 Dezembro 2023 • **Published:** DD Month YYYY

Abstract This study addresses the growing challenge of fake reviews on online platforms, a phenomenon that significantly affects consumer trust and the business environment. Focusing on the detection of fake reviews in Portuguese, an area still underexplored, the work concentrates on analyses of bars and restaurants in Brazil and Portugal. Employing machine learning techniques and natural language processing, the study distinguishes genuine reviews from fraudulent ones. For this purpose, a dataset was created through *web scraping* of *yelp.com.br*, on which an exploratory analysis was conducted, emphasizing textual and behavioral *features*. The results indicate that, with an adequately balanced dataset, it was possible to achieve an *F-Score* of 87.53%, a performance comparable to studies conducted in English. This work provides valuable *insights* for identifying fake reviews in Portuguese, contributing to research and practices in this field.

Keywords: Fake Reviews, Web Scraping, Machine learning, Natural Language Processing, Yelp

1 Introduction

The rapid growth of e-commerce, offering a vast selection of products and services online, has attracted an increasing number of consumers to the digital sphere. Online business platforms like *Amazon*, *Yelp*, and *TripAdvisor* provide customers with the opportunity to share reviews about their experiences, reflecting opinions on products and services acquired through ratings and comments. The content generated by these users plays an influential role in consumers' purchasing decisions and is crucial for maintaining the reputation and profits of sellers [Jindal and Liu, 2007].

The rise of e-commerce, while bringing numerous benefits, has also introduced the possibility of fraudulent practices, such as the creation of fake reviews. These reviews can distort consumers' perceptions of a product or service and directly affect sales and company reputation [Luca and Zervas, 2016]. The complexity in detecting these fake reviews in online environments underscores the need to develop effective techniques to combat them.

Existing approaches to identifying fake reviews generally rely on linguistic and behavioral characteristics. Linguistic characteristics refer to the textual content of the reviews. Behavioral characteristics, on the other hand, concern the patterns of behavior of the users posting the reviews, observable in elements like posting frequency, review history, review date, average rating, among other indicators of spam behavior [Sandulescu and Ester, 2015].

The task of identifying fake reviews is challenging, especially due to the constant evolution of techniques to create these reviews and the involvement of individuals specifically hired to write them deceptively. These reviews, often artificially boosted to influence consumer perception, are intentionally created to mislead, generating extremely positive or negative impressions about products or services. Analyzing the relationships between users and the collective behavior of reviewer groups can be an effective path to identifying these deceptive activities, where patterns of interactions and expressed sentiments reveal communities engaged in artificially boosting reviews [Choo *et al*., 2015].

Despite numerous studies over the years, it remains challenging to establish a clear distinction between legitimate and fraudulent reviews. This is due to the difficulty in objectively discerning between honest and dishonest opinions, and even human reviewers struggle to identify fake reviews [Crawford *et al*., 2015]. Therefore, it is essential to analyze not only the textual content of the reviews but also the characteristics of the user profiles that post them.

This project aims to explore supervised machine learning techniques, feature extraction, and natural language processing to identify fake reviews in Portuguese on the platform *yelp.com.br*. This area of research is less explored compared to investigations in English, which are widely represented in studies by [Birim *et al*., 2022], [Barbado *et al*., 2019], [Elmogy *et al*., 2021], and [Sihombing and Fong, 2019]. Although research in detecting fake reviews in Portuguese is less common, it is not nonexistent, as demonstrated by the works of [Lima, 2013], [Schuch, 2013], and [Cruz *et al*., 2020]. This gap is concerning, as misinformation and fake reviews are global problems, affecting consumers and businesses in all languages and regions. Furthermore, techniques developed for English may not be as effective in Portuguese, due to linguistic and cultural differences.

Considering the linguistic and cultural diversity, this research seeks to fill an important gap by applying machine learning and natural language processing techniques to identify patterns of inauthenticity in online reviews in Portuguese. The bar and restaurant sector in Brazil and Portugal was cho-

sen due to the richness of data available, essential for a detailed analysis of trends in reviews.

The aim of this work is to investigate the use of supervised machine learning and feature extraction in the context of the social platform *yelp.com.br*, using a dataset collected through *web scraping*, which includes textual and behavioral characteristics. It seeks to identify distinct attributes between true and false reviews, offering practical implications for online platforms.

The structure of this paper is organized as follows: Section 2 discusses Related Works, providing theoretical context and exploring previous research. Section 3, Theoretical Framework, delves into the concepts and theories underlying the research. Section 4 details the Dataset, addressing the collection, processing, and exploratory analysis. Section 5 describes the Methodology applied in the search for an efficient machine learning model to identify fraudulent reviews. Finally, Section 6 is dedicated to the Analysis of Results, discussing *insights* and practical implications of the study.

2 Related Works

The literature on the detection of fake reviews presents a variety of approaches, with studies such as [Birim *et al*., 2022], [Barbado *et al*., 2019], [Elmogy *et al*., 2021], and [Sihombing and Fong, 2019] focusing on the use of machine learning techniques to identify these reviews. Another interesting aspect is addressed by [Valdivia *et al*., 2019], who explore the inconsistencies in reviews, proposing a unified index to better correlate the user rating and sentiment analysis, thus contributing to an alternative understanding of product and service reviews on online platforms.

The study by [Birim *et al*., 2022] focuses on the detection of fake reviews on *Amazon*, using a balanced dataset of 10500 genuine and 10500 fake reviews. This work analyzes the efficiency of various classifiers, including *Decision Tree*, *Random Forest*, *Support Vector Machine (SVM)*, and *Neural Networks*, testing different combinations of features. Among the evaluated features are sentiment scores, topic distributions, cluster distributions, and a sparse word matrix, as well as reviewer-centered factors such as review length and the *Verified Purchase* feature. The results highlight that behavior-related features, especially the *Verified Purchase*, are crucial in distinguishing between authentic and fake reviews, especially when combined with textual aspects. When the sentiment score was removed from the RF classification model, which combined Topic Distribution and *Verified Purchase*, there was a slight improvement in the model. Notably, the RF model with Topic Distribution and *Verified Purchase* achieved the best results, with an F-score of 80.08%. The study used 5-fold cross-validation for training and testing.

In [Barbado *et al*., 2019], a detailed study on the detection of fake reviews was carried out, using a proprietary and balanced dataset extracted from *Yelp*. This dataset consists of reviews from four U.S. cities, focused on consumer electronics retailers, with an equal number of reliable and fake reviews for each city, where reviews filtered by *Yelp* were considered fake. When analyzing text-centered review features, the *TF-IDF* technique with bigrams was the most effective, but achieved an F-Score below 60%, indicating that textual features are not strong indicators of review authenticity in the consumer electronics domain. However, by focusing on user behavioral characteristics, the study, using the *AdaBoost* classifier, achieved an *F-Score* of 82% using 10-fold cross-validation. This result emphasizes the effectiveness of behavioral characteristics in detecting fake reviews and the importance of considering user behavior on the platform, in addition to textual content, to efficiently identify fraudulent reviews.

Another reviewed study was that of [Elmogy *et al*., 2021], which conducted a detailed analysis of fake review detection using a *Yelp* dataset focusing on textual content. This dataset included 5,853 reviews of 201 hotels in Chicago, classified by *Yelp* into 4,709 genuine and 1,144 fake reviews. The authors did not use cross-validation in their analysis but a train-test split method with proportions of 70/30% respectively. In addition to experimenting with a variety of classifiers, including *K-Nearest Neighbors (KNN)*, *Naive Bayes*, *Support Vector Machine (SVM)*, *Logistic Regression*, and *Random Forest*, the authors also considered *N-gram* language models, specifically *bi-gram* and *tri-gram*. Additionally, *metafeatures* such as the number of uppercase letters, number of punctuations, and number of emojis in each review, which are indicative of user behavior during the writing of their reviews, were considered. The research revealed that incorporating these features improved the performance of classifiers by 3.80%. The *KNN* (K=7) with *TF-IDF* and *tri-grams* showed the best F-Score among the tested models, with a value of 86.20%.

In [Sihombing and Fong, 2019], an investigation was conducted on the detection of fake reviews in a *Yelp* dataset. The original dataset initially presented a significant imbalance, with a much larger proportion of unfiltered reviews compared to filtered ones, totaling 53400 genuine reviews and 8141 fake reviews. To address this issue, the authors applied *oversampling* techniques, creating copies of the minority class. They explored various machine learning classification techniques, including *Logistic Regression*, *Gaussian Naive Bayes*, *Support Vector Machine*, and *XGBoost*, using both textual and behavioral characteristics in reviews. The trained dataset ended up with 53400 genuine reviews and 24421 fake reviews, with 16280 of the 24421 fake reviews coming from the *oversampling* process. The authors highlighted the superior performance of *XGBoost*, which achieved an approximate F-Score of 99%, and for the other classifiers used it was at most 78%. The authors did not describe the use of cross-validation during the study, nor did they provide additional information on how *XGBoost* achieved such efficacy. This lack of detail may suggest the possibility of *overfitting* in the *XGBoost* model. Finally, the detailed analysis of [Sihombing and Fong, 2019] describes challenges faced in the performance of the tested models due to the imbalance of the dataset.

In the study by [Valdivia *et al*., 2019], the research team focused on resolving the incongruences between user numerical ratings and sentiment analyses in review texts on *TripAdvisor*. They investigated the frequently observed discrepancies where users, despite presenting positive ratings, in-

cluded negative or neutral sentences in their textual reviews. To this end, they analyzed opinions on six Italian and Spanish monuments, applying eight Sentiment Analysis Methods (SAMs) to analyze correlations between the numerical polarity given by the user and the polarity expressed in the sentences of the reviews. As a solution, [Valdivia *et al*., 2019] proposed the Polarity Aggregation Model, which integrates the user's polarity and the polarity extracted by the SAMs. This model, calibrated through a β parameter, proved effective in balancing the two polarities, providing a corrected assessment that lies between the user's rating and the sentiment employed in the review detected by the SAM. The study revealed that this new aggregation model could be particularly useful for reassessing sentiments expressed in reviews of various monuments, however, the challenge of identifying a good β in other contexts may hinder the application of the model. The study by [Valdivia *et al*., 2019] highlights the importance of complementary approaches to ensure that reviews of products and services offered online are faithful to reality, aligning with the goal of identifying fake reviews.

The previously presented results offer a broad view of the approaches to combat fake reviews, from the analysis of textual and behavioral characteristics of users to the integration of sentiment analysis. The studies highlight the importance of detecting fake reviews, given the significant influence that reviews have on consumer decisions [Jindal and Liu, 2007]. Moreover, the reviewed studies show a marked trend in fake review research: the vast majority were conducted with content in English. This concentration in English reflects the widespread use of the language on global platforms and greater ease of access to data in this language. However, this linguistic limitation restricts the effectiveness of fake review detection techniques, as they do not consider the specific characteristics of other languages. This trend towards English reveals an important gap in the field of fake review studies, especially considering the global reach of e-commerce and online review platforms. Therefore, it is crucial to expand research to include other languages, such as Portuguese, to develop more comprehensive and effective detection methods, adapted to the linguistic and cultural peculiarities of each region.

3 Research Method

This work was meticulously conducted through several stages, encompassing everything from initial data collection to detailed exploratory analysis. The adopted methodology is illustrated in Figure 1, which schematizes the complete research process flow. This flow includes:

1. **Data Collection and Structuring:** Initially, web scraping was conducted for data acquisition. This step involved collecting and organizing raw data, establishing a solid foundation for subsequent analyses.
2. **Data Pre-processing:** Here, the collected data underwent a process of cleaning and standardization, ensuring the quality and uniformity of the data.
3. **Feature Extraction:** This phase was dedicated to identifying and extracting relevant features from the data, preparing them for modeling.
4. **Feature Classification:** The extracted features were then categorized into behavioral or textual, facilitating variable selection according to each phase of the research.
5. **Best Parameter Search:** The best features and hyperparameters for training were selected, according to the needs of each phase of the research, applying optimization techniques to improve model performance.
6. **Model Training:** With the optimized features and hyperparameters, the resulting models were trained.
7. **Analysis of Results:** Finally, the obtained results were analyzed, allowing detailed evaluations and interpretations of the models' performance.

Figure 1 below graphically represents this methodology, providing an integrated overview of the research process employed in this work.

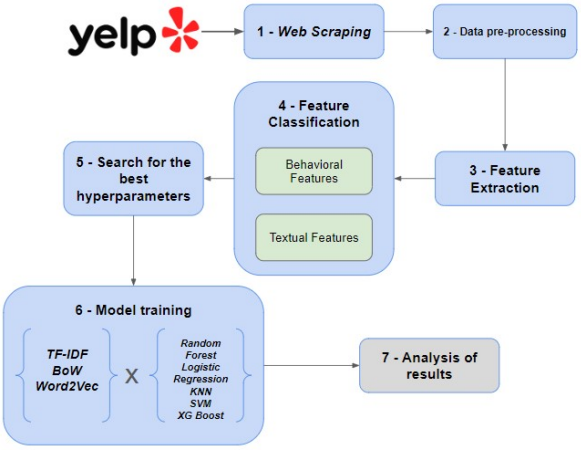

Figure 1. General structure of the work.

3.1 Data Collection

The dataset was collected using web scraping techniques with the *WebScraper* tool. The focus of data collection was on reviews related to restaurants and bars, with particular attention to reviews written in Portuguese. This focus on the food sector is due to the abundance of available data and the relevance of these establishments for online review analysis.

The choice of the *Yelp* platform as the primary data source is based on its unique features and applicability to machine learning algorithms. Yelp is notable for having a specific section on each establishment's page called 'non-recommended reviews'. This publicly visible area contains reviews filtered by the platform itself, and although visible to the public, are not recommended for users. The availability of these data allowed them to be considered 'false' for the purpose of this study, enabling the definition of a dataset with clear true and false labels. The relevance of this choice is corroborated by the study of [Mukherjee *et al*., 2013], which asserts the consistency of the Yelp filter in categorizing a review as false when it is indeed filtered by the platform. However, it is important to note that labeling as 'false' does not necessarily indicate a review was made with the intention to deceive. In some cases, the 'false' classification may represent a discrepancy between the review and the reported experience of

the product. Therefore, the term 'false' used here is a simplification, more appropriately considering these reviews as potentially misleading or inconsistent.

Despite a greater variety of information being available in true reviews, the selection focused on the common characteristics of both types, ensuring a fair and consistent approach in forming the dataset for model training. The collected information encompasses both textual and behavioral aspects of users. The extracted textual features include the content of the reviews, offering valuable insights into the writing style and language used by users. As for the behavioral characteristics, information was collected such as the number of friends of users on the platform, reflecting their level of social interaction; the total number of reviews made by each user, indicating their level of activity; and the number of photos posted, which can signal the user's visual engagement with the establishments. Additionally, the rating given by the user and the presence of a photo in the user's profile, elements that can influence the perception of credibility of the reviews, were captured. The date of posting of each review was also recorded, but this feature was not used in this study.

Figure 2 presents an illustrative example of a genuine review collected from a restaurant in São Paulo. Conversely, Figure 3 shows a case of a false review from the same establishment. Although both reviews received a rating of five, they exhibit significant differences. Additionally, Table 1 offers a comparative view of the initial dataset structure, displaying the features collected from each example as per Figure 2 and 3.

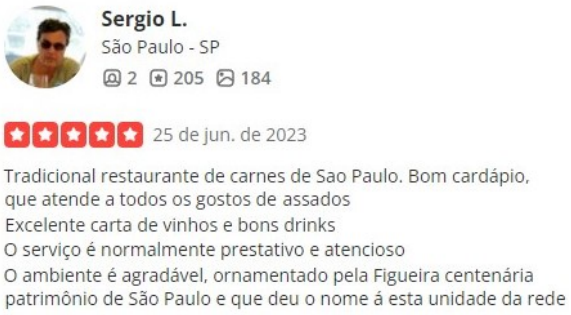

Figure 2. Information extracted from a true Yelp review.

Figure 3. Information extracted from a false Yelp review.

The differences observed between genuine and false reviews are striking and can be used as indicators in detecting inauthentic reviews. Firstly, the true review presents detailed and specific textual content, mentioning concrete aspects of the restaurant, such as the varied menu, the quality of wines and drinks, as well as the decoration and service. In contrast, the false review displays superficial and generic textual content, without detailing experiences or specific elements of the establishment.

Table 1. Comparison between a True and a False Review extracted from the Yelp platform.

True Review	False Review
Content: Traditional steakhouse in Sao Paulo. Good menu, catering to all tastes of grilled meats. Excellent wine list and good drinks. The service is usually helpful and attentive. The atmosphere is pleasant, decorated by the centennial fig tree, a heritage of São Paulo and that gave the name to this unit of the chain	**Content:** great food and service, highly recommend
Number of Friends: 2	**Number of Friends:** 0
Number of Reviews: 205	**Number of Reviews:** 2
Number of Photos: 184	**Number of Photos:** 0
Rating: 5	**Rating:** 5
User with Photo: Yes	**User with Photo:** Yes
Posting Date: 25/06/2023	**Posting Date:** 20/08/2023
False Review: No	**False Review:** Yes

Another relevant aspect is the user's interaction with the platform. In the true review, the user has a robust history, evidenced by a significant number of friends, reviews, and photos posted, suggesting consistent and authentic engagement with the platform. On the other hand, in the false review, the absence of friends and photos, along with an extremely low number of reviews, indicates an inactive or newly created profile, which may be a sign of inauthenticity.

These differences, combined with the general context of each review, provide valuable clues for identifying false reviews, significantly contributing to the accuracy of detection models based on data analysis and user characteristics.

In addition to the main dataset in Portuguese, a second dataset in English was collected. This English dataset comprised 26,692 true reviews and 14,567 false reviews of bars and restaurants in the USA. With the same structure as the Portuguese dataset, it underwent the same pre-processing procedures, except for *POS-Tagging*, as detailed in Section 3.2. The purpose of including this English dataset is to enable the validation of the final models developed for Portuguese.

3.2 Pre-processing

Pre-processing of textual data is a crucial step in linguistic data analysis, particularly in the context of machine learning. This phase begins with data standardization, which aims to homogenize formats and correct any inconsistencies. Standardization is fundamental to ensure that data are in a uniform format suitable for analysis, preventing errors in later stages [Speech, 2009].

During the pre-processing of the collected dataset, several actions were implemented to ensure its quality and accuracy. Initially, records without comments were removed, and null values were treated; moreover, a standardization of variables was carried out to uniformize the data. A critical step in preparing the datasets was implementing a linguistic filter with the *langdetect* library [Shuyo, 2010]. This filter ex-

clusively selected reviews in Portuguese for the Portuguese dataset and similarly, only reviews in English for the English dataset. This procedure proved vital, as preliminary analysis indicated the presence of reviews in various languages, suggesting reviews by foreigners in Brazilian and Portuguese establishments. Similarly, the English dataset also underwent this filter, ensuring linguistic consistency. In the final Portuguese dataset, 13,957 reviews were included, with 3,387 classified as false and 10,570 as true.

3.3 Feature Extraction

In the quest to understand and identify false reviews, this study is based on an in-depth analysis of user interactions with the platform and the content they share. Each piece of information plays a specific role in the dataset analysis, as detailed in Table 2 and described below.

The behavioral features (C), described by the number of friends (*qtd_friends*), the total number of reviews made by the user (*qtd_reviews*), the number of photos posted (*qtd_photos*), the rating given by the user (*rating*), and the presence of a photo in the user's profile (*user_has_photo*) offer a quantitative view of the user's activity on the platform.

The textual features (T), represented by the content of the reviews (*content*), are analyzed to extract language patterns and expressions that may be indicative of authenticity or falseness. To complement this analysis, the study incorporates behavioral *metafeatures* (MC), such as punctuation count (*punctuation_count*), capital letter count (*capital_count*), and total word count in the review (*word_count*), enabling a more detailed examination of the nuances in users' writing styles.

Table 2. Description of the types and characteristics of the dataset

Feature Name	Feature Type	Description
qtd_friends	C	Number of friends of the user on the platform.
qtd_reviews	C	Total number of reviews made by the user.
qtd_photos	C	Number of photos posted by the user.
rating	C	Rating given by the user to the establishment.
user_has_photo	C	Presence of a photo in the user's profile.
punctuation_count	MC	Count of punctuation in the review content.
capital_count	MC	Number of capital letters in the review content.
word_count	MC	Total number of words in the review.
content	T	Text of the review.
content_tagged	MT	Review with applied *POS-tagging*.

Furthermore, the removal of *stop words*, which are common words that generally do not contribute to the meaning of a text, such as prepositions and articles, was carried out. The removal of these words helps to reduce noise in the data, allowing machine learning algorithms to focus on words that are truly significant and informative. Additionally, the elimination of duplicate, incomplete data or data that does not add value to the proposed analysis was performed, to ensure that the dataset retains only the most pertinent and useful information [Kondrak, 2005].

Finally, after the removal of *stop words*, a textual *metafeature* (MT) (*content_tagged*) was added to the dataset, incorporating the grammatical classification of each word in the review, allowing for the observation of differences in meanings of the same words in different contexts. This *POS-tagger*, developed by [Inoue, 2019], was trained with the *NLTK* library and the *Mac-Morpho* corpus [Fonseca and Rosa, 2013], achieving an accuracy of 92.19%. The features are categorized into different types to facilitate analysis. In this process, the approach adopted was to concatenate the grammatical classification with the corresponding word, enriching the textual context. The behavioral features (C) refer to characteristics related to user behavior on the platform. The behavioral *metafeatures* (MC) are attributes derived from the behavioral features, providing additional insights. The textual features (T) are linked to the written content of the reviews. Finally, the textual *metafeatures* (MT) represent information extracted or derived from the text of the reviews.

3.4 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) conducted in this study addressed the comparison of characteristics between true and false reviews in the Portuguese dataset. To provide a more equitable analysis of feature distribution between true and false reviews, we applied density graphs, known as Kernel Density Estimates (KDE). These graphs offer a continuous and smooth view of data distribution, being useful in datasets with label imbalance, as described by [Silverman, 1986]. Moreover, KDE graphs allow for a clear observation of where there is a greater concentration of information by class. For example, suppose we observe that the peak of true reviews in relation to the number of friends was 0.02 on the y-axis for 0 friends, while for false reviews, the peak was 0.04 also for 0 friends. This suggests a higher density of false reviews for users without listed friends. It is important to emphasize that density values in a KDE graph represent probability estimates and not direct quantities. Thus, a higher value on the y-axis for false reviews indicates a tendency of higher probability of finding false reviews among users with 0 friends, but not necessarily double the number of true reviews, as per the example.

3.4.1 Feature Analysis

The exploratory analysis conducted in this study offers a deeper understanding of the behavioral characteristics of users on the Yelp platform for this dataset. In the first graph, Figure 4, we observe the distribution of the number of friends of users. Here, a higher prevalence of false reviews among users with fewer friends is notable, suggesting that a scarcity of social connections on the platform may be related to fraudulent behaviors. Figure 5 focuses on the number of reviews posted by users. This graph indicates that users with fewer prior reviews tend more to publish inauthentic reviews, highlighting a behavioral pattern where less active users on the platform are prone to engage in suspicious activities. Finally, the graph in Figure 6 analyzes the relationship between the number of photos posted by users and the authenticity of reviews. It is revealed that users who share few or no photos are more inclined to post suspect reviews, emphasizing that low visual engagement on the platform can be an indicator

of false reviews. Each of these aspects, analyzed individually, contributes to a more comprehensive understanding of the trends and characteristics associated with false reviews.

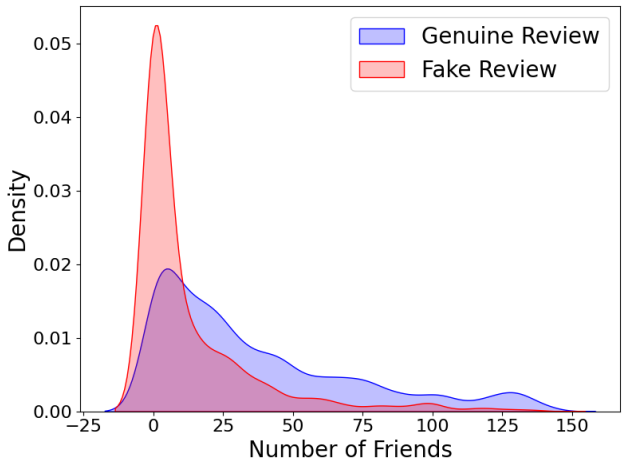

Figure 4. Density of the number of friends.

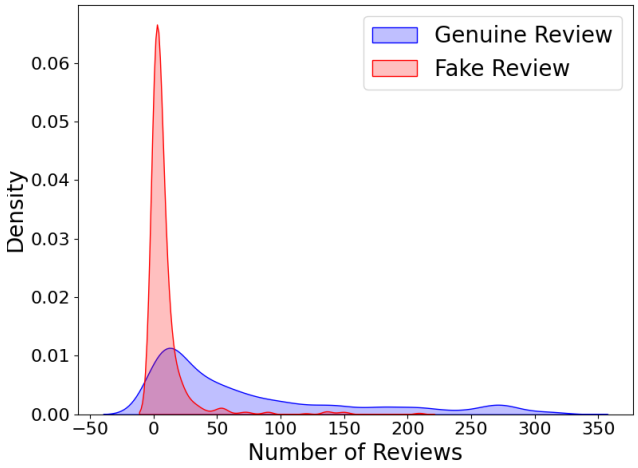

Figure 5. Density of the number of reviews.

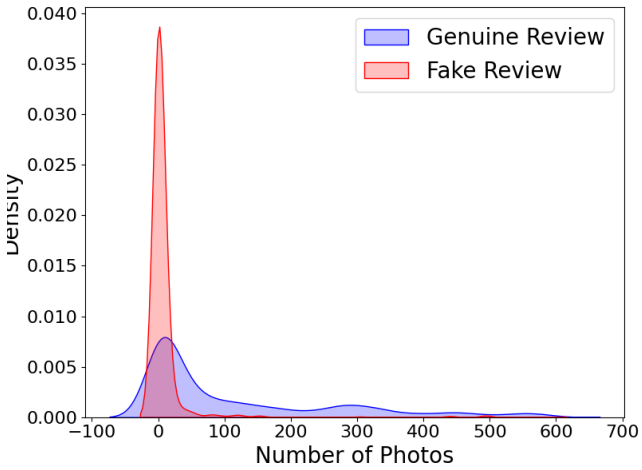

Figure 6. Density of the number of photos.

Figure 7 displays a density histogram comparing the presence of profile photos in true and false reviews. False reviews are more likely to be associated with users without a profile photo. This pattern suggests that profiles without photos may be more prone to producing inauthentic reviews. In contrast, true reviews tend to be from users with profile photos, indicating a higher level of authenticity.

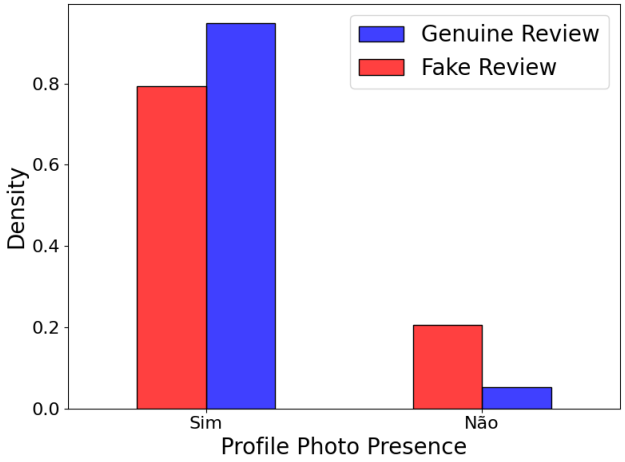

Figure 7. Density of users who have a profile photo.

The density histogram in Figure 8 highlights a relevant pattern in the distribution of review ratings. The analysis shows that reviews, especially false ones, tend to concentrate at the extremes of the scale, with significantly higher ratings at values 1 and 5. This behavior indicates that users who publish false reviews are more likely to assign extreme ratings - either the minimum or maximum score - compared to true reviews. Such a tendency for polarized ratings in false reviews may reflect an attempt to strongly influence the overall perception of an establishment, either positively or negatively.

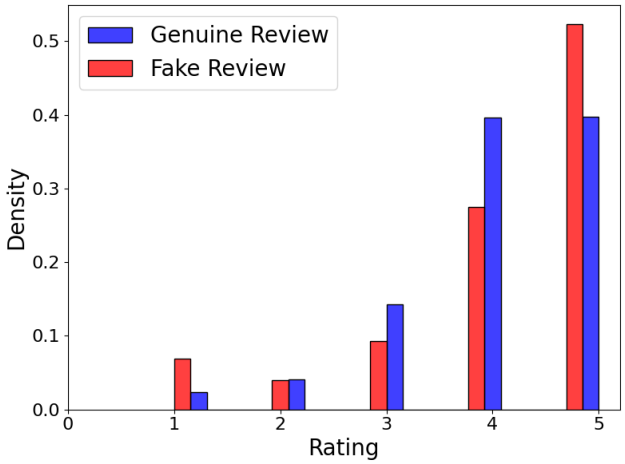

Figure 8. Density of review ratings.

In the correlation matrix presented in Figure 9, the textual metafeatures and the behavioral features are visually distinct in terms of correlation. The textual metafeatures, located in the lower right corner of the matrix, exhibit a strong positive correlation, approaching 1, reflecting their direct interrelation, as they are all intrinsically linked to the nature of textual content.

In contrast, the behavioral features, situated in the upper left corner of the matrix, also show a strong positive correlation within this group, indicating values close to 1. On the

other hand, there is low or no correlation between the behavioral features and the textual metafeatures, with values close to 0, indicating that these variables tend to operate independently of each other. This reinforces the lack of significant relationship between these two sets of variables. Notably, none of the relationships in the matrix presents a strongly negative correlation, approaching -1, highlighting that the variables, for the most part, do not inversely influence each other significantly.

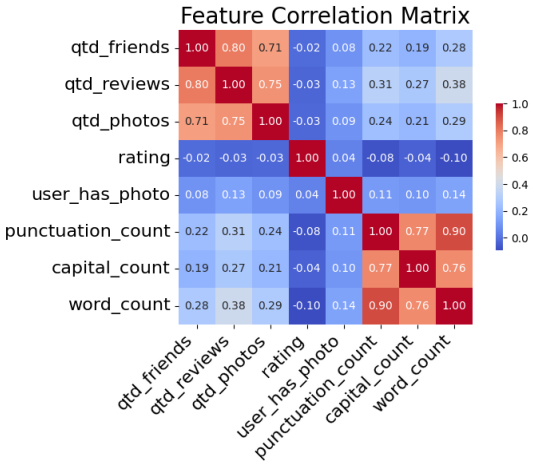

Figure 9. Correlation Matrix between Features.

3.4.2 Metafeature Analysis

The differences in the density graphs for the textual metafeatures offer other perspectives on the discrepancies between true and false reviews. This analysis focuses on the amount of punctuation, words, and capital letters in each review, metafeatures directly related to the textual content.

Figure 10 groups several density graphs, each exploring a different textual metafeature. In Figure 10 (a), we observe the density of the amount of punctuation used in reviews. This graph shows that, although there are differences between true and false reviews, the discrepancy is not as marked as in other behavioral characteristics. False reviews tend to use less punctuation, which may be a subtle indicator but not a clear differentiator. In Figure 10 (b), the analysis of the number of words in reviews reveals that false reviews are generally shorter than true ones. This trend, although notable, does not provide a distinctive enough pattern to be used as the sole criterion for identifying false reviews. Finally, in Figure 10 (c), the distribution of the use of capital letters is examined. Here, we see a tendency for lower use of capital letters in false reviews. Although this is an observable difference, it is not so pronounced as to act as a conclusive indicator by itself.

This joint analysis of the textual metafeatures illustrates that, despite there being subtle variations between true and false reviews, they are not as pronounced as the differences found in the behavioral characteristics. This suggests that for effective detection of false reviews, it is important to combine textual and behavioral features. However, care must be taken in the selection of textual metafeatures for the model, avoiding the inclusion of variables with limited distinctive relevance, to maintain the effectiveness of the detection model.

4 Approach and Model Selection

In this study, focused on identifying false reviews in Portuguese, three main approaches were explored: (i) based exclusively on the behavioral characteristics of users; (ii) exploring only the textual content of the reviews; and (iii) combining both approaches.

Figure 11 illustrates the adopted model selection diagram. In the steps indicated by the blue boxes, *GridSearchCV* [Bird *et al*., 2009] was used to determine the most relevant behavioral features, the most effective n-grams for each combination of vectorizer and classifier, and the ideal hyperparameters for classifiers focused on textual content. At this stage of the research, the Portuguese dataset was balanced, aligning it by the minority class. In the study of [Sihombing and Fong, 2019], a similar situation is reported, where results obtained with unbalanced data were considerably inferior compared to a balanced dataset. As a result, the dataset in this study now consists of 3387 true reviews and 3387 false reviews. To achieve this balance, a number of surplus true reviews were excluded. The selection of data to be removed was done by fixing the random state parameter to 42 in the *sample* method of the *Pandas dataframe* [Pedregosa *et al*., 2011], thus ensuring the reproducibility of the tests. In all steps of searching for better parameters, 5-fold cross-validation was used.

The choice of the *K-Nearest Neighbors (KNN)*, *Random Forest*, *XGBoost*, *Logistic Regression*, and *Support Vector Machine (SVM)* algorithms to identify false reviews was based on the efficacy described in the literature review of this study. In the review, the authors demonstrated the success of algorithms such as *Random Forest* [Birim *et al*., 2022], *AdaBoost* [Barbado *et al*., 2019], *KNN* [Elmogy *et al*., 2021], and *XGBoost* [Sihombing and Fong, 2019], and also tested *SVM* [Birim *et al*., 2022] and *Logistic Regression* [Elmogy *et al*., 2021] in classifying false reviews.

4.1 Selection of the Best Behavioral Features

The process of searching for the best behavioral features per classifier involved two steps. In the initial stage, the *Random Forest*, *Logistic Regression*, *SVM*, *KNN*, and *XGBoost* classifiers were optimized using *GridSearchCV* [Bird *et al*., 2009], focusing on behavioral features (C + MC). This process sought the best settings for each classifier, based on a specific set of predefined options. The optimized configurations for each classifier are detailed in Table 3.

Table 3. Best Hyperparameter Settings per Classifier.

Classifier	Hyperparameters
Random Forest	max_depth: None, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 1000
Logistic Regression	C: 250, penalty: l2, solver: lbfgs
KNN	n_neighbors: 13, metric: manhattan, weights: uniform, p: 1
SVM	C: 100, gamma: scale, kernel: rbf, max_iter: 5000
XGBoost	learning_rate: 0.01, max_depth: 11, n_estimators: 500

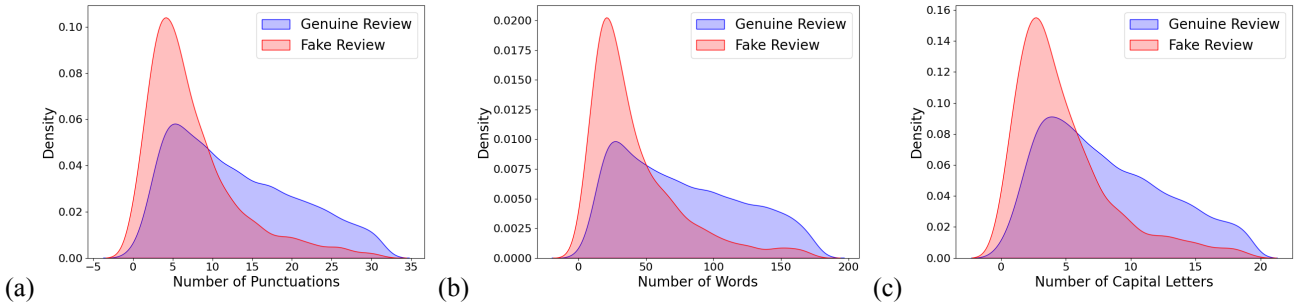

Figure 10. Density of the behavioral metafeatures.

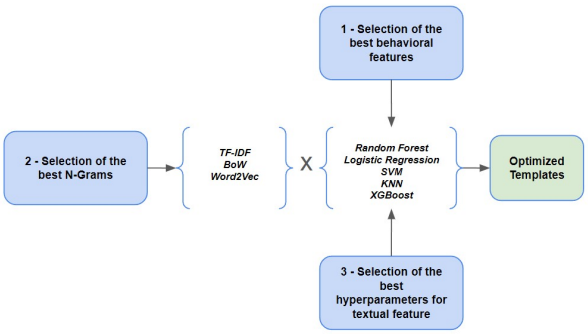

Figure 11. Model selection diagram.

In the study of [Elmogy *et al*., 2021], the authors observed that the sentiment score of reviews negatively influenced the model's performance. Therefore, aiming to identify if the removal of any feature from the dataset of this study would negatively influence the model, an iterative process was conducted in the second stage of selecting the best behavioral features. This process consisted of verifying the importance of each feature in relation to each classifier.

After establishing the best hyperparameters for the behavioral features (C) and behavioral metafeatures (MC), an iterative cycle of training and testing was proceeded with, using only these features. In each cycle, the F1-score was evaluated and the least important feature, identified by permutation techniques [Breiman, 2001], was removed. This process continued until only two features remained, thus identifying the most efficient feature combinations for each classifier. The feature sets that maximized the efficacy of each classifier can be found in Table 4.

Table 4. Best Behavioral Features per Classifier

Classifier	Average F1-Score	Best Features
Random Forest	**88.79%**	qtd_friends, qtd_reviews, qtd_photos
Logistic Regression	81.50%	qtd_friends, qtd_reviews, user_has_photo, word_count
KNN	84.94%	qtd_friends, qtd_reviews, qtd_photos
SVM	82.60%	qtd_friends, qtd_reviews, qtd_photos, punctuation_count, word_count
XGBoost	88.68%	qtd_friends, qtd_reviews, qtd_photos

Upon examining the most important features for each classifier, as illustrated in Table 4, a predominance of qtd_friends, qtd_reviews, and qtd_photos is observed. This tendency suggests that these features are crucial in optimizing the F1-score performance of the model. Such an observation aligns with the exploratory analysis, which had already pointed to significant differences in the behavior of these features. Furthermore, Table 4 highlights the ability of the selected models to classify reviews with an average F1-score above 80% based solely on behavioral features. The Random Forest and XGBoost stand out for their best efficacy, with an average F1-score of 88.79%, and 88.68% respectively. These results are consistent with those found in the studies of [Birim *et al*., 2022] and [Barbado *et al*., 2019], which reach around 80% with behavioral features.

4.2 N-Gram Selection

After defining the best features for each classifier, the next phase of the study focused exclusively on textual content. As adopted in the work of [Elmogy *et al*., 2021], where the authors employed the use of N-Grams and achieved improvements in model efficacy, this stage's main goal is to identify the most effective N-gram for each combination of classifier and vectorizer. To achieve this, the GridSearchCV method was employed to establish the N-grams that provided the best performance in terms of average F1-score.

Table 5 summarizes the results of this phase, presenting the optimal N-gram configurations for each vectorization scenario using TF-IDF and BoW. With the exception of the Word2Vec vectorizer, which is limited to uni-grams due to its intrinsic nature. Word2Vec is specifically designed to learn vector representations of individual words in their specific contexts, rather than focusing on sequences of words, a predominant feature in N-gram methods.

On the other hand, for other vectorizers, such as TF-IDF and BoW, various combinations of N-grams were evaluated, including six different variations: (1, 1), (1, 2), (1, 3), (2, 2), (2, 3), and (3, 3). The choice of these N-grams was made with the intent to explore the efficacy of different sizes of word sequences in capturing patterns in the textual content of reviews.

The relevance of this stage of the study lies in the understanding that the efficacy in identifying false reviews is influenced not only by the selection of features but equally by how textual content is processed and examined. The chosen N-gram configurations aim not only to enhance the performance of each model but also to optimize the training process, avoiding less efficient options.

Table 5. Best N-grams per Classifier and Vectorizer.

Vectorizer	Classifier	Best n-gram
TF-IDF	Random Forest	(1, 3)
	Logistic Regression	(1, 1)
	KNN	(1, 1)
	SVM	(1, 1)
	XGBoost	(1, 2)
BoW	Random Forest	(1, 2)
	Logistic Regression	(1, 3)
	KNN	(1, 1)
	SVM	(1, 2)
	XGBoost	(1, 3)

4.3 Hyperparameter Selection for Textual Feature

In this phase of the study, the central objective was to identify the most suitable hyperparameters of the classifiers for processing textual content. To achieve the ideal configuration of the classifiers, the GridSearchCV method was used with vectorizers parameterized with the best N-Grams selected earlier.

This meticulous process resulted in precise adjustments for each classifier in different vectorization contexts, reflecting the complexity and specificity of the challenge in identifying false reviews. The detailed results of this process, including the final adjusted hyperparameters for each classifier, are presented in Table 6. It is important to note that a lower performance was expected at this stage, especially considering the evidence in the literature indicating that the exclusive use of textual features is not the most promising approach in detecting false reviews [Barbado *et al*., 2019].

4.4 Selected Models

Finally, a selection of 15 distinct models was made, based on their proven efficacy in each step of the parameter definition process. This process involved a detailed analysis of possible combinations of behavioral and textual features, as well as the careful selection of appropriate N-grams for each vectorizer. The choice of these models was guided by the goal of maximizing efficacy in identifying false reviews, taking into account the particularities of the Portuguese language and the behavioral nuances of users.

The details of the best configurations for each combination of vectorizer and classifier were compiled in Table 7, summarizing the strategic choices made. This table presents a clear view of the most efficient combinations between vectorizers such as TF-IDF, BoW, and Word2Vec, and classifiers, including Random Forest, Logistic Regression, KNN, SVM, and XGBoost. This meticulous process of selecting the best hyperparameters and features was crucial to evaluate the most promising models, given the vast range of possible combinations.

5 Results Analysis

The tests were conducted using Anaconda version 2.5.1, employing the *Spyder* IDE 5.4.3 and *Python* version 3.11.5. A

Table 6. Performance of each Classifier and Vectorizer with the Best Hyperparameters and N-Grams.

Vectorizer	Classifier	Average F1-Score	Hyperparameters
TF-IDF	Random Forest	73.58%	max_depth: 1000, min_samples_leaf: 1, min_samples_split: 3, n_estimators: 500
	Logistic Regression	69.60%	C: 1, penalty: l2, solver: newton-cg
	KNN	69.07%	metric: euclidean, n_neighbors: 17, weights: uniform
	SVM	71.50%	C: 100, gamma: auto, kernel: rbf, max_iter: 2000
	XGBoost	72.38%	learning_rate: 0.01, max_depth: 15, n_estimators: 1000, min_child_weight: 10
BoW	Random Forest	73.63%	max_depth: 1000, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 1000
	Logistic Regression	72.01%	C: 1, penalty: l2, solver: newton-cg, max_iter: 2000
	KNN	67.08%	metric: euclidean, n_neighbors: 3, weights: uniform
	SVM	73.29%	C: 100, gamma: auto, kernel: rbf, max_iter: 5000
	XGBoost	66.86%	learning_rate: 0.01, max_depth: None, n_estimators: 500, min_child_weight: 1
Word2Vec	Random Forest	65.27%	max_depth: None, min_samples_leaf: 1, min_samples_split: 3, n_estimators: 1000
	Logistic Regression	61.67%	C: 500, solver: newton-cg, penalty: l2, max_iter: 2000
	KNN	59.11%	metric: euclidean, n_neighbors: 17, weights: uniform
	SVM	66.61%	C: 50, gamma: auto, kernel: poly, max_iter: 1000
	XGBoost	64.48%	learning_rate: 0.01, max_depth: 9, n_estimators: 500, min_child_weight: 10

crucial aspect of the study was the analysis of the impact of balanced and unbalanced datasets [Shah *et al*., 2018], considering the inherent imbalance in the Portuguese dataset. The focus of the final models was the integration of effective behavioral characteristics with optimized textual content for each classifier. The balanced Portuguese dataset was used as the basis for this integration, resulting in optimized combinations of behavioral features, N-grams, and hyperparameters.

Parallelly, during the data collection in Portuguese, a corresponding English dataset for bars and restaurants was collected, processed identically to the Portuguese set, to evaluate the impact of language change when applied to the models. The datasets evaluated in the selected models are as follows:

1. **Balanced Portuguese dataset (i):** Comprised of 3,387 true reviews and 3,387 false reviews.
2. **Balanced English dataset (ii):** This set consists of 3,387 true reviews and 3,387 false reviews.
3. **Balanced English dataset (iii):** Comprised of 14,597 true reviews and 14,597 false reviews.

Table 7. Resulting Models per Combination of Vectorizer and Classifier

Vectorizer	Classifier	Best N-gram	Best Behavioral Features
TF-IDF	Random Forest	(1, 3)	qtd_friends, qtd_reviews, qtd_photos
	Logistic Regression	(1, 1)	qtd_friends, qtd_reviews, user_has_photo, word_count
	KNN	(1, 1)	qtd_friends, qtd_reviews, qtd_photos
	SVM	(1, 1)	qtd_friends, qtd_reviews, qtd_photos, punctuation_count, word_count
	XGBoost	(1, 2)	qtd_friends, qtd_reviews, qtd_photos
BoW	Random Forest	(1, 2)	qtd_friends, qtd_reviews, qtd_photos
	Logistic Regression	(1, 3)	qtd_friends, qtd_reviews, user_has_photo, word_count
	KNN	(1, 1)	qtd_friends, qtd_reviews, qtd_photos
	SVM	(1, 2)	qtd_friends, qtd_reviews, qtd_photos, punctuation_count, word_count
	XGBoost	(1, 3)	qtd_friends, qtd_reviews, qtd_photos
Word2Vec	Random Forest	(1, 1)	qtd_friends, qtd_reviews, qtd_photos
	Logistic Regression	(1, 1)	qtd_friends, qtd_reviews, user_has_photo, word_count
	KNN	(1, 1)	qtd_friends, qtd_reviews, qtd_photos
	SVM	(1, 1)	qtd_friends, qtd_reviews, qtd_photos, punctuation_count, word_count
	XGBoost	(1, 1)	qtd_friends, qtd_reviews, qtd_photos

4. **Unbalanced Portuguese dataset (iv):** This set is formed by 10,570 true reviews and 3,387 false reviews.
5. **Unbalanced English dataset (v):** This set includes 10,570 true reviews and 3,387 false reviews.

After selecting the final models, the best numerical features for each classifier were standardized and transformed into sparse matrices to ensure compatibility with the vectorized textual matrices. Thus, textual and behavioral features were unified. These numerical matrices were then horizontally concatenated to the vectorized text matrices. This combination resulted in final matrices where the initial columns represent text features, while the last columns reflect the behavioral features of users. This procedure was consistently applied to all datasets, allowing the classification models to simultaneously process textual and behavioral information, which was essential for the final analysis of the models.

For the five datasets, models were individually trained and tested, maintaining exclusively the hyperparameters and features selected in the model optimization phase. This optimization phase employed the base dataset (i) in the search and definition of these parameters.

Figure 12 displays the top three F-Score performances for each dataset, organized by vectorizer and classifier. Each bar in the graph is accompanied by a marker at the top (error bar), representing the minimum and maximum performance variation of that model during cross-validation.

For the dataset that underpinned the search for the best parameters (i), more stable and homogeneous results were observed, highlighting *KNN* and *XGBoost* with all vectorizer options, surpassing the 85% F-Score mark. For the other models, despite scoring below 85%, they performed well compared to other datasets, probably because each model was optimized for that same dataset. Additionally, the F-Score variation in the models was low, with the exception of *SVM*, which tested along with *Word2Vec* showed abnormal fluctuations.

For the dataset (ii), which is in English and balanced in the same amount as dataset (i), the results stood out for *Logistic Regression* using *TF-IDF* and *BoW*, achieving about 83% F-Score. In this scenario, even with the general decrease in model performance, the best results approached those observed in dataset (i), maintaining consistency in cross-validation variation. Similarly, *SVM* showed instability when applied with *Word2Vec* and *TF-IDF*.

6 Analysis of Results

The tests were conducted using Anaconda version 2.5.1, with the *Spyder* IDE 5.4.3 and *Python* version 3.11.5. A crucial aspect of the study was the analysis of the impact of balanced and unbalanced datasets [Shah *et al*., 2018], given the inherent imbalance in the Portuguese dataset. The focus of the final models was the integration of effective behavioral characteristics with optimized textual content for each classifier. The balanced Portuguese dataset was used as the foundation for this integration, resulting in optimized combinations of behavioral features, N-grams, and hyperparameters.

Concurrently, during the data collection in Portuguese, a corresponding English dataset for bars and restaurants was gathered, processed identically to the Portuguese set, to assess the impact of language change when applied to the models. The datasets evaluated in the selected models are as follows:

1. **Balanced Portuguese dataset (i):** Comprising 3,387 true reviews and 3,387 false reviews.
2. **Balanced English dataset (ii):** Consisting of 3,387 true reviews and 3,387 false reviews.
3. **Balanced English dataset (iii):** Comprising 14,597 true reviews and 14,597 false reviews.
4. **Unbalanced Portuguese dataset (iv):** Formed by 10,570 true reviews and 3,387 false reviews.
5. **Unbalanced English dataset (v):** Including 10,570 true reviews and 3,387 false reviews.

After selecting the final models, the best numerical features for each classifier were standardized and transformed into sparse matrices to ensure compatibility with the vectorized textual matrices. Thus, textual and behavioral features were unified. These numerical matrices were then horizontally concatenated to the vectorized text matrices. This combination resulted in final matrices where the initial columns represent text features, while the last columns reflect the behavioral features of users. This procedure was consistently applied to all datasets, allowing the classification models to

simultaneously process textual and behavioral information, essential for the final analysis of the models.

For the five datasets, models were individually trained and tested, maintaining exclusively the hyperparameters and features selected in the model optimization phase. This optimization phase employed the base dataset (i) in the search and definition of these parameters.

Figure 12 displays the top three F-Score performances for each dataset, organized by vectorizer and classifier. Each bar in the graph is accompanied by a marker at the top (error bar), representing the minimum and maximum performance variation of that model during cross-validation.

For dataset (i), which underpinned the search for the best parameters, more stable and homogeneous results were observed, highlighting *KNN* and *XGBoost* with all vectorizer options, surpassing the 85% F-Score mark. For the other models, despite scoring below 85%, they performed well compared to other datasets, likely because each model was optimized for that same dataset. Additionally, the F-Score variation in the models was low, with the exception of *SVM*, which, when tested along with *Word2Vec*, showed abnormal fluctuations.

For dataset (ii), in English and balanced to the same extent as dataset (i), the results stood out for *Logistic Regression* using *TF-IDF* and *BoW*, achieving about 83% F-Score. In this scenario, even with the general decrease in model performance, the best results approached those observed in dataset (i), maintaining consistency in cross-validation variation. Similarly, *SVM* showed instability when applied with *Word2Vec* and *TF-IDF*.

For dataset (iii), which is also in English and balanced but more than four times larger than dataset (i), the results were notable for *Logistic Regression* with *BoW* and *TF-IDF*, and *XGBoost* with *TF-IDF*, all around the 85% F-Score mark. This dataset also showed a larger performance variation among classifiers compared to dataset (i), again due to dataset (i) being the base for parameter optimization. Additionally, the cross-validation variation in this dataset was lower, due to the larger amount of data. It was observed that *SVC*, despite being stable with the 3 vectorizers, had the worst performance in this dataset.

Observing the behavior of the models in the unbalanced dataset (iv), in Portuguese, there was a considerable worsening compared to dataset (i). The difference between the best F-Score of dataset (iii) and the best F-Score of dataset (i) reached 15.5%. Moreover, for the best models in dataset (iii), there was notable stability in cross-validation as in dataset (i). In this dataset, *SVM* also behaved abnormally when run with *Word2Vec*, nullifying some F-Scores during cross-validation.

Finally, dataset (v), with the same dimension as dataset (iv), shows a more intense performance difference among classifiers. However, as in scenario (iii) in English, *Logistic Regression* with *BoW* stood out, performing better than the *KNN* in dataset (iii). Except for this model, the others that stood out were *Logistic Regression* and *XGBoost* with *TF-IDF*, approaching 70%.

The impact of data imbalance on model performance in the context of this work is notable, and observing Figure 12, the change of language brings general differences in the models. However, when selecting a specific model, consistent performance is observed, reinforcing the reliability of the indicators of false reviews in both languages. In the English datasets, *Logistic Regression* showed the best results. In the Portuguese datasets, the *KNN* and *XGBoost* models stood out in efficacy. Analyzing all scenarios and Table 8, *XGBoost* is present as the third best in all scenarios, highlighting its robustness in the tested context. On the other hand, *SVM*, which showed several instabilities in the tests, seems sensitive to the amount of data available for training. Specifically, in dataset (iii), which has the largest volume of data, *SVM* showed more stable results. This suggests that a larger amount of data is needed to optimize the performance of this model.

The balancing of the datasets positively influenced the performance of the models, and the larger dimension of the datasets contributed to the reduction in metric variation, emphasizing the importance of a substantial volume of data for efficiency in detecting false online reviews. Moreover, revisiting Table 4, which shows the performance of the models considering only behavioral features, it was found that the best were *Random Forest* and *XGBoost*, with 88.79% and 88.68% respectively with the base dataset (i), and in models considering textual features, none surpassed this mark, aligning with the findings of the study by [Birim *et al*., 2022], where the authors observed that a feature based on textual content reduced model performance. Another interesting conclusion is that in Table 8, which shows the top three models for each dataset, in the unbalanced ones, despite having cases with high accuracy and precision, the recall is significantly reduced, indicating that many of the reviews that are actually false are not being correctly identified by the model.

7 Conclusion

This study addressed the identification of fake reviews in Portuguese from the *Yelp* platform, employing a combination of machine learning techniques and analysis of textual and behavioral characteristics. The effectiveness of the developed models was demonstrated, evidencing the ability to distinguish between authentic and inauthentic reviews using machine learning and natural language processing in a linguistically under-explored context, the Portuguese language. Notably, the results achieved were equal to or better than those reviewed in the literature using English language data with the same purpose: identifying fake reviews through machine learning and natural language processing. The results are in line with the findings of the studies by [Elmogy *et al*., 2021], [Birim *et al*., 2022], [Barbado *et al*., 2019], and [Sihombing and Fong, 2019], which also highlight the importance of behavioral features over textual features, a crucial aspect to achieve the presented results. Additionally, the study by [Sihombing and Fong, 2019] reported the problems caused by data imbalance, a phenomenon that was also observed in this study.

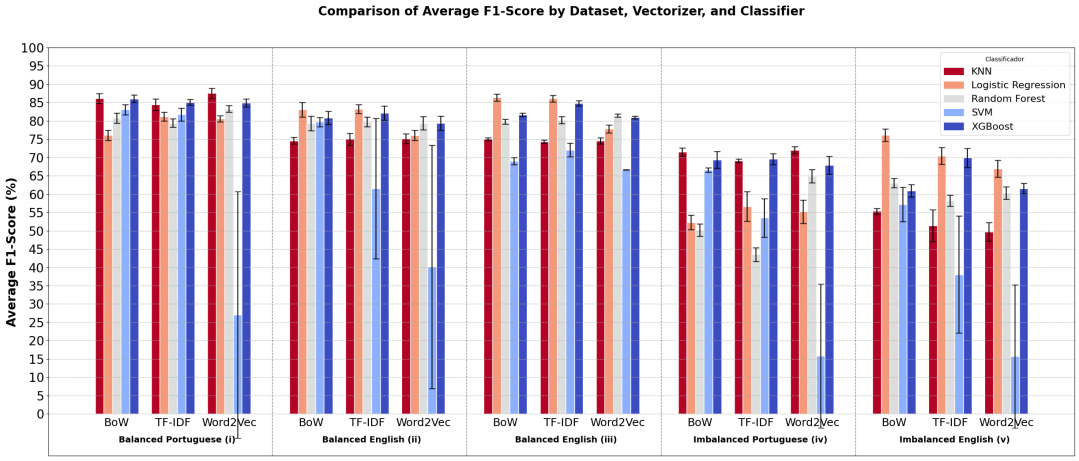

Figure 12. F1-Score for Each Scenario, Vectorizer, and Classifier

Table 8. Performance Comparison by Scenario, Vectorizer, and Classifier

Dataset	Size	Vectorizer	Classifier	Accuracy	Precision	Recall	F1-Score
Balanced Portuguese (i)	T: 3387 | F: 3387	Word2Vec	KNN	86.54%	81.54%	**94.48%**	**87.53%**
Balanced Portuguese (i)	T: 3387 | F: 3387	BoW	KNN	85.09%	80.46%	92.71%	86.15%
Balanced Portuguese (i)	T: 3387 | F: 3387	BoW	XGBoost	85.27%	81.70%	90.94%	86.06%
Balanced English (ii)	T: 3387 | F: 3387	TF-IDF	Logistic Regression	84.31%	89.56%	77.68%	83.19%
Balanced English (ii)	T: 3387 | F: 3387	BoW	Logistic Regression	82.95%	82.45%	83.73%	83.07%
Balanced English (ii)	T: 3387 | F: 3387	TF-IDF	XGBoost	81.74%	80.36%	84.06%	82.16%
Balanced English (iii)	T: 14597 | F: 14597	BoW	Logistic Regression	86.26%	85.96%	86.69%	86.32%
Balanced English (iii)	T: 14597 | F: 14597	TF-IDF	Logistic Regression	86.64%	89.71%	82.77%	86.10%
Balanced English (iii)	T: 14597 | F: 14597	TF-IDF	XGBoost	84.47%	82.79%	87.04%	84.86%
Unbalanced Portuguese (iv)	T: 10570 | F: 3387	Word2Vec	KNN	86.29%	71.31%	72.78%	72.03%
Unbalanced Portuguese (iv)	T: 10570 | F: 3387	BoW	KNN	84.08%	63.21%	82.40%	71.53%
Unbalanced Portuguese (iv)	T: 10570 | F: 3387	TF-IDF	XGBoost	85.58%	71.30%	67.96%	69.58%
Unbalanced English (v)	T: 10570 | F: 3387	BoW	Logistic Regression	**89.75%**	87.70%	67.20%	76.09%
Unbalanced English (v)	T: 10570 | F: 3387	TF-IDF	Logistic Regression	88.80%	**97.83%**	55.06%	70.45%
Unbalanced English (v)	T: 10570 | F: 3387	TF-IDF	XGBoost	86.56%	76.49%	64.39%	69.91%

8 Conclusion

This study addressed the identification of false reviews in Portuguese from the *Yelp* platform, employing a combination of machine learning techniques and analysis of textual and behavioral characteristics. The efficacy of the developed models was demonstrated, highlighting the ability to distinguish between authentic and inauthentic reviews using machine learning and natural language processing techniques in a linguistically underexplored context, the Portuguese language. Notably, the results obtained were equal to or better than those reviewed in the literature using data in the English language.

The findings of this study have relevant implications for the scientific community and professionals involved in online fraud detection. They offer new perspectives on the combination of textual and behavioral characteristics and emphasize the need to use balanced and representative datasets for training. Additionally, the results achieved with the validation of methods using English data are in line with the studies reviewed in this work, also contributing to the reliability of the results obtained with the Portuguese data and ensuring the robustness of the conclusions of this work.

This study indicates paths for future investigations, recommending the exploration of sentiment analysis and the improvement of lemmatization in Portuguese. It also suggests researching other approaches to using *POS-Tagging* and investigating other methods for extracting features from textual content to improve the performance of the validated models. These directions aim to enhance accuracy in identifying false reviews and expand the applicability of the models in various contexts and languages.

As a practical proposal, considering that behavioral characteristics were crucial in achieving good results, it is recommended that e-commerce platforms continuously improve their mechanisms for recording users' behavioral data, in addition to other information that facilitates fraud identification.

Finally, the importance of interdisciplinary collaboration and the continuation of research in this area is highlighted, given its growing relevance in the contemporary digital world. It is hoped that this study will inspire future investigations and significantly contribute to the field of natural language processing and online fraud detection.

Availability of data and materials

The entire project, including the *dataset*, results, and *web scraping* scripts used, is publicly available on *GitHub* as a contribution to the community. Interested parties can access the repository for more details and explore the project's resources:

- **GitHub Repository:** `https://github.com/lucaspercisi/yelp-fake-reviews-ptbr`

The repository includes:

- Dataset in Portuguese and English used in this work
- *Scripts* for *web scraping* specifically for the *Yelp* site used with the *WebScrapper* tool.
- Source code used in the development of this work.
- Files with the results of the training scenarios.

References

Barbado, R., Araque, O., and Iglesias, C. A. (2019). A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management*, 56(4):1234–1244.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Birim, Ş. Ö., Kazancoglu, I., Mangla, S. K., Kahraman, A., Kumar, S., and Kazancoglu, Y. (2022). Detecting fake reviews through topic modelling. *Journal of Business Research*, 149:884–900.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Choo, E., Yu, T., and Chi, M. (2015). Detecting opinion spammer groups through community discovery and sentiment analysis. In *Data and Applications Security and Privacy XXIX: 29th Annual IFIP WG 11.3 Working Conference, DBSec 2015, Fairfax, VA, USA, July 13-15, 2015, Proceedings 29*, pages 170–187. Springer.

Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., and Najada, H. A. (2015). Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2(1):1–24.

Cruz, B. d. P. A., Pimenta, S., Dutton, A., and Ross, S. D. (2020). Fake on-line reviews em restaurantes: intenção de boicote ou intenção de buycott de telespectadores do programa pesadelo na cozinha?

Elmogy, A. M., Tariq, U., Ammar, M., and Ibrahim, A. (2021). Fake reviews detection using supervised machine learning. *International Journal of Advanced Computer Science and Applications*, 12(1).

Fonseca, E. and Rosa, J. L. G. (2013). Mac-morpho revisited: Towards robust part-of-speech tagging. In *Proceedings of the 9th Brazilian symposium in information and human language technology*.

Inoue, M. (2019). Pos-tagger-portuguese-nltk. `https://github.com/inoueMashuu/POS-tagger-portuguese-nltk`. Acessado em: 11 de novembro de 2023.

Jindal, N. and Liu, B. (2007). Analyzing and detecting review spam. In *Seventh IEEE international conference on data mining (ICDM 2007)*, pages 547–552. IEEE.

Kondrak, G. (2005). N-gram similarity and distance. In *International symposium on string processing and information retrieval*, pages 115–126. Springer.

Lima, H. d. C. S. d. C. (2013). Detectando avaliações spam em uma rede social baseada em localização.

Luca, M. and Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):3412–3427.

Mukherjee, A., Venkataraman, V., Liu, B., and Glance, N. (2013). What yelp fake review filter might be doing? In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 409–418.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al*. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Sandulescu, V. and Ester, M. (2015). Detecting singleton review spammers using semantic similarity. In *Proceedings of the 24th international conference on World Wide Web*, pages 971–976.

Schuch, F. A. (2013). Detecção automática de spams de opinião em avaliações de produtos na língua portuguesa.

Shah, R., Khemani, V., Azarian, M., Pecht, M., and Su, Y. (2018). Analyzing data complexity using metafeatures for classification algorithm selection. In *2018 Prognostics and System Health Management Conference (PHM-Chongqing)*, pages 1280–1284. IEEE.

Shuyo, N. (2010). Language detection library for java.

Sihombing, A. and Fong, A. C. M. (2019). Fake review detection on yelp dataset using classification techniques in machine learning. In *2019 international conference on contemporary computing and informatics (IC3I)*, pages 64–68. IEEE.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Speech, J. D. M. J. (2009). Language processing: An introduction to natural language processing. *Computational Linguistics, and Speech Recognition*, 2.

Valdivia, A., Hrabova, E., Chaturvedi, I., Luzón, V. M., Troiano, L., Cambria, E., and Herrera, F. (2019). Inconsistencies on tripadvisor reviews: A unified index between users and sentiment analysis methods. *Neurocomputing*, 353:3–16.