NOVA IMS Information Management School

MDSAA

Master's Degree Program in

Data Science and Advanced Analytics

Automated Classification of Journalistic Texts

Developing a User-Centric Model

Tomás Costa França de Moura Vicente

Project Work

presented as partial requirement for obtaining a master's degree in data science and advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

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

Automated Classification of Journalistic Texts

Developing a User-Centric Model

by

Tomás Costa França de Moura Vicente

Project Work presented as a partial requirement for obtaining a master's degree in data science and advanced Analytics, with a specialization in Business Analytics.

Supervised by

Nuno António, PhD, NOVA IMS

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 2024

DEDICATION

I dedicate this work to my father.

ACKNOWLEDGMENTS

I would like to thank my supervisor for his guidance in orienting my project and highlighting his availability to help.

ABSTRACT

The journalism industry has undergone significant changes due to advances in technology. These changes have increased competition, making the use of data critical to understand user needs. One effective approach is to use clustering and classification techniques to segment users and gain deeper insights into their preferences and behaviors. This enables the creation of targeted content strategies that increase user engagement and satisfaction. This study aims to develop a classification model for journalistic content in text format, with a particular focus on understanding and meeting the needs of consumers in the digital age. This study employs data from the Global Media Group to apply the user needs model, categorizing news into eight distinct user need clusters. Machine Learning (ML) models, including Recurrent Neural Networks, Logistic Regression, and Random Forest, were evaluated with different word embedding techniques. Despite the challenges associated with data imbalance and the limitations of computational resources, the Logistic Regression model achieved the highest overall accuracy of 0.7. Results indicate that the introduction of biased manual classifications has a significant impact on the performance of the model, suggesting that future research should employ automated methods to ensure data consistency. The findings demonstrate the potential of ML in optimizing journalistic content classification, offering valuable insights for enhancing audience engagement and driving strategic decisions within media organizations. This study contributes to both academic and industry knowledge, emphasizing the importance of high-quality, balanced data and ethical considerations in the application of ML.

KEYWORDS

Text classification; Machine learning; Journalistic content; User needs; Data imbalance; Artificial Intelligence

Sustainable Development Goals (SDG):

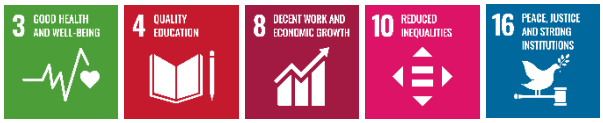

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
API	Application Programming Interface
AUC	Area Under the Curve
BOW	Bag of Words
Bs4	Beautiful Soup
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
CRISP-DM	Cross-Industry Standard Process for Data Mining
DistilBERT	Distilled BERT
DL	Deep Learning
DT	Decision Tree
FastText	Fast Text Embedding
FPR	False Positive Rate
F1-Score	Harmonic Mean of Precision and Recall
GPU	Graphics Processing Unit
GPT	Generative Pre-trained Transformer
GRU	Gated Recurrent Unit
GloVe	Global Vectors for Word Representation
GUI	Graphical User Interface
IDE	Integrated Development Environment
IQR	Interquartile Range
JtP	Jupyter Notebook
KNN	K-Nearest Neighbors
LSTM	Long Short-Term Memory
ML	Machine Learning

NB	Naive Bayes
NLP	Natural Language Processing
NN	Neural Network
PCA	Principal Component Analysis
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
ROC-AUC	Area Under the Receiver Operating Characteristic Curve
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
TPR	True Positive Rate
UGC	User-Generated Content
Word2Vec	Word to Vector
XGBoost	Extreme Gradient Boosting

1. INTRODUCTION

Journalism has been going through transformations since its creation, "First, the sporadic forerunners, gradually moving towards regular publication; second, more or less regular journals but liable to suppression and subject to censorship and licensing; and third, a phase in which direct censorship was abandoned but attempts at control continued through taxation, bribery, and prosecution" as introduced by Britannica (2024). Today, press companies are one of the pillars of a developed civilization, being one of the most influential media in any country and therefore coveted by people who want to benefit from that influence. Meanwhile, on December 5th, 1969, the first Internet connection marked the beginning of a new era, opening many doors and threatening the future of various industries, including journalism. This technology democratized sources of information, allowing everyone to disseminate information globally at any time. Some industries did not survive this technological revolution, others adapted, and some were replaced. This has happened to journalism, which has seen a drastic decline in newspaper consumption, the replacement of physical media by digital media, and the emergence of a high level of low-cost competition focused on distributing news via digital platforms. As a result, press companies have had to reinvent themselves, adapting to new technologies and to the new needs of consumers of journalistic content (Canavilhas & Rodrigues, 2013).

This project, together with data from the Global Media Group, aims to fulfill the quest to understand the newly adapted needs of consumers and accompany them through an Artificial Intelligence algorithm that classifies the news by labelling it, based on Dmitry Shishkin's eight user needs: "*Keep Me Engaged*", "*Update Me*", "*Educate Me*", "*Give Me Perspective*", "*Entertain Me*", "*Inspire Me*", "*Connect Me*" and "*Help Me*" (Shishkin et al., 2021). This algorithm allows organizations to automatically classify their content and derive valuable insights that can lead to greater engagement and new customers. Labelling is critical because it increases engagement, enables targeted content, provides valuable insights, automates management, and increases both user retention and acquisition.

Specific research questions addressed in this project are: Is the training set representative and sufficient for the classification algorithm? A small or unbalanced training set can make the model ineffective. This leads to the next question: Which type of news is least represented in the training set? Identifying under-represented categories helps to ensure a balanced data set, and therefore the quality of the training set. Another key question is which machine learning model performs best at classifying journalistic content into the defined user needs. Selecting the optimal model is critical to achieving accurate classifications. This is related to the question: What are the most effective ways to evaluate the performance of these algorithms? Defining the best evaluation metrics ensures proper model selection. After all, is the overall accuracy of the classification model greater than 0.5? Models are only useful if they can outperform random chance. These questions ensure the quality and effectiveness of the model.

2. LITERATURE REVIEW

2.1. INTRODUCTION

2.1.1. Overview

The evolution of text classification algorithms has been significantly influenced by advancements in ML, Natural Language Processing (NLP), and the increasing availability of digital text data over the past few decades.

In the early stages of text classification, the predominant approach was based on manual methods, where human experts categorized documents based on predefined criteria. As the quantity of digital text data increased, the necessity for automated approaches became evident. Initial automated methods were based on statistical techniques, such as Naive Bayes and Decision Trees. These methods were built on simple probabilistic models and heuristic rules, which provided a foundation for more complex algorithms.

The introduction of ML algorithms marked a significant shift in text classification. Support Vector Machines (SVM) emerged as a highly effective tool, offering superior accuracy by optimizing the distance between classes. Support Vector Machines are well-suited to text classification tasks due to their ability to handle high-dimensional data, which is often the case in such applications (Joachims, 1998).

Another significant advance was the development of techniques for representing features. The Bag of Words (BOW) model, which represents text as a collection of individual words, was one of the earliest and simplest methods. Despite its limitations in capturing context, BOW established the foundation for more sophisticated approaches, such as Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF represents an improvement upon BOW, as it weights words based on their importance across documents, thereby enhancing the representation of text (Sparck Jones, 1972).

The appearance of neural networks and Deep Learning (DL) has revolutionized text classification. Recurrent neural networks and their variants, such as long short-term memory (LSTM) and gated recurrent unit (GRU), addressed the limitations of traditional models by capturing sequential dependencies in text. These models have significantly improved the handling of context and long-range dependencies (Schmidhuber & Hochreiter, 1997).

Convolutional Neural Networks (CNN), although initially designed for image processing, have been successfully applied to text classification by treating text as a temporal sequence. Convolutional Neural Networks ability to capture local patterns has made them effective in a range of NLP tasks. The attention mechanism enables models to focus on the most pertinent elements of the input sequence, thereby enhancing the performance of Recurrent Neural Networks and LSTM models (Bahdanau et al., 2014). Transformer models, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and

Generative Pre-trained Transformer (GPT) (Radford et al., 2018), have established new benchmarks in text classification by leveraging self-attention mechanisms to capture global context and relationships within text.

The concept of transfer learning, whereby models that have been pre-trained on large datasets are fine-tuned for specific tasks, has further enhanced text classification. Models such as BERT and GPT-3, which have been pre-trained on vast amounts of data, achieve state-of-the-art performance across a range of NLP tasks with minimal task-specific training (Radford et al., 2019).

2.2. EVALUATION OF EXISTING RESEARCH

2.2.1. Strengths of Existing Algorithms

The algorithms currently in use for text classification have demonstrated several strengths, contributing to significant advancements in the field of NLP processing and related disciplines. Modern algorithms, in particular DL models such as Recurrent Neural Networks with long short-term memory and gated recurrent units, and convolutional neural networks, have achieved high levels of accuracy and precision in text classification tasks. These models can capture complex patterns and interdependencies within text data, increasing performance when compared to traditional ML models such as Naive Bayes and Support Vector Machines (Schmidhuber & Hochreiter, 1997). Transformer models, including BERT and GPT, have transformed the field of text classification through the utilization of self-attention mechanisms, enabling the capture of global dependencies and contextual information within the text. These models have established new benchmarks in various NLP tasks, providing a profound understanding of the text's context and semantics (Vaswani et al., 2017). Many modern algorithms are highly scalable and can efficiently handle large datasets. Techniques such as transfer learning permit models to be pre-trained on extensive datasets and fine-tuned on smaller, task-specific datasets, thereby reducing the necessity for extensive computational resources and labeled data for each new task (Radford et al., 2019).

DL models, in particular transformer-based models, have demonstrated versatility across different NLP tasks, including text classification, sentiment analysis, named entity recognition, and more. This adaptability makes them valuable tools for a variety of applications in the field.

2.2.2. Weaknesses and Limitations

Despite the strengths, existing algorithms for text classification also have several weaknesses and limitations that need to be addressed.

DL models, particularly transformer models like BERT and GPT, require significant computational resources for training and inference. This includes high-performance Graphics Processing Unit (GPU) and large memory capacities, which may not be accessible to all researchers or organizations (Strubell et al., 2019).

These models often require large amounts of labeled data for training to achieve optimal performance. In many cases, obtaining such large, labeled datasets can be challenging and time-consuming, limiting the applicability of these models in domains with scarce data.

DL models, especially those with complex architectures, often lack interpretability. Understanding the decision-making process of these models can be difficult, making it challenging to trust and validate their predictions, particularly in critical applications (Lipton, 2016).

These models can inherit and even amplify biases present in the training data. Ensuring fairness and reducing bias in model predictions remains a significant challenge, as biased models can lead to unfair or discriminatory outcomes (Bolukbasi et al., 2016).

DL models are prone to overfitting, especially when trained on small datasets. Overfitting occurs when a model learns to perform well on the training data but fails to generalize to new, unseen data, leading to poor performance in real-world applications.

2.2.3. Methodological Approaches

Similar research on text classification algorithms takes a different approach, as shown in Table 1, each with its strengths and applications, both traditional and DL models were employed.

Table 1 - Methodological Approaches

Research	Method
(Joachims, 1998)	Transform text into high-dimensional feature vectors and use SVMs to find optimal hyperplanes for categorization.
(Archad et al., 2021)	Compare classifiers (Naive Bayes, Linear SVM, Logistic Regression, Word2Vec, Doc2Vec) on multi-class text data.
(Schmidhuber & Hochreiter, 1997)	Use memory cells to retain information over long sequences, addressing the vanishing gradient problem.
(Howard & Ruder, 2018)	Fine-tune pre-trained language models for specific text classification tasks using transfer learning.

2.3. PATTERNS, THEMES, AND DEBATES IN TEXT CLASSIFICATION

2.3.1. Patterns and Trends

A significant trend in text classification is the shift from traditional ML models to DL approaches, which have become the standard due to their superior performance and ability to handle complex text data. Modern text classification increasingly focuses on understanding the context within the text. Transformer models, which use attention mechanisms to capture long-range dependencies and contextual information, exemplify this trend. These models have established new benchmarks across a range of NLP tasks (Vaswani et al., 2017).

Transfer learning, where models are pre-trained on large datasets and fine-tuned for specific tasks, has become a common approach. This approach leverages the extensive knowledge embedded in pre-trained models such as BERT, GPT, and RoBERTa, improving performance on specific text classification tasks with limited data.

Word embedding techniques, such as Word2Vec, GloVe, and contextual embeddings derived from transformer models, have enhanced the representation of text data. These embeddings capture semantic relationships and provide dense, informative vector representations, contributing to improved classification accuracy (Mikolov et al., 2013).

There is a growing awareness of the importance of fairness and the mitigation of bias in text classification models. A growing number of researchers are focusing their attention on the identification and rectification of biases present in training data and model predictions, to ensure the ethical and equitable development of AI systems (Bolukbasi et al., 2016).

2.3.2. Themes and Significant Findings

Research consistently demonstrates that DL models outperform traditional models in terms of accuracy and scalability.

The effectiveness of feature representation techniques is a recurring theme. TF-IDF and word embeddings have been demonstrated to be essential in enhancing the performance of classification models. Embedding techniques that capture context and semantics have been particularly significant in advancing the field. While DL models offer high accuracy, their interpretability remains a challenge. This has led to a body of work focused on developing methods to make these models more transparent and understandable, addressing the nature of neural networks (Lipton, 2016). The challenge of imbalanced data is a common theme. Techniques such as resampling, class weighting, and synthetic data generation are frequently employed to address this issue and to improve model performance and reliability.

The choice of evaluation metrics is critical in assessing model performance. The most employed metrics in this context are accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Researchers have emphasized the

importance of selecting metrics that align with the specific goals and context of the classification task (Powers & Ailab, 2011).

2.3.3. Debates and Controversies

The relationship between model complexity and interpretability is a topic of ongoing debate in the field of ML. One of the primary debates in text classification is the trade-off between model complexity and interpretability. Although complex models, such as deep neural networks, can achieve high levels of accuracy, they frequently lack transparency, making it challenging to comprehend the rationale behind their decision-making processes. This has prompted discussions on the necessity for interpretable models in applications where explainability is important (Lipton, 2016).

Another significant debate concerns the issue of bias and fairness in text classification models. Researchers have expressed concern that models may inherit and amplify biases present in training data, potentially leading to unfair or discriminatory outcomes. Efforts are ongoing to develop methods to detect, mitigate, and eliminate bias in AI systems (Bolukbasi et al., 2016).

The ethical implications of utilizing large datasets for the training of text classification models continue to be a topic of ongoing debate. The issues of data privacy, consent, and the ethical use of AI technologies are of critical importance to researchers and practitioners alike (Floridi et al., 2018).

Furthermore, there is a debate regarding the generalization capabilities of text classification models. While some models demonstrate efficacy across a range of tasks, others are designed for specific applications. The challenge of balancing the need for generalizable models with the benefits of specialized models remains a significant issue in the field (Howard & Ruder, 2018).

The selection and interpretation of evaluation metrics are subjects of contention. While accuracy is a common metric, it can be misleading in imbalanced datasets. Other metrics, such as precision, recall, F1-score, and AUC-ROC, are also employed, but the optimal metric for a given task remains a matter of contention among researchers (Powers & Ailab, 2011).

2.4. GAPS IN THE LITERATURE

2.4.1. Identifying Gaps in Current Text Classification Research

Despite the advancements in DL models, there remains a significant gap in the interpretability and transparency of these models. Understanding decision-making processes is challenging. This lack of transparency can be problematic, especially in critical applications where understanding the rationale behind predictions is essential (Lipton, 2016).

Recent research has identified the presence of biases in text classification models, which can result in unfair or discriminatory outcomes. However, there is still a need for more robust

methods to detect, mitigate, and eliminate bias in these models. Ensuring fairness in AI systems is an ongoing challenge that requires further exploration and innovative solutions (Bolukbasi et al., 2016).

While transfer learning has helped with some challenges related to data scarcity, there is still a need for high-quality labeled datasets to train and evaluate text classification models. Many domains lack sufficient data, which limits the applicability and generalizability of existing models. Additionally, the quality and diversity of training data plays a crucial role in model performance, highlighting the need for better data collection and annotation practices.

Imbalanced data is a common issue in text classification, where some classes have significantly more instances than others. Various techniques have been proposed to address this issue. However, there is still a need for more effective and scalable solutions to handle imbalanced datasets.

The selection of appropriate evaluation metrics remains a topic of debate. There is a need for a consensus on the most suitable metrics for different text classification tasks to ensure a comprehensive evaluation of model performance (Powers & Ailab, 2011).

2.4.2. Opportunities for Future Research

Future research should focus on developing models that balance accuracy with interpretability. Interpretable models will be crucial for applications requiring explainability and accountability (Ribeiro et al., 2016). Research should continue to develop and refine methods for detecting and mitigating biases in text classification models. Ensuring ethical AI systems will be a critical area of focus (Floridi et al., 2018). Efforts should be made to improve the quality and availability of labeled datasets. High-quality data is essential for training robust and generalizable models. Consequently, future research should explore innovative techniques to address the challenges posed by imbalanced datasets. This includes advanced resampling methods, cost-sensitive learning, and anomaly detection techniques to improve model performance on underrepresented classes. Furthermore, there is a need to standardize the evaluation metrics used in text classification research. The development of guidelines for metric selection based on task-specific requirements and the promotion of the use of comprehensive evaluation frameworks will ensure the consistent and reliable assessment of model performance (Sokolova & Lapalme, 2009).

2.4.3. Contributions of the Current Study

The present study seeks to address several gaps identified in the existing literature and make meaningful contributions to the field of text classification, with a particular focus on the Portuguese language.

The study will investigate and identify the most suitable text vectorization methods for the Portuguese language. By testing various text representation techniques, including BOW, TF-

IDF, and word embedding methods like Word2Vec, and GloVe, among others, will be evaluated to ascertain their capacity to capture the nuances and semantic relationships inherent in Portuguese text. This analysis will provide valuable insights into optimal text representation practices in Portuguese, thereby facilitating more accurate and efficient text classification.

This study evaluates both traditional and advanced models for text classification. The research will provide insights into the relative strengths and weaknesses of each approach. This comparative analysis will highlight the performance differences and identify the most effective models for text classification tasks in the Portuguese language.

A key contribution of this study is the emphasis on using appropriate evaluation metrics to assess model performance. Traditional metrics such as accuracy, precision, recall, and F1-score will be complemented by metrics that account for imbalanced datasets and model fairness, such as AUC-ROC, Cohen's Kappa, and Balanced Accuracy. This comprehensive evaluation strategy will guarantee a robust assessment of model performance.

3. METHODOLOGY

3.1. INTRODUCTION TO METHODOLOGY

For this project It was decided to choose a methodology known as CRISP-DM. It is a known methodology in the data mining industry, which focuses on six phases, the last being the implementation of the model. There are several reasons to choose this methodology, such as the experience of the author implementing this methodology, and, most importantly, the perfect fit this project has with this method. This method requires a strong grasp of the purpose of this project as problem-solving and business contextualization, it requires a very good understanding of the data and a thorough analysis, and it is flexible to any changes that may be made.

3.2. RESEARCH DESIGN

This section presents an opportunity to elaborate on the concept model applied in this project. The CRISP-DM framework comprises six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. The workflow follows the order shown in Figure 1.

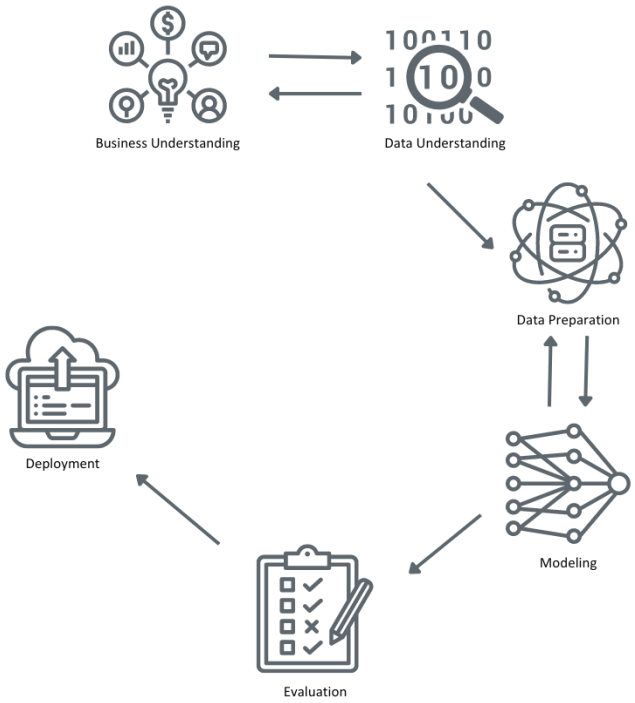

Figure 1 - Crisp-DM

CRISP-DM starts with an approach, Business Understanding, this first phase is crucial as it forces us to stop and understand the ambition of the project, its significance in the context in which we have it, the concrete objectives to determine the success of the project and a plan of what we want to concretely obtain these results. It is interesting because it requires, right

from the start, several skills, such as problem analysis, creativity, business focus, results orientation, planning and organization, critical sense, and detail, among others, or being a fusion between hard and soft skills, which together lead to a good understanding of the framing of the problem within the company's needs.

Following the planning phase, the subsequent step was the Data Understanding phase. The initial step involved the collection of data through web scraping from a specific website. This entailed the description of the data, including the identification of null entries, unique values, data types, and missing values, both in relative and absolute terms. Subsequently, it was conducted a statistical analysis of the data, examining the distributions of each variable, the occurrence of unique values in categorical features, and the relationships between variables. To assess the quality of the data, it has been employed the use of boxplots to identify any outliers or potential issues. Finally, correlation analyses were performed to gain insight into the relationships between variables. This ensured a comprehensive understanding of the data before proceeding to the data preparation phase.

As a common fact, data rarely comes in perfect shape. Therefore, this third phase is responsible for preparing the data by starting with cleaning data, getting rid of duplicates and rows that have too many missing values, checking outliers, and making decisions regarding them. Following that, Feature Engineering was performed. The tasks in this part depend on the project itself, but it is common to find tasks such as categorical features encoding, normalization, and creating new variables.

The modeling phase marks an important transition from understanding and preparing data to the actual work in the model. This stage involves various steps that have a high responsibility in the model performance, starting from data separation into training, Validation, and Test. Every decision has an impact on the results of the model, so it is expected to go back and forth, select different models, and try different hyper-parameters until the best model possible for this project was got.

Once the model is prepared, it is crucial to evaluate its performance on a dataset that the model has not seen before. The importance of the chosen metrics depends on the specific problem being addressed and the types of errors that may occur. This evaluation allows us to predict how the model will perform with future data and ensure that any mistakes made do not negatively drastically impact business results.

Finally, we proceed to the deployment phase. This phase focuses on deploying the developed model into operational systems and integrating it into the existing infrastructure. This involves packaging the model into deployable artifacts, integrating it with production systems, and establishing mechanisms for model monitoring and maintenance. After execution, the model is ready for use. It is important to keep in mind that updates are necessary from time to time to improve its performance through further training.

It can be concluded that the CRISP-DM methodology provides a structured framework for this project. By following the six distinct phases, a model capable of delivering the expected results was built. The simplicity of the methodology allows for easy understanding and review of the building process, leading to various benefits. The most important of these benefits is the ability for anyone to improve the model.

3.3. METHODOLOGY APPLICATION

3.3.1. Business Understanding

This section explores the process of understanding the business context and objectives of the text classification project for journalistic content. The Business Understanding phase of the CRISP-DM methodology is crucial as it lays the foundation for defining the problem statement, identifying success criteria, and formulating a plan to achieve desired outcomes.

The initial stage of the Business Understanding phase required a thorough comprehension of the problem statement and its relevance within the wider business context. In this case, the objective is to categorize news into eight user-need labels. By achieving this, the aim is to address the challenge of enhancing customer engagement by introducing a pertinent variable that will provide additional perspective into news engagement and customer behavior analysis. This will offer valuable insights to help solve the problem. This project aims to address the need for automating the categorization of news articles to make the task faster, cheaper, more consistent, and less prone to human error.

The problem statement is clearly defined, and the next step was to identify the specific objectives of the text classification project. These objectives were determined through collaboration with stakeholders. The main objective is to develop an algorithm that classifies news based on eight user-need labels.

To measure the success of the classification model, clear and quantifiable criteria were established. These benchmarks evaluate the effectiveness and impact of the model. Key success criteria include:

- General Accuracy: To evaluate the overall correctness of the model.
- Balanced Accuracy: To calculate the average accuracy obtained in each class.
- Label Precision: The proportion of true positive instances among the predicted positives for each label.
- Label Recall: The proportion of instances that are correctly identified as positive for each label.
- Label F1-Score: The harmonic means of precision and recall for each label.
- Confusion Matrix: A visual representation of the confusion between the predicted and actual labels.
- Cohen’s Kappa: To evaluate the level of agreement between two raters.

Based on the identified problem statement, objective, and success criteria, and plan of action, a comprehensive plan was formulated to guide the execution of the text classification project. The plan included the following steps:

- Data acquisition: Getting the data through web scraping.
- Data understanding: Check the problems that might come up with the data.
- Data preprocessing: Process the data using specific techniques.
- Modelling: Split data, find the best model and its hyperparameters.
- Evaluation: Evaluate the model's performance based on some indicators mentioned in the success criteria.
- Deployment: Deploy the model into the systems of the company.

As described by Schröer et al. (2021) "The business situation should be assessed to get an overview of the available and required resources. The determination of the data mining goal is one of the most important aspects in this phase. First the data mining type should be explained (e.g. classification) and the data mining success criteria (like precision). A compulsory project plan should be created.". Once these components are in place, you can move on to the Data Understanding phase.

3.3.2. Data Exploration

This section provides a comprehensive understanding of the data used in the text classification project. The Data Understanding phase of the CRISP-DM methodology involves loading and collecting data, describing data characteristics, exploring data distributions, assessing data quality, and analyzing data relationships.

The initial step in the Data Understanding phase involved retrieving textual data through web scraping techniques. Web scraping is the process of programmatically extracting information from websites, offering several advantages for gathering textual data. To retrieve textual data from websites, the "*Requests*" and "*Beautiful Soup*" (bs4) libraries in Python were utilized. The Requests library simplifies the process of making HTTP requests to web servers (Yun et al., 2019) Web scraping allows access to unstructured textual data found on web pages, such as news articles, blog posts, and product descriptions.

One of the main benefits of web scraping is the automation of the information retrieval process. The dataset provided includes links to one thousand news articles, making manual retrieval unfeasible. Therefore, this technique enabled me to save resources and time. Additionally, web scraping enables customizable data selection, allowing us to target specific sections of a webpage or filter data based on predefined criteria. In this case, the webpages were from the same website, so the criteria were the HTML section related to the body where the article was written. This flexibility ensures that the collected textual data aligns with the research objectives and requirements of the text classification project.

After retrieving the data, its characteristics using the "*Pandas"* library for data manipulation and analysis were retrieved. This library provided powerful tools for describing the dataset, including the number of articles, distribution of classification labels, and article lengths. In addition to "*Pandas*", together with "*NumPy"* library for numerical computing, which efficiently handled large arrays of numerical data. To visualize data distributions, we used the "*Matplotlib"* and "*Seaborn"* libraries. "*Matplotlib"* is a versatile plotting library that enables the creation of a wide range of static, interactive, and animated visualizations. "*Seaborn"*, which is built on top of "*Matplotlib*", provides a high-level interface for creating informative and attractive statistical graphics. The libraries allowed for the creation of visualizations, including histograms, word clouds, and topic distributions, to gain insights into the underlying patterns and themes present in the dataset.

Upon analyzing the data, unique values of the categorical features and their frequencies throughout the dataset were examined. Given the high number of unique values, it was not practical to analyze each one individually. However, certain unique values were identified, such as "*pagepath*", that should not appear more than once. This indicated that some news articles were being retrieved multiple times, which was not relevant to my study. Next, the distribution of each label was checked by plotting a pie chart. This revealed that the dataset was imbalanced. To further investigate, a random sample of 40 news articles was selected and printed. During this process, it was discovered that some "*New_Articles*" contained only the text "*Página Não Encontrada*" ("page not found"). This suggested that some "*pagepaths*" were broken or not working properly, rendering them useless for model training.

To assess data quality, various techniques were used, including statistical analysis and visualization, to identify and address issues such as missing values, duplicate entries, and outliers. Statistical methods from the "*SciPy*" library were employed to perform data quality checks and ensure the reliability and consistency of the dataset. Missing values, duplicates, and outliers are common data quality issues that can affect the integrity and reliability of the dataset, ended up finding several missing values in the "*user-need*" variable. Missing values are data that are absent in one or more fields or records within a dataset. These values can be attributed to several factors, including data entry errors, equipment malfunctions, or survey non-responses. In this instance, a missing data plot was employed to ascertain the probable cause of this missing data. It appeared that the absence of finishing the task of manually classifying the news, once the last 165 observations had missing labels, was the probable cause. Missing values can distort statistical analyses and ML models, leading to biased results and inaccurate predictions. Duplicates occur when two or more records in a dataset have identical values across all or most fields. Duplicate records can cause inflated counts and skew statistical analyses, potentially leading to incorrect conclusions about the underlying data distribution.

At last, data relationships were analyzed, explored correlations between variables in the dataset by using Cramer's V coefficient to measure the strength of association between

categorical variables, as supported by Rayward-Smith (2007) “This test is widely used as a measure of correlation between two sequences of nominal values. It always delivers a value between 0 and 1.”. Luckily no strong correlation was found between variables, which is a positive output for this activity.

Overall, the use of these libraries simplifies all tasks, and statistical knowledge is mandatory to understand the data. Once this phase is complete and all problems are identified, we can proceed further with the CRISP-DM method.

3.3.3. Preprocessing

During the Data Preparation stage of my text classification project, it is crucial to pay meticulous attention to detail to ensure that the dataset is thoroughly refined, optimally structured, and free from errors for subsequent analysis and modeling. This stage comprises two pivotal components: Data Cleaning and Feature Engineering, each playing a critical role in enhancing the dataset's quality, integrity, and utility. As supported by Kantardzic (2019) “The most important step in data science is to prepare the data. Data preparation is the process of cleaning, processing, and transforming raw data for analysis. From this stage, the errors in the data can be effectively handled by cleaning, identifying the missing values, handling outliers, etc. Hence, this chapter discusses the methodologies used to prepare the data using the Pandas package in “Python””.

Initially, the missing values issue was addressed. Specifically, the rows where the "*user-need*" column had missing values (165 rows) were excluded, as these entries lacked essential information and were irrelevant for model training. Additionally, the 9 rows where the page path was not functioning were also removed, as these entries did not provide valid news articles.

Subsequently, transformations were applied to both the general data frame (df) and the exclusive content data frame (df_exclusive). In the general data frame, all entries containing exclusive content, as identified by the presence of the string "*Já é assinante*?" in the "*News_Article*" column, were removed. Then, the same exclusive content entries from the data frame containing exclusive content were appended to the general data frame, resulting in the addition of 136 exclusive content entries to the general data frame, which accounted for 17% of the general data frame.

To address the inconsistencies in label names, the letter case of labels was standardized to ensure uniformity across categories, for example, the label "*Connect me*" was changed to "*Connect Me*" and similar adjustments were made to other label categories. This step aimed to improve data consistency and facilitate subsequent analysis.

Following this, the text was cropped to remove irrelevant sections and noise, focusing on the main article text. The content above the string "*ASSINAR JN DE HOJE*" and below "*TÓPICOS:*", which were non-essential for model training, were discarded. Additionally, articles with

approximately 30 words, which often corresponded to empty or incomplete articles, were identified, and removed.

Lowercasing was then performed to standardize the text and improve normalization. This process also reduced the vocabulary size, facilitating better generalization and avoiding redundancy in the data. Subsequently, punctuation was removed to further normalize the text, reduce dimensionality, improve tokenization, and enhance model interpretability. Stopword removal was applied to eliminate common words that do not carry significant meaning, such as "*e*", and "*o*", among others, this step aimed to reduce noise, improve model interpretability, and speed up preprocessing. As stated by Lazarinis (2007) ”It has been claimed that a word which appears in 80% of the documents in a document collection is useless for purposes of retrieval”.

For lemmatization, the “*Stanza*” library was employed, under research indicating its effectiveness for Portuguese text. Lemmatization was selected due to its capacity to reduce words to their base forms, thereby enhancing analysis. As stated by Lazarinis (2007) “Lemmatization involves the reduction of words to their respective headwords (i.e. lemmas). In the linguistic dictionaries every entry corresponds to a lemma that defines a set of words with the same lexical root.”.This process resulted in improved text coherence and semantic understanding. Due to the time-intensive nature of lemmatization, the data was exported to Excel for further processing.

The handling of numerical characters involved the identification and handling of the 4750 occurrences of numerical values within the text, so it was decided to normalize the numbers that were written in a textual format by identifying them and replacing them with the same format of numerical representation, although they have kept as a string. Special characters, present in 15 articles, were meticulously examined. However, no removal was necessary as they constituted normal Portuguese language accentuation.

HTML tags were not present in the text data, simplifying the preprocessing steps. Finally, rare words were addressed using the “*Natural Language Toolkit*”, identifying rare words, and applying appropriate handling strategies to reduce noise and enhance model performance.

To summarize the preprocessing part analyzed the balance of user-need labels through the df, affirming that, “*Update Me*” has 411 occurrences, “*Give Me Perspective*” has 87 occurrences, “*Divert Me*” has 58 occurrences, “*Keep Me Engaged*” has 57 occurrences, “*Help Me*” has 37 occurrences, “*Inspire Me*” has 19 occurrences, “*Connect Me*” has 12 occurrences, and “*Educate Me*” has 11 occurrences, confirming after preprocessing that the dataset is imbalanced, and expecting the model to not have such good accuracy in labeling the least frequent labels.

After all the techniques performed the text was ready for modeling, for example, a text that would be at first “*O despiste de dois Ferrari, no último sábado, contra muro de uma vivenda agitou a pacata região de Osimo, em Itália.”* (The crashing of two Ferraris into the wall of a

villa last Saturday has shaken the peaceful region of Osimo in Italy), would be transformed to *"despiste 2 ferrari último sábado contra muro 1 vivenda agitar pacato região osimo italia"* (crash 2 ferrari wall villa saturday shake peaceful region osimo italy), being the text suitable for modeling.

3.3.4. Modelling & Evaluation

The modeling and evaluation processes were iterative, involving repeated adjustments to parameters, data preprocessing refinements, model retraining, and re-evaluation until an optimal performance was achieved. This approach was informed by similar studies in literature and the application of state-of-the-art embedding techniques. As stated by Jackson (2002) "Typically, several techniques can be applied to the same data mining problem type. Some techniques require a specific form of data. Therefore, stepping back to the data preparation phase is often needed. This approach was informed by similar studies in literature and the application of state-of-the-art embedding techniques.

Before model testing, the dataset was divided into two distinct sets: a training set comprising 80% of the data and a testing set comprising the remaining 20%. Given the imbalanced nature of the dataset, a stratified split was employed to ensure that the training and testing sets were representative of the overall class distribution. Furthermore, class weights were calculated for the least representative labels and adjusted according to the model results for each label to effectively address the imbalance.

Initially, Recurrent Neural Networks with Long Short-Term Memory and Gated Recurrent Unit architectures were implemented using the "*Keras*" library, as supported by Shen & Zhang (2016) "LSTM is used in our comparison since LSTM is comparable to GRU and the two both outperform basic RNN". The decision to use neural networks was based on their strong performance in similar literature, as affirmed by Lai et al. (2015) "the experimental results show that the neural network approaches outperform the traditional methods for all four datasets" . For Recurrent Neural Networks, which are designed for sequential data, various word embedding techniques were tested, including:

- GloVe
- Word2Vec
- FastText
- BERT

Although TF-IDF is typically employed in conjunction with traditional ML models, it was also utilized as an input for Recurrent Neural Networks to gain insight into its potential advantages. In addition, pre-trained models such as BERT, DistilBERT, and GPT were evaluated to determine their efficacy in capturing semantic meanings and contextual information.

In addition to neural networks, traditional ML models were subjected to testing:

- Logistic Regression

- Random Forest
- Gradient Boosting
- XGBoost
- Ensemble Models

To assess the efficacy of these models, TF-IDF was employed due to its simplicity and effectiveness in non-sequential data scenarios. However, to exhaustively investigate the potential of these models, combinations of BOW, and GloVe (the most successful performer in previous Recurrent Neural Networks tests), other than TF-IDF were also considered. This approach was inspired by studies suggesting that BOW and embedding methods can also output good results with traditional models. As mentioned by Ong et al. (2020) "Text featurization converts narrative text into structured data. Examples of text featurization methods include Bag of Words (BOW), Term Frequency-Inverse Document Frequency (TF-IDF) and word embeddings. Word embedding methods, including Word2Vec and Global Vectors for Word Representation (GloVe), learn a distributed representation for words".

To achieve optimal performance for each model, a systematic approach involving exhaustive search through a manually specified subset of the hyperparameter space was employed, as implemented by GridSearch. The optimal combinations were selected based on performance metrics obtained during cross-validation.

Models were evaluated according to several metrics: General Accuracy, Balanced Accuracy, Label Precision, Label Recall, Label F1-score, Confusion Matrix, and Cohen’s Kappa.

In addition to the 5 previously mentioned metrics, the misclassified observations were printed and examined in detail to draw insights and improve the model's performance.

The objective of the study was to identify the optimal approach for the classification task. To this end, various combinations of models and embedding techniques were tested, and GridSearch was employed for hyperparameter optimization. The iterative process of model adjustment, training, and evaluation ensured that the final models were robust and effective, leveraging the strengths of both neural networks and traditional ML methods.

4. RESULTS AND DISCUSSION

4.1. INTRODUCTION TO RESULTS AND DISCUSSION

This section aims to offer a clear and detailed account of the effectiveness of the models and the implications for the classification of journalistic content. Any problems encountered during the evaluation process and their significance are also discussed, providing a complete and transparent overview of the study's findings.

4.2. MODEL PERFORMANCES

4.2.1. Recurrent Neural Networks

The Recurrent Neural Network models were evaluated using a variety of word embedding techniques. The principal objective was to calculate the balanced accuracy and Cohen's Kappa metrics for each model when combined with each embedding technique. The optimal Recurrent Neural Networks architecture includes a single dense input layer with a dropout rate of 0.2, a single LSTM layer, a single GRU layer, two dense layers, both with a dropout rate of 0.2, and an output layer. The facts are presented in tabular form in Table 2 and Table 3.

Table 2 - RNN Model Architecture

Parameter	Dense 1	Dense 2	Dense 3	Output
Units	300	64	32	8
Activation	Relu	-	-	softmax
Dropout	0.25	0.25	-	-

Table 3 - RNN Model Architecture (LSTM & GRU)

Parameter	LSTM	GRU
Units	64	64
Return_sequences	True	False
Dropout	0.15	0.15

In contrast to the findings of existing literature, as displayed in Table 4, the best-performing recurrent neural network model employed the TF-IDF as its input. The model achieved a general accuracy of 0.66, indicating that 66% of the total predictions made by the model were correct. A balanced accuracy of 0.35 was observed, suggesting that the model's performance varies significantly across different classes, with some classes being accurately predicted while others were not. Finally, a Cohen's Kappa of 0.37 was calculated, indicating fair agreement between the predicted and actual labels, given the imbalanced nature of the data.

Table 4 - Vectorization Results on RNN

Input	Overall Accuracy	Balanced Accuracy	Cohen's Kappa
GloVe	0.525	0.249	0.105
Word2Vec	0.568	0.318	0.202
FastText	-	-	-
Bert	-	-	-
TF-IDF	0.662	0.353	0.371

Despite demonstrating the highest performance among Recurrent Neural Networks models in this study, Figure 2 shows that the model failed to correctly classify the "*Connect Me*" and "*Divert Me*" labels, as illustrated in the confusion matrix plot. The confusion matrix reveals specific issues with the "*Connect Me*" and "*Divert Me*" labels, which all the models consistently misclassified. This indicates that while the model performs adequately overall, it is less effective when dealing with less frequent or more complex categories.

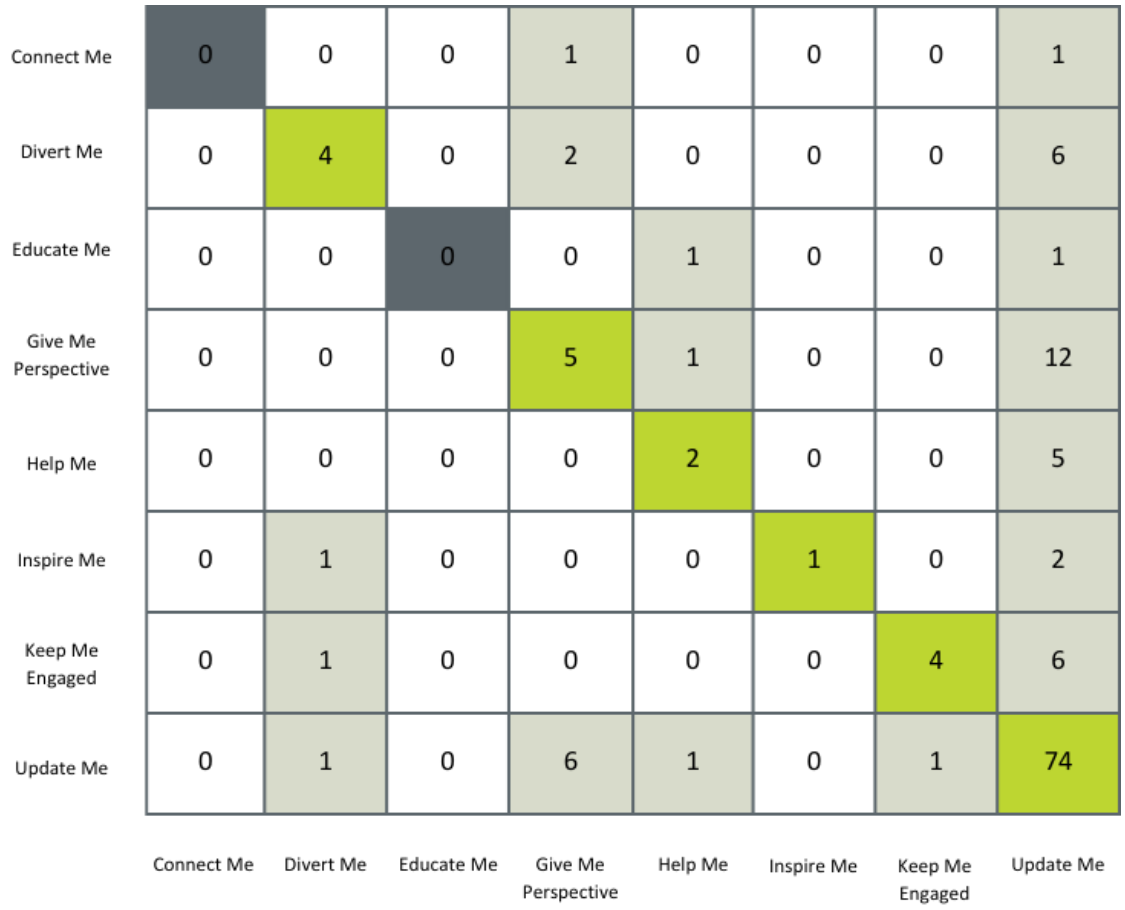

Figure 2 - Confusion Matrix RNN

Due to the limited computational resources available, it was not possible to conduct evaluations using more sophisticated word embeddings, such as FastText and BERT. Furthermore, attempts to run BERT and GPT pre-trained models also failed. Despite efforts to utilize cloud-based platforms to overcome these limitations, these solutions were unable to support the computational load required for these models. To address these challenges, a smaller and faster version of BERT, known as DistilBERT, was employed. However, the results obtained with DistilBERT were not as satisfactory as those obtained with the TF-IDF-based model.

4.2.2. Traditional Models

In addition to evaluating Recurrent Neural Networks, traditional ML models were tested to determine their effectiveness in classifying journalistic content. Among the three types of embedding methods tested, TF-IDF emerged as the most effective, as shown in Table 5.

Table 5 - Vectorization Results on Logistic Regression

Input	Overall Accuracy	Balanced Accuracy	Cohen's Kappa
GloVe	0.389	0.238	0.188
BOW	0.676	0.331	0.393
TF-IDF	0.705	0.362	0.457

As shown in Table 6, from all the models evaluated, logistic regression proved to be the most accurate, achieving an overall accuracy exceeding 0.7, indicating that 70% of the total predictions made by the model were correct. The Logistic Regression model achieved a balanced accuracy of 0.36, indicating that its performance varied significantly across different classes. Some classes were accurately predicted, while others were not. Additionally, a Cohen's Kappa of 0.45 was calculated, indicating fair agreement between the predicted and actual labels, given the imbalanced nature of the data.

Table 6 - Traditional Models Evaluation

Input	Overall Accuracy	Balanced Accuracy	Cohen's Kappa
GBoost	0.619	0.283	0.211
XGBoost	0.647	0.205	0.234
Random Forest	0.640	0.189	0.189
SVM	0.683	0.318	0.369
Logistic Regression	0.705	0.362	0.457
TF-IDF	0.705	0.333	0.408

Although the Logistic Regression model demonstrated the highest performance, it was unable to correctly categorize the "*Connect Me*" and "*Divert Me*" labels, as illustrated in Figure 3 highlights specific issues with these labels, which were consistently misclassified by all the models. This suggests that while the Logistic Regression model performs adequately overall, it is less effective when dealing with the "*Connect Me*" and "*Divert Me*" categories.

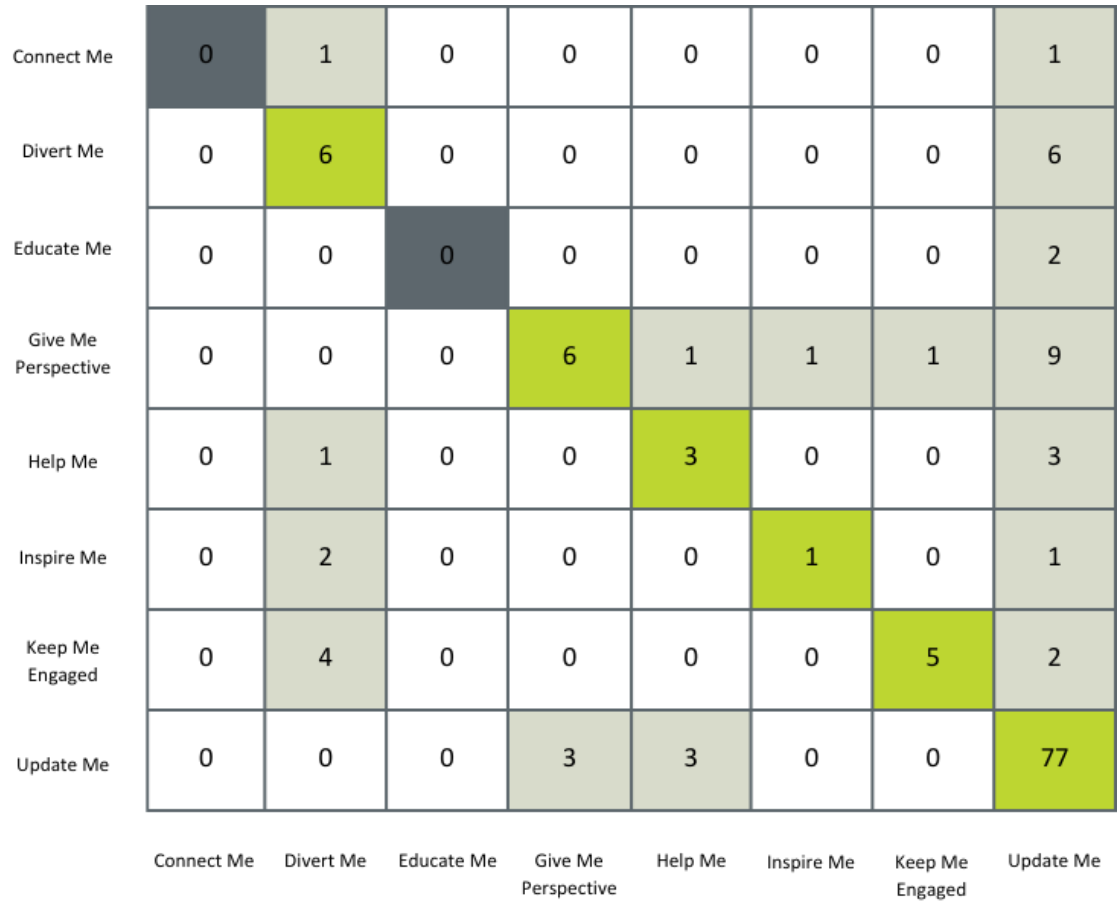

Figure 3 Confusion Matrix Traditional Models

4.3. INTERPRETATION OF RESULTS

4.3.1. Root Cause of Imbalance

Although the Logistic Regression model demonstrated a general accuracy of 0.7, which surpassed the performance of Recurrent Neural Networks, which are strongly recommended by the literature, several issues indicate that the dataset imbalance significantly affects model performance. It is particularly unusual for Recurrent Neural Networks to underperform compared to traditional methods. Moreover, the consistent misclassification of certain categories, such as "*Connect Me*" and "*Divert Me*", suggests deeper issues rooted in the dataset's composition.

The most significant issue identified is the severe imbalance in the dataset. For example, the "*Inspire Me*" label has only four test observations, which is only slightly more than the "*Connect Me*" and "*Educate Me*" labels. Despite this imbalance, the "*Inspire Me*" label achieved a precision of 0.25, which is surprisingly high given its low representation. This discrepancy prompted a more detailed examination of the dataset, to identify the root cause of the misclassification issues.

It has been evidenced in the literature that the application of manual classification frequently introduces bias, which can result in the creation of imbalanced datasets. This bias has the potential to significantly impact model performance, as evidenced by the results. A study on characterizing bias in classifiers indicates that real-world data, especially when manually labeled, tends to reflect systemic human biases and inconsistencies (McDuff et al., 2019). This finding aligns with our observation that the dataset's imbalance is likely a consequence of the manual labeling processes employed.

The analyses of the misclassified examples revealed that articles labeled "*Connect Me*" and "*Educate Me*" were often misclassified due to their ambiguous content or manual misclassification, which led to errors.

The "*Connect Me*" label is used when users express a desire to feel empathy, which motivates them to act. This approach often employs ideas or experiences to build affinity with the subject or topic. As present in Shishkin et al. (2021) "The big difference between this and the Help me need is that it's societal, not personal.". One of the "*Connect Me*" misclassified articles describes the case of a horse named Hippie that suffered from health issues and required an expensive operation. The owner and friends launched a fundraising campaign to cover the veterinary bills. The narrative emphasizes the emotional bond between the owner and the horse, describing Hippie as a family member. One reason for this issue is that the article mentions medical terms and fundraising, overlapping with themes from other categories like "*Help Me*" or "*Inspire Me*". The emotional appeal and personal story may have influenced the model to align with the "*Inspire Me*" or "*Help Me*" categories. The model may lack sufficient training examples that capture the nuances of articles aimed at connecting individuals through personal stories involving animals.

The Educate Me label indicates that the articles frequently address scientific research, studies, projects, and educational content to enhance knowledge and comprehension. The misclassified article reports on a tragic incident where a rockfall at a beach in Albufeira resulted in multiple fatalities, including a family from Porto. It provides comprehensive data regarding the victims, the precipitating cause of the rockfall, and the subsequent response from the relevant authorities. Furthermore, the article addresses the broader implications of the incident, including safety concerns and the government's response. The article primarily reports a tragic event, which is consistent with the focus of an "*Update Me*" or "*Help Me*" label on breaking news and emergency response. The educational aspect of the article may be eclipsed by the immediacy of the tragedy. The emotional tone and detailed narrative may have prompted the model to categorize the article as "*Inspire Me*" or "*Help Me*". This is potentially because the article was not correctly classified as "*Educate Me*" news, as it did not align with the keywords associated with that category.

A more detailed examination of the data set revealed that 61-page paths (indicating the same article) were, in fact, distinct articles, with three different user-need labels. This indicates that 3 out of 30 unique news articles were incorrectly classified manually. This indicates a 10%

error rate among duplicates, which provides substantial evidence for the presence of bias and errors in the manual classification task. Such inconsistencies result in significant issues during the training and evaluation of the model, as the model learns from erroneous and biased data.

To address these issues, future research should avoid manual classification and instead utilize automated or semi-automated methods to ensure consistency and reduce bias. One promising approach, in this case, is the use of models to create a training set with hypothetical news articles that include keywords that well define each label. These articles should be created proportionally to the original data set.

In conclusion, while the Logistic Regression model performed reasonably well, the root cause of misclassification and imbalance issues lies in the dataset's biased and inconsistent manual labeling. These issues can be addressed through improved data collection and labeling strategies, which will result in the development of more accurate and reliable classification models.

4.3.2. Results of Research Questions

The "*Inspire Me*," "*Connect Me*," and "*Educate Me*" labels were the least represented in the training set, resulting in a high frequency of misclassification. The underrepresentation of these categories revealed a significant imbalance in the dataset, which impeded the model's capacity to learn and accurately classify these types of news.

The training set was subject to significant limitations due to its imbalance and the introduction of manual classification biases. These issues resulted in suboptimal learning outcomes, particularly affecting the model's ability to accurately classify underrepresented categories, such as "*Connect Me*" and "*Divert Me*."

The most effective method for evaluating algorithms was found to be a combination of overall accuracy, balanced accuracy, and Cohen's Kappa. These metrics offered a comprehensive understanding of the model's performance, while also identifying specific areas requiring improvement. A detailed error analysis was also instrumental in identifying areas of weakness.

The Logistic Regression model demonstrated the highest overall accuracy, exceeding 0.7, exceeding the 0.5 threshold, indicating that the model performed significantly better than would be expected by chance and was effective in correctly predicting most journalistic content classifications. The model demonstrated superior performance to Recurrent Neural Networks and other traditional models, including Random Forest and Gradient Boosting. Despite its robust performance, the model encountered difficulties with imbalanced data and misclassification of specific categories.

5. CONCLUSION

This project establishes a classification model for journalistic content based on the eight user needs (Shishkin et al., 2021) in his user needs model project. In the highly competitive news industry, it is imperative to extract value from every available resource. The value of data in enabling meaningful and profitable decisions is a key motivation for this project. The objective is to automate the classification process to facilitate the retrieval of valuable insights from consumers by media companies. This differential approach is intended to be combined with other valuable information.

The classification model developed achieved a general accuracy of 0.7. Despite this reasonable performance, it is evident that the provided data used for training introduced biases and resulted in an imbalanced dataset. This imbalance likely affected the model's performance, highlighting the need for balanced, properly classified data as a training set to improve the model's metrics.

The project work contributes to the field by applying ML tools, together with text mining techniques, to classify journalistic news articles. The application of this technology in this market is valuable, particularly for automated classification. However, it is important to avoid bias and ensure a good balance and diverse data.

This work contributes to the body of knowledge on the development of a ML model for media companies to classify news into eight user needs. It also contributes to the academic community by exploring interesting and real-world case studies, such as the use of imbalanced and biased data, and the most effective strategies for addressing these challenges in similar contexts.

The implementation of an automated classification model within the journalistic industry could facilitate improvements in working processes, resulting in enhanced efficiency and accuracy. This would enable a greater focus on data analysis, leading to more informed strategic decisions and ultimately, higher audience engagement. Furthermore, the reduction in error rates associated with the information obtained would enhance the overall accuracy of the model, thereby reducing the potential for misinformation.

One of the limitations of this study is the inconsistency present in the provided data, which resulted in an imbalanced dataset and affected model performance. This bias introduced systematic errors that potentially skewed the results, limiting the model's effectiveness in certain categories. Furthermore, the computational resources available were limited, which constrained the number of parameters or models that could be tested. Specific constraints included limited GPU availability and memory, which restricted the complexity of models that could be employed. Future research should focus on applying unsupervised learning to classify news articles, with particular attention paid to the balance of the training set. Additionally, it

would be beneficial to explore the integration of keyword-based approaches to improve classification reliability and readability by the model.

In conclusion, while the current model is of sufficient quality to overcome manual or random classification, it is important to focus on the quality of the training data, by applying unsupervised learning techniques, such as clustering methods or topic modeling, and to use more powerful and resource-intensive models to achieve the best possible accuracy in automating the process of classifying news articles. This study paves the way for further improvements in the application of ML to text classification. Finally, ethical considerations are mandatory when building such models, and when using such information, it is important to keep the data inside the organization.

BIBLIOGRAPHICAL REFERENCES

Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural Machine Translation by Jointly Learning to Align and Translate*. http://arxiv.org/abs/1409.0473

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*. http://arxiv.org/abs/1607.06520

Britannica, T. E. of E. (2024). journalism. In *Britannica*. https://www.britannica.com/topic/journalism

Canavilhas, J., & Rodrigues, C. (2013). The citizen as producer of information: a case study in the Portuguese online newspapers. In *Observatorio (OBS*) Journal* (Vol. 7, Issue 1). http://obs.obercom.pt.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*, 4171–4186.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Howard, J., & Ruder, S. (2018). *Universal Language Model Fine-tuning for Text Classification*. http://arxiv.org/abs/1801.06146

Jackson, J. (2002). Data Mining; A Conceptual Overview. *Communications of the Association for Information Systems*, *8*. https://doi.org/10.17705/1cais.00819

Joachims, T. (1998). *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. https://doi.org/https://doi.org/10.1007/BFb0026683

Kantardzic, M. (2019). PREPARING THE DATA. In *Data Mining: Concepts, Models, Methods, and Algorithms* (pp. 33–60). Wiley. https://doi.org/https://doi.org/10.1002/9781119516057.ch2

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent Convolutional Neural Networks for Text Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence* (1st ed., Vol. 29, pp. 2267–2273). https://doi.org/https://doi.org/10.1609/aaai.v29i1.9513

Lazarinis, F. (2007). Lemmatization and stopword elimination in Greek Web searching. *EATIS*, 1–4. https://doi.org/https://doi.org/10.1145/1352694.1352757

Lipton, Z. C. (2016). *The Mythos of Model Interpretability*. http://arxiv.org/abs/1606.03490

McDuff, D., Ma, S., Song, Y., & Kapoor, A. (2019). *Characterizing Bias in Classifiers using Generative Models*. http://arxiv.org/abs/1906.11891

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. http://arxiv.org/abs/1301.3781

Ong, C. J., Orfanoudaki, A., Zhang, R., Caprasse, F. P. M., Hutch, M., Ma, L., Fard, D., Balogun, O., Miller, M. I., Minnig, M., Saglam, H., Prescott, B., Greer, D. M., Smirnakis, S., & Bertsimas, D. (2020). Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. *PLoS ONE*, *15*(6). https://doi.org/10.1371/journal.pone.0234908

Powers, D. M. W., & Ailab. (2011). EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION. *International Journal of Machine Learning Technology* , *2*(1), 37–63. https://doi.org/https://doi.org/10.48550/arXiv.2010.16061

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, *1*(8), 9–9. https://github.com/codelucas/newspaper

Rayward-Smith, V. J. (2007). Statistics to measure correlation for data mining applications. *Computational Statistics and Data Analysis*, *51*(8), 3968–3982. https://doi.org/10.1016/j.csda.2006.05.025

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Schmidhuber, J., & Hochreiter, S. (1997). LONG SHORT-TERM MEMORY. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/https://doi.org/10.1162/neco.1997.9.8.1735

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Shen, L., & Zhang, J. (2016). *Empirical Evaluation of RNN Architectures on Sentence Classification Task*. https://doi.org/https://doi.org/10.48550/arXiv.1609.09171

Shishkin, D., Verhoeven, R., ten Teije, S., & Kuntze, E. (2021). *user needs model 2.0*. https://smartocto.com/blog/explaining-user-needs/

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

sparck jones, karen. (1972). A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Strubell, E., Ganesh, A., & McCallum, A. (2019). *Energy and Policy Considerations for Deep Learning in NLP*. http://arxiv.org/abs/1906.02243

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. http://arxiv.org/abs/1706.03762

Yun, S., Seung, L., & Woon Woo, Y. (2019). A Scraping Method of In-Frame Web Sources Using Python. *Proceedings of the Korean Institute of Information and Commucation Sciences Conference*, 271–274. https://koreascience.kr/article/CFKO201916842431266.page

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa