



RECOGNIZING REFUND VALUES IN COMPLAINTS AGAINST BRAZILIAN TELECOMMUNICATION COMPANIES

By

LEANDRO DE LIMA LIRA

Master Thesis in Modeling, Data Analysis
and Decision Support Systems

Supervised by:

Professor Dr. João Manuel Portela da Gama (FEP)

Dr. Thiago Pereira de Brito Vieira (ANATEL/Brazil)

Faculty of Economics

University of Porto

2021

Acknowledgements

I would like to express my gratitude to God, for allowing me to execute this work, as to my family and my friends for encouraging me. To Anatel and my colleagues in this institution, for supporting me in this period of studies. I am also so thankful to my supervisors, Dr. João Manuel Portela da Gama and Dr. Thiago Pereira de Brito Vieira for the patience, encouragement and advices given during all this work.

Abstract

The Brazilian telecommunications sector represents one of the largest markets in the world. The National Telecommunications Agency (ANATEL), the regulatory institution for telecommunications services in Brazil, provides a customer service platform to users of these services, called "Anatel Consumidor", which receives approximately 3.5 million complaints per year, which are stored in text format. Many of these complaints relate to customer refund cases.

In this scenario, the work aims to identify situations in which consumers request a refund and how much the customers are requiring, as well to identify situations in which companies provide some refund, and how much is the amount.

This research uses real and sensitive data, adopting Natural Language Processing (NLP) techniques with supervised learning from unlabeled data, which involves an effort of interaction with Anatel specialists.

The work follows the CRISP-DM methodology applied to NLP, building predictive models to classifying complaints using own heuristics and word embeddings. Besides that, the work uses Named Entity Recognition (NER) to identify amounts complained and refunded.

In summary, the research presents a concrete case study applying Natural Language Processing on complaints from the telecommunications sector, involving thematics of classification, sample labeling, regression, named entity recognition (NER), and development of own heuristics. In addition, for the public sector, the work presents an example of how to quantify part of the financial return that an institution generates for society, as well as the automation of activities that, due to their volume, would not be feasible to be done manually.

Keywords: complaints in the telecommunications sector, classification, quantification of social return, refund to consumers, natural language processing, named entity recognition, word embeddings.

Resumo

O setor de telecomunicações brasileiro representa um dos maiores mercados de consumo de serviços no mundo. A Agência Nacional de Telecomunicações (ANATEL), instituição reguladora dos serviços de telecomunicação no Brasil, disponibiliza uma plataforma de reclamações aos usuários destes serviços, chamada "Anatel Consumidor", por onde se realiza uma média anual de quase 3,5 milhões de reclamações, as quais ficam armazenadas em formato de texto livre. Muitas destas reclamações se referem a casos de reembolso aos consumidores.

Neste cenário, o trabalho visa identificar situações em que os consumidores solicitam reembolso e quanto solicitam, bem como as situações em que as empresas fornecem algum reembolso e qual o valor.

Trata-se de um trabalho realizado sobre uma base real, com dados sensíveis, utilizando técnicas de Processamento de Linguagem Natural com aprendizado supervisionado a partir de dados não rotulados, o que envolveu um esforço de interação com os especialistas da Anatel.

O trabalho seguiu a metodologia CRISP-DM aplicada a NLP, tendo sido construídos modelos preditivos para classificação das reclamações utilizando heurística própria e word embeddings. Além disto, o mesmo trabalho também faz uso de Reconhecimento de Entidades Nomeadas para identificação de valores cobrados e ressarcidos.

Em seu conjunto, o trabalho fornece um estudo de caso concreto aplicando Processamento de Linguagem Natural sobre dados reais em reclamações do setor de telecomunicação, envolvendo temáticas de classificação, rotulação da amostra, regressão, reconhecimento de entidades nomeadas e desenvolvimento de heurísticas próprias. Além disto, para o setor público oferece um exemplo de como se quantificar parte do retorno financeiro que uma instituição gera para a sociedade, bem como de automação de atividades que pelo volume seriam inviáveis de serem feitas manualmente.

Palavras-chave: reclamações no setor de telecomunicação, classificação, quantificação de retorno social, reembolso aos consumidores, processamento de linguagem natural, reconhecimento de entidades nomeadas, word embeddings.

Contents

1	Introduction and Motivation	1
1.1	Thesis structure	1
1.2	Understanding the problem	1
1.3	Proposals	9
1.4	Objectives	9
1.5	Notation	10
2	Literature Review	11
2.1	Corpus and the Concern with specific idiom	11
2.2	Regular Expressions	12
2.3	Natural Language Processing	13
2.3.1	Word Embedding	13
2.3.2	BERT	13
2.3.3	NLP toolkits	14
2.4	Named Entity Recognition	14
2.4.1	Different approaches in NER	15
2.5	Quantification in NLP	16
2.6	Classification Algorithms	16
2.7	Regression Tasks	18
2.8	Measures	18
2.8.1	Classification Measures	18
2.8.2	NER Measures	19
2.8.3	Regression Measures	20
2.9	Comparative with other studies	20
3	Methodology	21
3.1	Data Model	21
3.1.1	Untagged and Subjective Real Data	21
3.2	About Software	22
3.3	Software Methodologies	22
3.4	Adoption of a specific Data Mining project framework	22
3.5	About Source Code	23

4	Case Study	25
4.1	The Workflow	25
4.2	Exploratory Data Analysis	26
4.2.1	Defining the sample size to validation	27
4.2.2	Preliminary Results	28
4.3	Preprocessing text	34
4.4	Modeling	37
4.4.1	Structure of the problem	37
4.4.2	Train and Test datasets	39
4.4.3	Modeling the Classification Tasks	39
4.4.4	Tasks and opportunities to improve the model	41
4.4.5	Modeling the Regression/NER Tasks	43
5	Evaluation and Analysis of Results	47
5.1	Evaluation and Results in Classification Problems - Task 1 and Task 3	47
5.1.1	Evaluation in Classification Tasks	47
5.1.2	Results using the heuristic	48
5.1.3	Results using Word2Vec + metadata	48
5.1.4	Results using word2Vec + Heuristic	49
5.2	Evaluating and Results in Regression/NER Problems - Task 2 and Task 4	50
5.2.1	Evaluation in Regression/NER Tasks	50
5.2.2	Results using the first algorithm	52
5.2.3	Results using the second algorithm (considering the "weight" of the value)	53
5.3	Choosing the Models	54
6	Conclusions, Contributions and Future Works	56
6.1	Conclusions	56
6.2	Contributions	57
6.3	Future Work	57
	Bibliography	59

List of Figures

1.1	Complaint screen.	3
2.1	Application of QSearch (Ho, Pal, Kleer, Berberich, & Weikum, 2020)	16
2.2	Text Classification Task	17
2.3	Confusion Matrix	18
3.1	Project life cycle based on CRISP-DM.	23
4.1	NLP Workflow for the project	25
4.2	Classification Sheet Format	26
4.3	Ideal Sample Size Formula	28
4.4	Histogram of values complained by customers	32
4.5	Boxplot of values complained by customers	32
4.6	Histogram of values refunded by the companies	33
4.7	Boxplot of values refunded by the companies	33
4.8	Preprocessing Text Tasks	34
4.9	Prediction Task Types	38
4.10	Features as inputs to train classification models	41
4.11	Example of the application of the first heuristic to detecting the value refunded	45
4.12	Example of application of the second heuristic, with weights to the numbers	46

List of Tables

1.1	Number of requests submitted to Anatel Consumidor	2
1.2	Scenario 1: Cases in which there is no relationship with a claim for refund	4
1.3	Scenario 2: Cases where customers ask for a refund, but the company does not say if has refunded	5
1.4	Scenario 3: Cases where the company communicates the refund, expressing the amount	6
1.5	Scenario 4: Cases in which the company communicates some refund, but without express the amount	7
1.6	Scenario 5: Cases involving reference to values, but not related to refunds	8
2.1	Some regular expression meanings (Jurafsky, 2000)	12
2.2	Examples of regular expressions	12
2.3	Tuples identified by an NER task	15
2.4	NER Ruled Example for entity type <i>Value</i>	15
4.1	Result of Lowercasing and Cleaning subtasks over a complaint	36
5.1	Scores for Task 1 just using the heuristic	48
5.2	Scores for Task 3 just using the heuristic	49
5.3	Scores for Task 1 using Word2Vec + Metadata	49
5.4	Scores for Task 3 using Word2Vec + metadata	49
5.5	Scores for Task 1 by classifiers, using Word2Vec + Heuristic + metadata	50
5.6	Scores for Task 3 by classifiers, using Word2Vec + Heuristic + metadata	50
5.7	Scores for Task 2, considering the first algorithm	52
5.8	Scores for Task 4, considering the first algorithm	53
5.9	Scores for Task 2 considering the second algorithm	53
5.10	Scores for Task 4, considering the second algorithm	54

List of Acronyms

ANATEL	Brazilian National Agency of Telecommunication
BERT	Bidirectional Encoder Representations from Transformers
CRISP-DM	Cross Industry Standard Process for Data Mining
FEP	Faculty of Economics of the University of Porto
IEEE	Institute of Electrical and Electronics Engineers
ITU	International Telecommunication Union
MAD	Median Absolute Deviation
MAE	Mean Absolute Error
MSE	Mean Squared Error
NER	Named Entity Recognition
NLP	Natural Language Processing
RegEX	Regular Expression
RUP	Rational Unified Process
RMSE	Root Mean Squared Error
XP	Extreme Programming

Chapter 1

Introduction and Motivation

The telecommunication sector covers several services in the world: mobile communications, internet, satellites, radio, streaming and etc. More important, the majority of the population use these services, including all geographic, ethnic and social aspects. As evidence of this statement, according to statistics provided by the International Telecommunication Union (ITU), at the end of 2019 more than 51 percent of the global population was connected to the internet, and there was more than 6.6 billion mobile access, by phones, in the world. (Union, 2020)

In this context, the consumer telecommunication market is one of the biggest in the world, and consequently generates a huge number of complaints against the companies.

This work addresses a problem involving telecommunication complaints in Brazil. According to ANATEL, the Brazilian regulatory telecommunication institution, the monetary value that the customers ask to be refunded in their complaints, as well the value that the companies refund, are unknown. Not to mention that even the identification of the complaints that contain some ask for a refund is also a manual task.

1.1 Thesis structure

This thesis contains 5 main sections. The first one describes the details of the problem. The second section covers the literature review and the analysis of possible approaches to treat the problem. The third chapter creates and justifies a strategy to implement the solution, based on CRISP-DM framework over a Natural Language Processing problem. The fourth chapter covers the problem modeling, as a description of the problem solution. The fifth chapter analyses the results obtained, and the conclusions chapter describes the gains the work presents and future work.

1.2 Understanding the problem

As a reflection of a new model for regulating public services in Brazil, linked to the privatization of the telecommunications sector, in 1997 the National Telecommunications Agency (ANATEL) was created, with the responsibility of promoting competition and regulation in this

sector.

Internet, telephony, pay-TV, satellite communication, broadcasting services and others are under Anatel regulatory authority, which involves everything from the approval of equipment (telephones, antennas, drones, cables, etc.) until verification of the quality of the services.

Anatel created and has maintained a system, called "Anatel Consumidor" (Anatel Customer, in English), where the customers can submit requests about telecommunication services. Table 1.1 shows the number of requests submitted by customers in the last years, according to data published by Anatel. (de Telecomunicações, 2021)

Table 1.1: Number of requests submitted to Anatel Consumidor

Year	Quantity
2015	4.132.965
2016	3.962.021
2017	3.447.450
2018	2.982.349
2019	3.018.934
2020	3.007.538
Average 2015-2020	3.425.210

There are 4 ways to customers submit the requests:

- a) *In-person at any Anatel's office*, where the customer goes to an Anatel office and exposes to an attendant the request;
- b) *By Anatel Call Center*, where the customer makes a call and talk with an attendant;
- c) *By Anatel Web Site*, where the customer writes himself the request;
- d) *By Anatel APP*, where the customer writes himself the request.

The figure 1.1 shows the structure of a request in the Anatel Web Site. Here we identify some meta data related to the classification of the request, like request type, telecommunication service and subject matter.

The records are divided into four types: denunciation, complaint, suggestion and request for information. Anatel says that approximately 98% of requests are from the type *complaint*. (de Telecomunicações, 2021)

It is important to notice that, despite the different classifications, once the content of the request is input and stored in a free text field, it is possible to have requests where the content refers to different subjects and types of classification. For example, in a request from type "complaint" and subject "billing" it is possible to have content that includes also a suggestion, a denunciation and cover another type of subject.

As described in table 1.1, Anatel receives on average more than 3 million complaints from consumers in the telecommunications sector each year, through the "Anatel Consumidor" system. The registration of complaints can be made directly by consumers (through a smartphone application or website), through a telephone call to an available call center, or through personal

Campos com (*) são de preenchimento obrigatório!

Dados da solicitação

Protocolo na Prestadora(*)

Telefone com problema(*)

CPF do Assinante (*)

Nome do Assinante (*)

Reclamado (*)

Localização (*)

Descreva o Problema (*)

No caracteres: 1436

Figure 1.1: Complaint screen.

attendance at an Anatel unit (a situation where an employee will register the complaint in that system). In any of the forms of opening the complaint, it will be stored in Anatel Consumidor as a textual form, that is, there will be a record in the institution’s database with the textual report of the complaint. Likewise, the operators’ responses to the complaint, as well as the record of the service evaluation by Anatel itself, are also recorded in a textual form. At this point, it is important to highlight that, despite the different ways of entry result in a text in the database, there is a substantial difference between text wrote by an attendant and the text wrote by the customer himself. The attendant tends to be more clear and objective in writing than the customer, which can represent results below the expectations when processing complaints that customers have written themselves.

A large number of these requests are complaints that refer to “refund” problems, where the consumer claims that he was the victim of an undue charge or that for some other reason the telecom company owes him some monetary compensation. In this context, the Anatel business area presents the following problem:

- a) How much is requested by consumers complaining to Anatel Consumidor ?
- b) How much do the companies say they refund? ?

In order to understand the complexity besides these two questions, it is necessary to analyze some scenarios in complaint examples extracted from Anatel Consumidor.

Table 1.2: Scenario 1: Cases in which there is no relationship with a claim for refund

Idiom	Customer Complaint
English Version	<p>Even though ANATEL filed a complaint and was considered to be "well resolved", the operator ***** insists on offering plans in an INSISTENT AND ABUSIVE way using VIRTUAL sellers, in an ELECTRONIC way, through a VIRTUAL CALL ASSISTANT' ***** "through various numbers originating in the State of São Paulo (DDD 011) and the number ***** in order to DECEIVE THE CUSTOMER using the DDD device of the same city, masking a local call number . It asks for IMMEDIATE measures, since the complaint filed with ANATEL and falsely considered as "settled well" was not complied with.</p>
Portuguese	<p>Mesmo aberto reclamação na ANATEL e dada como "resolvida procedente", a operadora ***** insiste em ofertar planos de forma INSISTENTE E ABUSIVA usando vendedores VIRTUAIS, de forma ELETRÔNICA, através de uma ASSISTENTE VIRTUAL CHAMADA "*****" através de variados números originados do Estado de São Paulo (DDD 011) e do número ***** com o intuito de ENGANAR O CLIENTE usando o artifício do DDD da mesma cidade, mascarando um número de ligação local. Pede providências IMEDIATAS, já que não foi cumprido a reclamação aberta na ANATEL e falsamente dada como "resolvida procedente".</p>
	<p>Company answer</p>
English Version	<p>In response to complaint ID: ***** contact Mr. ***** on March 1, 2018 at 10:10 am, on the phone: *****. Under protocol: ***** , Regarding your request for marketing calls. Where we request the exclusion of numbers ***** to receive marketing calls. Customer aware of the deadline of 10 working days for the calls to cease. If there are any doubts or need further clarification in relation to this closing, the consumer protection agency relationship area remains at your disposal for further telephone contact, forward request via e-mail: ***** *** (forward e-mail with one of the holder's details).</p>
Portuguese	<p>Em atenção à reclamação ID: ***** em contato com Sr. ***** dia 01/03/2018 às 10:10, no telefone: *****. Sob protocolo: ***** , Referente a sua solicitação de ligações de marketing. Onde solicitamos a exclusão dos números ***** para recebimento de ligações de marketing. Cliente ciente do prazo de 10 dia uteis para as ligações cessarem. Casos tenha ficado alguma dúvida ou precise de mais esclarecimentos em relação a este fechamento, a área de relacionamento de órgãos de defesa do consumidor permanece a sua disposição para novo contato telefônico, encaminhar solicitação através do e-mail: ***** (encaminhar e-mail com um dos dados do titular).</p>

Table 1.3: Scenario 2: Cases where customers ask for a refund, but the company does not say if has refunded

Idiom	Customer Complaint
English Version	<p>I made a ***** plan in February, I was never able to use the chip, because there is no network in my house, I want to cancel this plan and I want to return the amounts paid, since I was never able to use the service! reimburse me I will sue ***** for moral damages, because that is absurd! I WANT CANCELLATION AND REFUND on the Agency Account: **** account: ***** (*****)</p>
Portuguese	<p>Fiz um plano da ***** em fevereiro, nunca consegui utilizar o chip, pois não tem rede na minha casa, quero cancelar esse plano e quero a devolução dos valores pagos, já que nunca consegui utilizar o serviço! Caso não me reembolsem eu vou entrar com uma ação contra a ***** por danos morais, por que isso é um absurdo! QUERO CANCELAMENTO E REEMBOLSO na Conta agência: **** conta: ***** (*****)</p>
English Version	<p>Company answer</p> <hr/> <p>Please note ID ***** and Internal Protocol ***** Contact Mr. ***** on 06/07/2019 at 16:22 on the phone ***** to apologize and inform that our goal is to provide a according to the expectations of our customers. In contact with the customer, we inform that the cancellation of the plan was made, as well as the exemption from the loyalty penalty, making it clear that there is no procedure for returning the amounts paid regarding the invoices, as we were able to contact the customer on the claimed line, in the system no no signal has been identified and the customer has 7 working days to cancel if there is any discrepancy, the more the customer has the active plan for 4 months, the line is now on prepaid as requested by the same, if there is any doubt or In the event of your request, we ask that you contact us through ***** ...”</p>
Portuguese	<p>Em atenção a ID ***** e Protocolo interno ***** Em contato com o Sr. ***** na data de 07/06/2019 às 16:22 no telefone ***** para nos desculpar e informar que nosso objetivo é proporcionar um atendimento de acordo com as expectativas de nossos clientes. Em contato com o cliente informamos que foi feito o cancelamento do plano e também a isenção da multa de fidelização, deixando claro que não existe procedimento de devolução dos valores pagos referente as faturas, pois conseguimos contato com a cliente na linha reclamado, no sistema não foi identificado ausência de sinal e cliente tem 7 dias uteis para efetuar o cancelamento caso ocorra alguma divergência, mais cliente esta com o plano ativo a 4 meses, a linha agora se encontra no pré-pago como solicitado pela mesma, caso ocorra alguma duvida ou intercorrência referente à sua solicitação solicitamos que entre em contato conosco através do *****”</p>

Table 1.4: Scenario 3: Cases where the company communicates the refund, expressing the amount

Idiom	Customer Complaint
English Version	<p>...IN THE MONTH OF DECEMBER, I RECEIVED AN INVOICE INVOICE OF 117.30, WHICH SHOULD BE ONLY R\$ 51.73, EQUIVALENT TO *****. THE AMOUNT OF R\$ 65.57 WAS CHARGED, UNDERSIDENTLY, FOR SERVICES THAT WERE NOT HIRED! IN THE MONTH OF NOVEMBER, ***** WAS R\$ 87.87 AND ***** R\$ 26.40. (PAY INQUIRY). FROM THIS DONE, I AM REQUIRING THE RETURN OF THE PAID AMOUNT REFERRING TO THE UNDUE PAYMENT, IN CHARGE, ACCORDING TO CDC, IN THE IMPORTANCE OF R\$ 52.80...</p>
Portuguese	<p>...NO MES DE DEZEMBRO, RECEBI FATURA COM O VALOR INDEVIDO DE 117,30, O QUE DEVERIA SER APENAS R\$ 51,73, EQUIVALENTE AO *****. FOI COBRADO O VALOR DE R\$ 65,57, INDEVIDAMENTE, POR SERVIÇOS QUE NAO FORAM CONTRATADOS! NO MES DE NOVEMBRO, O ***** FOI DE R\$ 87,87 E ***** R\$26,40. (PAGO INDEVIDAMENTE). DESTA FEITA, VENHO REQUERER A DEVOLUÇÃO DO VALOR PAGO REFERENTE A COBRANÇA INDEVIDA, EM DOBRO, CONFORME CDC, NA IMPORTÂNCIA DE R\$ 52,80. ...</p>
<u>Company answer</u>	
English Version	<p>In consideration of ID ***** ,referring to contestation protocol n° (Not informed), we contacted Mr. ***** (Holder) on 01/23/2019, at 18h47min., On the phone claimed ***** ,to inform that, after analyzing the dispute, we found that, in the invoice 12/2018 of R\$ 117.30 due on 1/16/2019, there was an undue charge of R\$ 64.52, referring to the undue value of the contracted service ***** . Thus, an adjustment of R\$ 64.52 was made on the 12/2018 invoice of R\$ 117.30 and a second corrected copy was sent in the amount of R\$ 52.78, with maturity extended to 02/02/2019. We concluded that the reimbursement was made in the amount unduly paid of R\$ 52.80, which will be credited in the next invoice.</p>
Portuguese	<p>Em atenção à ID ***** , referente ao protocolo de contestação n° (Não informado) ,contatamos o Sr. ***** (Titular) em 23/01/2019, as 18h47min., no telefone reclamado ***** , para informar que, após análise da contestação, verificamos que, na fatura 12/2018 de R\$117,30 com vencimento em 16/01/2019, ocorreu a cobrança indevida de R\$64,52, referente ao valor indevido do serviço ***** contratado.Assim, foi realizado ajuste no valor de R\$64,52 na fatura 12/2018 de R\$117,30 e foi encaminhada 2ª via corrigida no valor de R\$52,78, com vencimento prorrogado para 04/02/2019.E concluímos que foi realizado o ressarcimento no valor pago indevidamente de R\$52,80, que será creditado na próxima fatura. ...</p>

Table 1.5: Scenario 4: Cases in which the company communicates some refund, but without express the amount

Idiom	Customer Complaint
English Version	<p>Good afternoon. Incredible as every time we have a problem with *****, it becomes a total neglect. It had been a while since I had no phone line and therefore no internet. I filed a complaint yesterday 02/18 and the person was so kind that I found it strange. He scheduled a technician for assistance in 24 hours and for a change, as usual in *****, he did not show up. When calling hj again, automatically at Ura it says that I will repair it tomorrow until 11 am or that is outside the 24 hours that they set out to solve the problem and when talking to the attendant ***** hj, 02/19, according to protocol *****, I was informed that you have to wait. How long will we consumers have to endure this neglect on account of the operators ??? I don't settle for that. What is the solution??? And so I go without a phone!</p>
Portuguese	<p>Boa tarde. Incrível como todas as vezes que temos problema com a *****, vira um descaso total. Há algum tempo que não ficava sem linha telefônica e conseqüentemente sem internet. Registei reclamação ontem 18/02 e a pessoa foi tão gentil que estranhei. Agendou técnico para atendimento em 24 horas e para variar, como é de costume na *****, não apareceu. Ao ligar hj de novo, automaticamente na Ura já fala que reparo amanhã até as 11 horas ou seja fora das 24 horas que se propuseram à resolver o problema e ao falar com a atendente ***** hj, 19/02, conforme protocolo *****, fui informada que tem que aguardar. Até quando nós consumidores vamos ter que aguentar esse descaso por conta das operadoras??? Eu não me conformo com isso. Qual a solução??? E assim sigo sem telefone!</p>
	<p>Company answer</p>
English Version	<p>Please note ID: ***** protocol ***** as per contact with Mrs. ***** 02/25/19 16:48 at number *****, it confirms the operation of the service, we make the credit refund on the next invoice for 3 days without using the service. If you have any questions or need further clarification regarding this closing, the Consumer Protection Bodies relationship area remains at your disposal for further telephone contact, forward request via e-mail **** *. Any doubt I am available.</p>
Portuguese	<p>Em atenção a ID:***** protocolo ***** conforme contato com Sra ***** 25/02/19 16:48 no número *****, a mesma confirma o funcionamento do serviço, realizamos o ressarcimento em credito na próxima fatura referente a 3 dias sem a utilização do serviço. Caso tenha ficado alguma dúvida ou precise de mais esclarecimentos em relação a este fechamento, a área de relacionamento de Órgãos de Defesa do Consumidor permanece a sua disposição para novo contato telefônico, encaminhar solicitação através do e-mail *****.Qualquer dúvida estou a disposição.</p>

Table 1.6: Scenario 5: Cases involving reference to values, but not related to refunds

Idiom	Customer Complaint
English Version	<p>... On 10/14 I got in touch because I noticed that ***** unilaterally canceled the 15 Gigabyte Bonus and implemented a plan called ***** , which is more limited than the Bonus I contracted and started to charge more expensive than the contracted, I complained and they granted a discount, however they have been charging R\$ 279.97 and not R\$ 239.99. I complained again on 04/20, about the value, they propose a new discount only they want me to sign a loyalty contract, which in addition to seem illegal, is unfair, because the mistake is theirs and not mine, they are breaking the contract. Before going to court, I would like ANATEL's intersection in this case. Sincerely *****.</p>
Portuguese	<p>... Em 14/10 entrei em contato pois verifiquei que a ***** , unilateralmente, cancelou o Bonus contratado de 15 gigabytes e implantou um plano chamado ***** , que é mais limitado que o Bonus que contratei e passou a cobrar mais caro do que o contratado, reclamei e eles concederam um desconto, no entanto vêm cobrando R\$ 279,97 e não os R\$ 239,99. Voltei a reclamar no dia 20/04, sobre o valor, eles propõe um novo desconto só que querem que eu assine um contrato de fidelização, o que além de me parecer ilegal, é injusto, pois o erro é deles e não meu, eles estão descumprindo o contrato. Antes de entrar na justiça, gostaria de interseção da ANATEL nesse caso. Atenciosamente *****.</p>
	<p><u>Company answer</u></p>
English Version	<p>... Therefore, we correct the invoices with due date ***** , R\$ 254.97, adjustment in the amount of R\$ 14.98, due on 12/2018 invoice ***** , R\$ 267.40, adjustment in the amount of R\$ 27.41, due on 01/2019 invoice ***** , R\$ 265.54, adjustment in the amount of R\$ 25.62, due on 02/2019 invoice ***** , R\$ 264.98, adjustment in the amount of R\$ 24.55, maturity on 03/2019 invoice ***** , R\$ 296.62, adjustment in the value of R\$ 41.18, due on 4/4/2019 invoice ***** , R\$ 279.97, adjustment in the amount of R\$ 39.98 and due on 5/5/2019 invoice ***** , R\$ 287.79 adjustment in the amount of R\$ 47.80. ...</p>
Portuguese	<p>.. Sendo assim corrigimos as faturas com vencimento em fatura ***** , R\$ 254,97 , ajuste no valor de R\$14,98, vencimento em 12/2018 fatura ***** , R\$ 267,40 , ajuste no valor de R\$27,41 , vencimento em 01/2019 fatura ***** , R\$ 265,54, ajuste no valor de R\$ 25,62 , vencimento em 02/2019 fatura ***** , R\$ 264,98, ajuste no valor de R\$24,55, vencimento em 03/2019 fatura ***** , R\$ 296,62 , ajuste no valor de R\$41,18, vencimento em 04/2019 fatura ***** , R\$ 279,97 , ajuste no valor de R\$39,98 e vencimento em 05/2019 fatura ***** , R\$ 287,79 ajuste no valor de R\$47,80. ...</p>

Scenario 1, presents in table 1.2, shows cases in which there is no relationship with a claim for refund. Scenario 2, presents in table 1.3, shows cases where the customer asks for a refund, but the company does not say anything about it. Scenario 3, present in table 1.4, exposes cases where the company communicates the refund, expressing the amount. Scenario 4, present in table 1.5, shows situations in which the company communicates a refund, but without express the amount. Finally, scenario 5, present in table 1.6, shows situations involving values, but that are not related to refunds.

After the analysis of these scenarios, we can observe several particularities :

- We need to analyze both the complaint and the company answer because in some cases the customer is not clear about his intentions, and despite that, the company refund him even without a clear asking;
- The refund values are described in different ways, using just numbers, with or without currency symbols, using mathematical operators present in expressions like 'in double' or 'difference';
- We need to analyze the text in a semantic level, in order to identify if the value described refers to a refund or not, and to identify if the refund was in fact realized.

1.3 Proposals

The work presents a problem that combines non-tagged data, supervised learning, classification and regression tasks, named entity recognition, and the development of own heuristics.

Hence, the proposal of this work is the adoption of recognized methodologies, like CRISP-DM, to a real problem involving natural language processing over real data of Brazilian telecommunication sector, through the use of a combination of well known techniques and tools.

1.4 Objectives

The main objectives of this work regarding Anatel, other public institutions and the academic community are :

1. To present a concrete case study applying Natural Language Processing over sensitive data in real telecommunication complaints involving classification, regression and Named Entity Recognition (NER) tasks;
2. To be an example of how to quantify part of the financial return that the existence of a public institution generates for society;
3. To present analysis about different ways the customer express their asking for a refund, as well the different ways the companies express their refunding;
4. To present statistical analysis about customer complaints and companies answers over a sample extracted randomly and classified by Anatel specialists;

5. To construct heuristics for identifying customer complaints involving asking for refund and companies answers involving refunding;
6. To construct heuristics for identifying how much customers have asked for a refund and how much companies have refunded;
7. To construct an automated process to identify situations where the customers have asked for a refund, and identify the value asked;
8. On the other hand, to construct an automated process to identify situations where the companies have refunded something, and identify the value refunded;
9. To compare the results between simple heuristics versus more sophisticated representation models;
10. To be, with a didactic and accessible language even for non-specialized people, a reference for data science teams in the public sector about opportunities to apply data mining in their institutions.

1.5 Notation

The monetary values in this work are expressed using Brazilian currency, called *REAL*. The *REAL* is represented by the symbol "R\$", and adopt "," to split decimals and "." to split hundreds, thousands, millions and so on.

For instance, the value R\$ 2.523.105,87 represents 2 million, five hundred and twenty-three thousand, one hundred and five *reais* and eighty-seven cents.

Chapter 2

Literature Review

In order to identify the literature studies related to the problem, we realize searches in following academic platforms: google scholar, Web Of Science, Scopus and IEEE.

The main terms we use in searches are the following: "extração elementos relevantes", "information retrieval in text", "information retrieval monetary values", "natural language processing monetary values", "information retrieval complaints communications", "information retrieval complaints systems", "recognizing monetary values from text", "text mining for monetary values", "entity recognition monetary values", "recognizing prices in text", "retrieval prices in text", "quantity search", "refund named entity recognition", "survey named entity recognition", "ner quantity search", "ner quantities", "NLP quantities", "statistic relevant samples size" and similars.

We figured out that the academic community addresses the problem better using the terms "quantity search", "NER quantities", "numeracy" and related.

We also identify that the literature recognizes the issue in the analysis as a Natural Language Processing (NLP) problem. NLP refers to a large and interdisciplinary field of concepts and techniques in order to recognize the natural way humans talk in different communication channels: speech, write and so on (Jurafsky, 2000),

2.1 Corpus and the Concern with specific idiom

In its turn, the literature highlights the importance of adopting an approach observing the specific idiom spoken, like English, Arabic, Portuguese etc. Not just because there are different words, but also because each idiom represents particular ways of express meaning (Ferreira, 2019) and (de Aquino Silva, da Silva, Dutra, & de Araujo, 2020).

Usually NLP systems implement a pipeline (a sequence of activities), where each task has as input the output of the previous task. In this context, first tasks problems will spread in all the pipeline.

In addition, because of the specificity of each language, the NLP approaches concern in observe since the first task the specific idiom. We do this, among other things, adopting a *Corpus* of the specific language. A *Corpus* is a collection of spoken or written natural language data that

is analyzed with linguistic criteria in order to understand better the idiom and construct better NLP systems (Thanaki, 2017). In practice, a Corpus is the main input of an NLP project.

2.2 Regular Expressions

One of the most known tools in computer science, Regular Expression (RegEX) is a language that consists in defining patterns in text. A regular expression, or RE, describes strings of characters (words or phrases or any arbitrary text)(Kaur, 2014).

According to this language, patterns are specified using special characters. Table 2.1 presents some of them, and their respective meaning.

Table 2.1: Some regular expression meanings (Jurafsky, 2000)

Character	Regular-expression meaning
.	Any character, including white-space or numeric
?	Zero or one of the preceding character
*	Zero or more of the preceding character
+	One or more of the preceding character

In this context, and remembering that the currency in Brazil is the "Real", represented by the symbol 'R\$', table 2.2 presents examples of regular expression that could be used to establish text patterns of interest :

Table 2.2: Examples of regular expressions

Regular Expression	Pattern represented
<code>\\R\$[0-9]*\\,[0-9]{2}</code>	Reais with 2 optional decimal places, like "R\$32,25"
<code>\\[0-9]*\\,[0-9]{2} \\reais\\</code>	With no currency symbol, like "32,25 reais"

In a first view, we could address our problem by adopting regular expressions, identifying and constructing several patterns corresponding to different ways of express money in the Portuguese language and Brazilian currency.

However, these approaches would conduct to innumerable problems. First of all, once the data are introduced through free text fields filled by the telecommunication users, there are countless possibilities of text representing monetary values, including cases with errors in orthography.

Additionally, the monetary values can be expressed not only with numbers but also with words (for instance "thirty-five dollars and two cents"), which increases considerably the complexity of the problem to solve with regular expressions.

Moreover, the problem involves not just identify monetary values, but also identify if these values are in the context of complaints that ask for some refund. This issue adds to the task a need for semantic interpretation because there are cases where the monetary values described are not associated with refunds.

Therefore, to the best of our knowledge, the use of regular expressions would not cover properly the complexity of the problem, despite we can adopt to solve part of that.

2.3 Natural Language Processing

As discussed before, Natural Language Processing (NLP) is an important and growing branch in data science, once focus on understanding the natural way humans communicate with each other in different ways: speech, text, signals etc. In the next sections, we present some of the important NLP techniques that have utility to our work.

2.3.1 Word Embedding

It is convenient to highlight the evolution of the way we analyze and represent the words and expressions in NLP tasks. First of all, it is understood that similarities are very important attributes to identifying the meaning of a word or a phrase. Based on the similarities concept, the context (neighbor words) in which the word appears in similar situations will define the semantic meaning of the word.

In the context of the neural networks approaches in NLP, in the last years the use of *Word Embeddings*, as a way to represent text (a word or sentence) in a vector in space, has emerged and obtained better results.. This vector stores in dimensions the context the word is presented, where each dimension can represent multiples meanings (Mikolov, Chen, Corrado, & Dean, 2013). Three of the most known word embeddings approaches are *Word2Vec*, *FastText*, and *Doc2Vec* (Gualberto, 2020).

In the first one, Word2Vec, relationships between words, their meanings and uses in similar contexts with other words are stored in spatial vectors. In this context, the vector space will contain closer each one that vectors of words used in a similar context.

In the second, FastText, a word is represented as a bag of character n-grams. In its turn, each of these character n-grams is represented by a vector, and the words are represented as the sum of these representations (Bojanowski, Grave, Joulin, & Mikolov, 2017).

By the end, Doc2Vec is an unsupervised learning approach that implements distributed representations of sentences and documents proposed. This algorithm learns a fixed-length dense vector representation for any variable-length pieces of texts (Gualberto, 2020).

2.3.2 BERT

The most recent literature also emphasizes Bidirectional Encoder Representations from Transformers (BERT) as one of the most efficient (probably the most) language representation models nowadays.

BERT is presented as a conceptually simple and empirically powerful language, that obtained the best scores on eleven natural language processing tasks (Devlin, Chang, Lee, and Toutanova (2018)).

The creators of BERT argue that current models (before BERT) had the limitation to be unidirectional, and this limits the choice of architectures that can be used during pre-training (Devlin et al., 2018). They emphasize that the others approach restrict the possibilities of the pre-trained representations. Whith this in mind, they propose a bidirectional representation, in order to obtain gains provided by the fine-tuning approaches.

Because of the simplicity and powerful results, BERT has been strongly adopted. As evidence of this, in October 2019, Google announced an important update in its environment: the adoption of BERT in its search algorithm (Nayak, 2019).

Despite represents the *State of the Art* in NLP, we do not adopt BERT in this work, what represents an opportunity for improvement in future work.

2.3.3 NLP toolkits

In order to help the NLP system developers, several toolkits exist implementing the pipeline tasks. As discussed before, we need to observe if the toolkit provides support to the specific idiom of the data that will be analyzed.

Another important point to observe in the choosing of the toolkit is the software or computational language in which it is implemented: Python, R, Java etc.

Regarding this topic, in order to follow the Anatel computational environment, we decide to search for NLP toolkits that observe the following requirements:

- support for Brazilian Portuguese idiom;
- implemented in Python language;
- implements NER tasks.

In this context, we identify the following toolkits: Spacy, NLTK, NLPyPort (Ferreira, 2019) and Stanza (Qi, Zhang, Zhang, Bolton, & Manning, 2020).

2.4 Named Entity Recognition

One important branch of Natural Language Processing (NLP), in general words Named Entity Recognition (NER) consists in define target entities (like name, address, price, etc.) and identify in the text these elements.

We can also describe NER as one of the tasks in the pipeline the represents NLP activities. Once NER usually is not one of the first tasks in the NLP pipeline, we need to execute before some NLP preprocessing task to be able to obtain the results expected by the NER task.

In a formal definition, we can describe NER as a task where, given a sequence n of tokens $S = (w_1, w_2, \dots, w_N)$, we expect to output a list of tuples (Is, Ie, t) , each of which is a named entity mentioned in S . Here, Is contained in $[1, N]$ and Ie contained in $[1, N]$ are the start and the end indexes of a named entity mention; t is the entity type from a predefined category set. (Li, Sun, Han, & Li, 2020)

For instance, defining *Person*, *Location* and *Date* as types of entities, we can identify the application of NER in the phrase:

”Vasco da Gama was born in Sines, Portugal, in 1469.”

Where $S = \{w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9, w_{10}, w_{11}, w_{12}\}$ and $w_1=$ ”Vasco”, $w_2=$ ”da”, $w_3=$ ”Gama”, $w_4=$ ”was”, $w_5=$ ”born”, $w_6=$ ”in”, $w_7=$ ”Sines”, $w_8=$ ”, $w_9=$ ”Portugal”, $w_{10}=$ ”, $w_{11}=$ ”in”, $w_{12}=$ ”1469”, $w_{13}=$ ”.”

In this context, table 2.3 shows the tuples detected by the NER task

Table 2.3: Tuples identified by an NER task

Tuple	Value
$\langle w1, w3, Person \rangle$	Vasco da Gama
$\langle w7, w7, Location \rangle$	Sines
$\langle w9, w9, Location \rangle$	Portugal
$\langle w12, w12, Date \rangle$	1469

2.4.1 Different approaches in NER

In the beginning, NER were implemented using handcrafted rules, lexicons, orthographic features and ontologies.(Nadeau & Sekine, 2007) However, in the past years, NER has been improved and implemented using several strategies, some of which are presented in this subsection.

(Li et al., 2020) propose this division to the traditional NER approaches:

a) Rule-based Approaches: In this approach, it is necessary to define by hand syntactic-lexical patterns to recognize the entities. In general, a rule follows the structure: *ContextualPattern* \rightarrow *Action*.

Table 2.4 presents an example of rule, in pseudo-code:

Table 2.4: NER Ruled Example for entity type *Value*

Rule 1: ValueByCurrency
If token[i] is "R\$" and (token[i+1] is number) { match }
Action: match.NamedEntity = {kind=value, rule="ValueByCurrency"}

Ruled-based approaches are good strategies in situations with a well-defined domain, where the lexicon is exhaustive. On the other hand, these systems cannot be used in other domains.

b) Unsupervised Learning Approach: In this case, the strategy consists of, from a huge corpus, extract entities by similarity. Here, techniques like clustering, lexical patterns and statistics can be used to associate words or expressions to the named entities.

c) Feature-based Supervised Learning Approaches. By this strategy, someone annotates data samples, identifying examples of the named entities. As a result of this action, features are created representing the training data.

After the training step, the algorithm will be applied over unseen data. Therefore, it is a case of a multi-class machine learning approach.

Despite these 3 approaches described before are consistent, dependent on the nature of the problem, in the few years are emerging new approaches based on the use of Deep Learning, that have raised the performance of NER tasks (Yadav & Bethard, 2019).

One important characteristic of the neural networks in NER tasks is that they require minimal feature engineering effort because these models are not domain dependant. Moreover, once this approach has presented better results for complex domains, it comes been very popular in the last few years.

2.5 Quantification in NLP

NER is a very powerful and useful technique to identify monetary values in a text, once it identifies easily numbers preceded by tags that represent currency, like ”\$”. Despite this facility, the problem grows when the task involves interpreting the semantic of this value in the text and its numeracy. Besides that, the problem becomes bigger when the task involves making mathematical operations between different values in the text.

Recent studies describe the problem of comparing, sorting, and adding numbers in natural language, even for embedding models that represent the State of Art in NLP, like BERT. The study concludes that the encode numeracy problem in NLP presents good results just in pre-trained datasets, and that is difficult to extrapolate beyond the values seen during training. Besides that, the authors highlight that the findings involving numeracy problems are still incipient and represents a fruitful area for future research (Wallace, Wang, Li, Singh, & Gardner, 2019).

In the same sense, (Ho et al., 2020) highlights the challenge of numeracy in NLP and propose a prototype (*QSearch*) to solve NLP problems involving numeracy that involves detecting values using NER and relevant tokens that will help to understand the meaning of the number in the context, generating triples (e,q,X) - (*entity*, *quantity*, *context*), called *Qfact*. Figure 2.1 shows the application of the QSearch over a sentence.

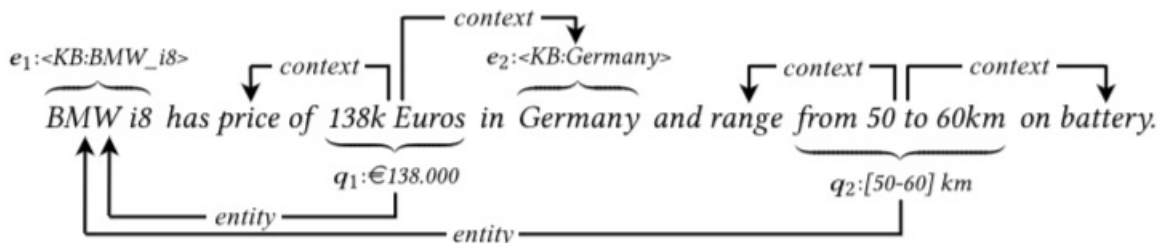


Figure 2.1: Application of QSearch (Ho et al., 2020)

2.6 Classification Algorithms

There are important steps in our work that involves identifying which complaints involve an ask for a refund (by the customer’s complaints side), and identify in which cases companies have refunded something.

These classification steps are important because there are situations that involve values, but that are not related to ”refund”. On the other hand, some situations involve asking for a refund, but where the values were not expressed. Once analyzing the companies answers, some situations involve values, but that are not related to refund, and there are situations where the company has refunded something, but without express the amount.

Classification is a well-addressed topic in data science and has a large use in bank fraud systems, spam detection, medical diagnosis, and others.

We identify as classification tasks the split of the customer complaints in classes "ask for refund" and "does not Ask for refund", as well the split of the companies answers in the classes "Company refunded" and "The company did not refund".

There are several classification algorithms, each one with its implementation strategies, where the problem features will define the algorithm that better fit the classifier model. Because of this, it is very common to apply more than one classification technique to the same problem, in order to compare the results and choose one.

Some of the best-evaluated classification algorithms are Support Vector Machine (SVM), the Naïve Bayes Classifier (NBC) , k-nearest neighbor (KNN), Multiple Instance Learning (MIL), Decision Tree (DT), Random Forest (RF), XGradient Boosting (XBG) (Kowsari et al., 2019).

In general lines, each of this algorithms uses features of the data as entry and output the the classes each case belongs. Classification tasks are part of the supervised learning concept. In supervised learning approach the algorithms are trained with target values (the classes) tagged previously, in order to understand the patterns in the data and be able to make predictions. Figure 2.2 presents this workflow:

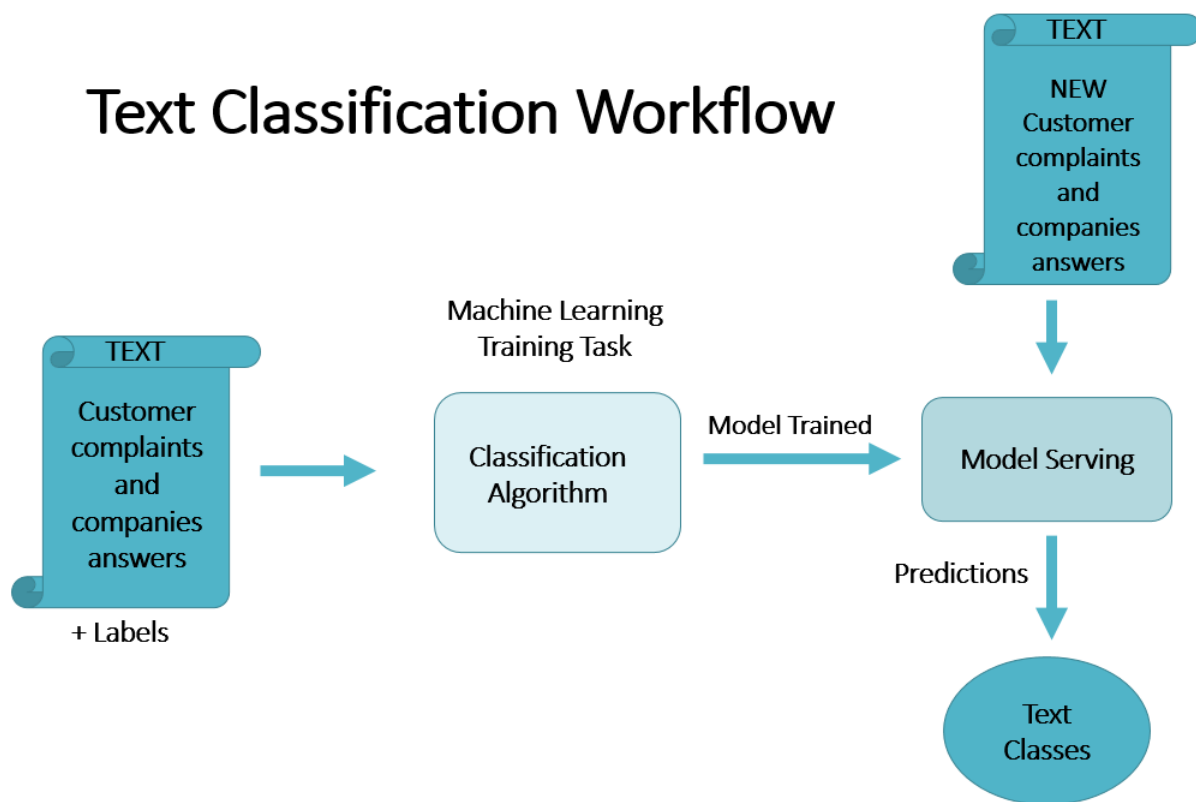


Figure 2.2: Text Classification Task

Because each algorithm fits better the model depending on the feature characteristics (the problem nature), in this work we use some of these classifiers in order to compare the results, fit the model and improves the scores of the classification task, in an incremental approach.

2.7 Regression Tasks

We can define regressions problems as prediction tasks where the target value is not a class or type, but a number. In the definition of our problem, it is possible to identify this nature, once we want to identify the monetary value asked for a refund by the customer, and the monetary value refunded by the company.

By definition, the algorithms for regression predictions work with a set of features as input and have as output a numeric value. This numerical value is obtained by a function that will generate outputs in a continuous and ordered set. (DA GAMA et al., 2017, pp. 71–72)

Hence, it is possible to identify that the problematic nature of this work has a difference with general regression problems because to identify the values complained and refunded, we need to identify which value, or values, inside text represent the target value, what in some cases will be obtained just after some calculations over it.

2.8 Measures

We must evaluate the efficiency of the tasks in order to identify the best algorithms and models, and then take decisions about the next steps constantly until chose the model to be deployed.

Because of the nature of each task, exist a group of possible measures that will be appropriate. And depending on the business problem nature, we will choose one or a combination of the measures.

2.8.1 Classification Measures

Regarding classification tasks, the understanding of the measures requires us to present the concept of the confusion matrix. This matrix, described in figure 2.3, gives us a way to interpret the results of a prediction task.

		Predicted Values	
		Negative	Positive
Actual Values	Negative	TN	FP
	Positive	FN	TP

Figure 2.3: Confusion Matrix

In order to understand these classification measures we need to observe the following concepts:

- True Negative (TN) represents a case in which neither the real value, neither the predicted value are positives;

- False Positive (FP) represents a case in which the value predicted as true is not true in the real world;
- False Negative (FN) represents a case in which the predicted value returns negative but it is true (exists) in the real world;
- True Positive (TP) represents the case in which the positive value predicted really is positive in the real world.

In the confusion matrix, the True Negative (TN) and True Positive (TP) represent the right predictions, while False Negative (FN) and False Positive (FP) represent the wrong predictions.

With these statements in mind, we can present the definition of the common classification measures below:

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN} \quad F - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Moreover, the *Accuracy Score* is an important score in classification tasks, and represents the simple percentage of correct classifications.

2.8.2 NER Measures

In general lines, we can evaluate the performance of NER tasks by comparing the outputs provided by the software against the human annotations (Li et al., 2020). There are different methodologies available to do this, which one of the most adopted refers to "*exact match evaluation*". According this strategy, both *boundary detection* and *type identification* NER subtasks are evaluated at the same time, in order to obtain final scores. In this context, we use the traditional measures of False Positives (FP), False Negatives (FN) and True positives (TP) to obtain the following NER measures: *Precision*, *Recall*, and *F-score*.

To better understand these measures in the context of NER, we need to observe the particularities in the concepts that differs from the traditional classification tasks:

- False Positive (FP) represents a case in which the entity returned by the NER task is not true in the real world;
- False Negative (FN) represents a case in which the entity is not returned by the NER task but it is true (exists) in the real world;
- True Positive (TP) represents the case in which the entity returned by the NER task really exists in the real world.

Hence, in NER context *Precision* represents the the number of entities a system predicted correctly divided by the number that the system predicted. In its turn, *Recall* represents the number of entities a system predicted correctly divided by the number that were identified by the human annotators. By the end, *F-score* is a measure that combines precision and recall (Yadav & Bethard, 2019).

2.8.3 Regression Measures

The tasks involving detecting the monetary value complained and refunded need a measure that points to the deviation between the predicted and the real value. In this context, regression tasks present appropriate measures to evaluate them.

Four of the most common measures for regression problems are the Mean Squared Error (MSE), the Root-Mean-Squared-Error (RMSE), the Mean-Absolute-Error (MAE), and the Median Absolute Deviation (MAD).

The first one, MSE, computes the average of the squared difference between the real value and the value predicted. The RMSE computes the square root of the average square difference between the real value and the value that was predicted. The MAE computes the absolute difference between the real value and the value predicted. By the end, the MAD is computed based on the median (instead of the mean) of the values and is more resistant to outliers than the other measures.

2.9 Comparative with other studies

There are studies related to telecommunications and energetic sector complaints, but not covering the quantification refund issue (Sousa et al., 2020), (Hui-nan, Jiang-ning, & Yan-zhong, 2006) and (de Freitas Junior, 2013).

Still using customer data, (Grljević & Bošnjak, 2018) focus on sentiment analysis, but not covering classification tasks over telecommunication customer complaints.

There are also studies that cover the retrieval of monetary values in different contexts, but not in compliant systems (Banerjee, Chakrabarti, & Ramakrishnan, 2009) and (Ho et al., 2020).

As discussed before, the issue of recognizing monetary values in a text, and at the same time interpret the meaning of this value in the context of the message, represents a branch of research still incipient.

On the other hand, the task of classification text in two classes (YES or NO) is well addressed in the literature, reaching the semantic level of meaning by using modern techniques like *word embeddings*.

In this context, this work represents a new object study to the academic community, once address together topics involving real and sensitive telecommunication complaints data, supervised learning in data initially not tagged, customer text classification, companies answers classification, monetary quantification from the text wrote by customers and companies.

As a consequence of the singularity of this work, the treatment of the problem will cover different study fields, using a combination of different approaches inside the large NLP subject.

Chapter 3

Methodology

3.1 Data Model

The data subject of the studies are real, stored inside Anatel databases. The data are stored in text fields in a relational database. Once the data represent free text expressions from customers and companies, there is a lot of sensitive information inside it. Therefore, we need to obtain permission by Anatel to respect the Brazilian General Data Protection Regulation, adopting in special the following practices:

- a) maintaining data processing inside Anatel environment;
- b) guarantee anonymization of the data;
- c) use for the specific research proposal;
- d) store the database operations used in the research.

In terms of volume, "Anatel Consumidor" receives an average of 3.425.210 requests per year, which represents more than 9384 per day (de Telecomunicações, 2021).

As will be better exposed later, after analyzing manually a random sample of 400 requests, Anatel specialists figured out that 68% of cases involve some complaint related to refund, either because the customer has asked for it or because the company has refunded.

3.1.1 Untagged and Subjective Real Data

We observe that, besides being real and sensitive, the data provided to this work is subjective and untagged. Untagged because no one (a person or a machine) has tagged this data before, and subjective because the interpretation to point if a complaint include or not asking for a refund can change from a person to another.

Once we adopt supervised learning to train the models, we need an effort of interactions with Anatel specialists, in order to define objective criteria to tag the complaints and answers and to construct a tagged test dataset with statistical significance to the study.

In this context, we had 6 meetings and more than 10 interactions (by e-mail and chats) with the Anatel specialists in order to define the criteria and tagged the test dataset.

We observe that the analysis of each case (customer complaint and company answer) has a mean human cost of 5 minutes. This measure represents an estimate of effort at least 2000

minutes to Anatel specialists tag the 400 test cases.

3.2 About Software

To choose the languages and tools for the work we observe the Anatel technological environment, in order to follow the organization patterns, once all the tasks will be processed inside the institution, and also because the activities resulted from this research will be continued by the organization after finishing in the academic environment.

With this in mind, we adopt *Python* as computational language, and *SQL Server* as Data Base Management System. As discussed before, once is relevant to observe the idiom to the NLP tasks, we also adopt just NLP tool kits that are implemented in Python and also support Portuguese languages and NER tasks: Spacy, NLTK, NLPyPort (Ferreira, 2019) and Stanza (Qi et al., 2020).

3.3 Software Methodologies

Software development methodologies represent a big topic of studies in the software engineering branch. On the one hand, we identify classical and robust process methodologies, like Rational Unified Process (RUP), that are references not just in literature but also in big projects that focus on documentation and formal requirements. On the other hand, we identify very slim and agile methodologies, like Extreme Programming (XP), that focus on fast adapting, fast development and scenarios with undefined requirements.

We observe also that usually some elements of each methodology are combined in order to adapt parts of them to a concrete project (Pereira & de Brito Vieira, 2006).

Despite the existence of several methodologies focused on software development, we decide to adopt a methodology specialized for data mining projects, as we present follow.

3.4 Adoption of a specific Data Mining project framework

Data Mining projects involve several particularities that distinguish them from industrial or software development projects. Because of this, it was developed and has been usually adopted specific methodologies for this kind of project. The most usual methodology adopted has been the CRISP-DM (CRoss Industry Standard Process for Data Mining) (Wirth & Hipp, 2000).

CRISP-DM previews several levels following a hierarchical model, departing from *phases*, that generate *generic tasks*, that will be described in the *specialized tasks*, that in its turn will be implemented by the *process instances*.

Still, CRISP-DM describes that a data mining project has a cyclic nature, divided into six phases: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modelling*, *Evaluation* and *Deployment*.

In this work, we adopt the strategy to design and implement the solution to the problem based on the CRISP-DM framework. In our case, the first and second phases (Business Understanding and Data Understanding) of CRISP-DM cycle were merged, but the activities continue to exist.

Our strategy consist in execute 4 iterations of 1 month each, in order to run the life cycle 4 times. The figure 3.1 shows the phases.

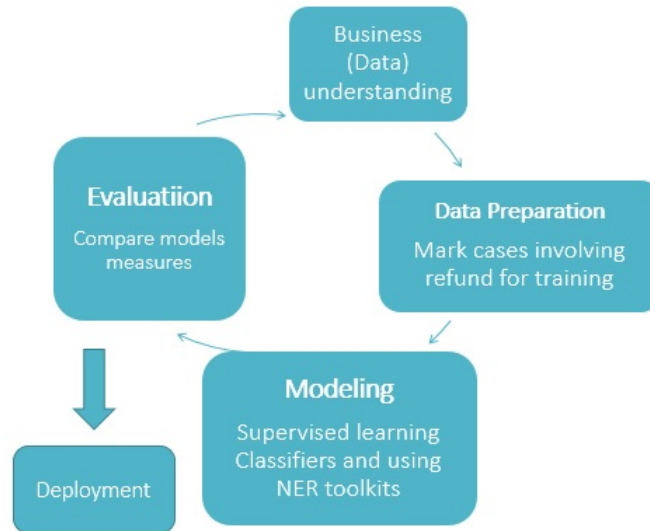


Figure 3.1: Project life cycle based on CRISP-DM.

The first phase, *Business (Data) Understanding*, is dedicated to identify and validate assumptions about the data with the business area. In the second phase, *Data Preparation*, data structures are created containing both metadata and the content of the complaints and company answers. In this phase is also selected a sample and tagged refund cases to be the input to the supervised models. The third phase, *Modeling*, is where we use the NLP tool kits to construct a model to train and test over the sample. In the fourth step, *Evaluation*, we compare the model results and figure out improvement opportunities. Depending on the results obtained, in the *Deployment* phase the model can be prepared to be used by the business area.

In practical terms of iterations, we use the 4 cycles to execute iterations with the following purposes: *1st Iteration*: Focused in creating and configure the development environment in Anatel, and test a simple solution for just 100 or 200 cases from 2020; *2nd Iteration*: Focused in implement improvements in the model and also test a second NLP tool kit, choosing one; *3rd Iteration*: Focused in tagging more data for training and test and check the results; *4th Iteration*: Focused in improvements to optimize the models.

Therefore, the use of this iterative and incremental approach, provided by CRISP-DM, allows each iteration to decrease the risks of the project, compare the results of the classification and NER tasks with the real classification and values of refund, fit the models and improves the final results.

3.5 About Source Code

The source code of this work is available at the following link:

https://github.com/leandrolirabsb/refund_identification_and_quantification.git

The project contains Python code in Jupyter Notebook format. Once the data is sensitive, we removed from the source code the steps related to loading data from the Anatel database, in order to preserve the names of database columns, tables and instances.

Despite this, we provide also the embedding models and embedding features that are the input to the classifiers, as well the test labels. Hence, it is possible to perform some experiments even without textual and personal data.

Chapter 4

Case Study

In this chapter, we present the way we use the methodology and tools to treat the problem.

4.1 The Workflow

The organization and development of an NLP project can follow the phases presented on Cross-Industry Standard Process for Data Mining(CRISP-DM), as shows the figure 4.1.

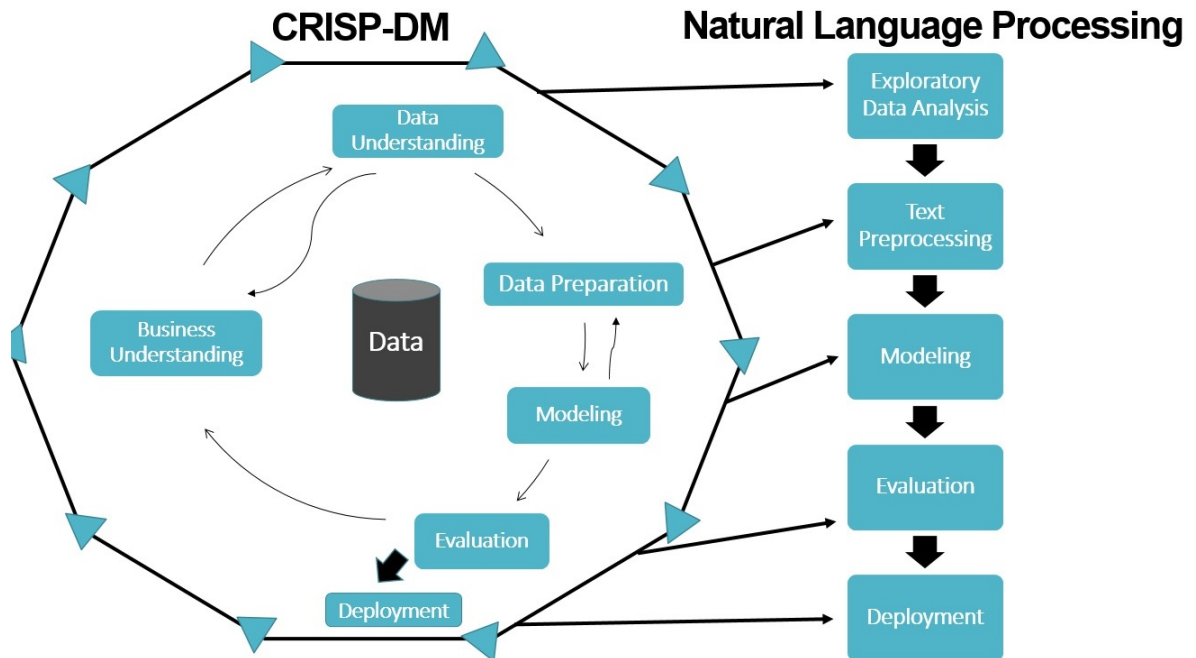


Figure 4.1: NLP Workflow for the project

In an iterative and incremental approach, the development of each of these steps uses as entry the findings and outputs from previous steps. In this context, Exploratory Data Analysis in the NLP process represents the *Business and Data Understanding* steps in CRISP-DM. The Text

Preprocessing represents the *Data Preparation* step in CRISP-DM, the Modeling/Pattern Recognition in NLP represents the *Modeling* step, the Evaluation/Model Selection in NLP represents the *Evaluation* step in CRISP-DM, and the Deployment in NLP represents the same *Deployment* step in CRISP-DM.

We adopt this methodology in this work, and we present in the next sections the tasks we perform in each step.

4.2 Exploratory Data Analysis

This step contains the analysis of the data, to identify particularities that will define actions in the next steps.

In order to do this, we tagged initially 1200 complaints and answers. This classification happened after several meetings with Anatel Business Area to define the classification criteria.

Regarding the problem definition, the Anatel Business area initially address us two questions : "How much is requested by consumers complaining in "Anatel Consumidor"?" and "How much the company says it refund?". Based on this we structure the problem into four questions:

- a) Has the customer asked for a refund?
- b) How much has the customer asked for as refund?
- c) Has the company refunded the customer?
- d) How much has the company refunded?

In this context, it is important to highlight that questions "b" and "d" are dependent on the questions "a" and "c", because does not make sense to search for monetary values in cases where the customer has not asked for a refund and in cases where the company has not pointed a refund.

As a consequence of this problem structure, we define a sheet using both data collected from the database and fields that will represent the target values (the answers to the four problem questions). We use this sheet to tag manually 1200 complaints and company answers ourselves.

Figure 4.2 shows the fields of the sheet.

CLASSIFICATION SHEET SHAPE						
Id	Refund Asked ?	Complaint	Value Complained	Company answer	Company Refunded ?	Value Refunded
	1- Yes 0- No		value (Ex 20,00) or 0- Not possible to identify		1- Yes 0- No	value (Ex 20,00) or 0- Not possible to identify

Figure 4.2: Classification Sheet Format

In this table, the white columns represent data there are selected from the database. In case, the ID of the complaint, the text of the complaint itself and the answer to the complaint, given by the telecommunication company.

On the other hand, the grey columns represent the four fields that are filled by the human specialists, the target values.

Based on the table, in a first analysis, we identify that the solution of the problem involves tasks of 2 different natures: *Classification and Regression*. First, the fields 'Refund Asked ?' and 'Company Refunded ?' represent classification problems, where the task is to identify in which class (YES or NO) the text is. On the other hand, the fields 'Value Complained' and 'Value Refunded' represent regression problems, where the task is to identify a value from the text. It is also important to point that there are interdependent tasks: the regression task that involves finding out the 'value complained' just makes sense in case the customer has asked for a refund. In the same way, the regression task of finding out the 'value refunded' just makes sense in case the company had the decision to refund the customer.

As part of the Business Understanding CRISP-DM step, the Anatel business area presents the following assumptions:

a) Any monetary complaint must be considered. In this context, if the customer complained against mistakes in the bill, for instance, it will be considered. The idea behind this interpretation is that, if the customer would not have complained, he would lose money.

b) The company's answers must be analyzed even if the customers have not asked for a refund. In this context, in some cases, the company refunds even without an asking from the customers.

c) The values complained or refunded, in some situations must be obtained by approximation. For instance, in case the customer asks for 'R\$ 100,00 plus interest', so we consider R\$ 100,00

d) Once the predictions will be used to support public policies decisions, the classification predictive models must avoid false-positive predictions, and the regression models must also avoid making predictions in cases the value is not clear. This behavior permits take decisions supported by statements like 'at least in 60% of complaints the customers ask for some refund' and 'at least R\$ 100.000,00 were refunded to the customers in April 2020'.

4.2.1 Defining the sample size to validation

Despite the classification we did, it was necessary to define the sample to validate the solution implemented. In order to do this, Anatel specialists that work in the data curator area for customer complaints classified another sample.

Regarding the sample size for validation, as we define in the Introduction chapter, the annual average of complaints in the period 2015-2020 is 3.425.210 (approximately 3 million and a half). (de Telecomunicações, 2021)

Once Anatel business area defines that the universe of this study would be the set of complaints over the year of 2020, in case 3.007.538, we adopt this number as the population size.

(Agranonik & Hirakata, 2011) highlight that when the population size is known, it is possible to use the formula in figure 4.3 to calculate the minimum sample size in order to have statistical significance in the study.

Once the proportion expected is unknown, the literature recommends applying the value $p = 0.5$ (50%). (Agranonik & Hirakata, 2011) Considering as population size the value 3.007.538, and applying a confidence level of 95%, with 5% of margin of error, the minimum sample size to support the statistical significance of the study must be 385 complaints.

$$n = \frac{p(1-p)Z^2N}{\varepsilon^2(N-1) + Z^2p(1-p)}$$

where:

- n : sample size;
- p : proportion expected;
- Z : Value of normal distribution given a confidence level
- N : population size
- ε : confidence interval size (margin of error)

Figure 4.3: Ideal Sample Size Formula

Therefore, in order to obtain results with statistical significance, the Anatel specialists classified 400 complaints and respective companies answers. We allocate these 400 cases to be the test sample of the work, and we choose the occurrences randomly from complaints over the year 2020.

4.2.2 Preliminary Results

As a result of the effort in the manual classification of complaints, we have two samples:

- The first, containing 1200 customer complaints and company answers selected from the dates 11/05/2020 and 12/05/2020, classified by ourselves; and
- The second, containing 400 customer complaints and company answers selected randomly from the year 2020, classified by the Anatel business area.

Based on the analysis of these two tagged samples, seen in 4.2.2 we can observe some preliminary results:

a) We did not find any monetary value in complaints or answers wrote in words, like "2 hundred fifty reais". Because of this, we ignore the treatment of situations like that in this work.

b) Analyzing the complaints from the customer point of view in table 4.2.2, we identify that, for the sample size of 1200, in 48,33% of complaints the customers ask for a refund. Also, we can identify the value they are asking for in 28% of cases (or 13% if we consider all the complaints). For the sample size of 400, in 58,50% of complaints the customers ask for a refund, and we can identify the value they are asking for in 25% of cases (or 15% if we consider all the complaints)

Table 4.2.2: Customer Complaints Statistics

1200 Classified by ourselves				
Ask For Refund ?	Quantity	% Ask For Re- fund	It is possible to identify the value	% is possible to identify the value
Yes	580	48.33% of total	160	28% over cases that ask for refund, and 13% over all complaints
No	620			
Total	1200			
400 Classified by Anatel Specialists				
Ask For Refund ?	Quantity	% Ask For Re- fund	It is possible to identify the value	% is possible to identify the value
Yes	234	58.50% of total	58	25% over cases that ask for refund, and 15% over all complaints
No	166			
Total	400			

c) Analyzing the answers for complaints, (table 4.2.3), we verify, in the sample of size 1200, that in 47% of cases the companies refund some value to customers, and in 64% of these cases we can identify the value that is refunded. For the sample size of 400 in 46% of cases the companies refund some value to customers, and in 62% of these cases we can identify the value that is refunded.

Table 4.2.3: Companies Answers Statistics

1200 Classified by ourselves				
Has the company refunded?	Quantity	Percentage	It is possible to identify the value	% is possible to identify the value
Yes	566	47% of cases the company refunded	363	64% of cases where company refunds it is possible to identify the value
No	634			
Total	1200			
400 Classified by Anatel Specialists				
Has the company refunded?	Quantity	Percentage	It is possible to identify the value	% is possible to identify the value
Yes	184	46% of cases the company refunded	114	62% of cases where company refunds it is possible to identify the value
No	216			
Total	400			

This analysis of the data able us to identify the following particularities:

1. For the sample size of 1200:

- (A) In 14% of cases the customer asks for a refund but does not receive any refund;
- (B) In 12% of cases the customer does not ask for a refund but receives some refund from the companies;
- (C) In 35% of cases the customer asks for a refund and receives a refund;
- (D) In 39% of cases the customer does not ask for a refund and also does not receive a refund from the company;
- In 61% of cases the refund subject is present (situations A + B + C), either the consumer requesting a refund or the company granting it.

2. For the sample size of 400:

- (A) In 22% of cases the customer asks for a refund but does not receive any refund;
- (B) In 9% of cases the customer does not ask for a refund but receive some refund from the companies;
- (C) In 37% of cases the customer asks for a refund and receive a refund;

- (D) In 32% of cases the customer does not ask for a refund and also does not receive a refund from the companies;
- In 68% of cases the refund subject is present (situations A + B + C), either the consumer requesting a refund or the company granting it.

In general, it is possible to identify the similarity in the statistics obtained over the sample of size 1200 and the sample of size 400. The small inequalities can be justified by two reasons: a) The second sample was classified by the Anatel specialists, while the first one was classified by myself, despite observing the conventions constructed with Anatel business area; b) the first one sample was collected just from 2 dates, 11/05/2020 and 12/05/2020, while the second sample was collected randomly over all the year of 2020.

	Sample size of 1200	Sample size of 400
(A) Customer ASKS but DO NOT receives refund	14%	22%
(B) Customer DOES NOT ASK but RECEIVES refund	12%	9%
(C) Customer ASKS and RECEIVES refund	35%	37%
(D) Customer NEITHER ASK NEITHER RECEIVE refund	39%	32%
% cases involves some situation of refund (A+B+C)	61%	68%

During the exploratory data analysis, we identify words and expressions that point to an asking for a refund in the customer complaints. On the other side, we identify also words and expressions that point to a refund in the companies answers. These *special words* are useful in order to construct heuristics to help the predictions, what we expose in *Modeling* step.

Regarding the monetary values complained by customers and refunded by the companies, based on the sample size of 400 cases (classified by Anatel specialists and selected randomly over the year 2020) we separate the cases where it is possible to identify the values complained (by the customers) or refunded (by the companies), in order to make some analysis.

In histogram present in figure 4.4 we identify a concentration of the value complained between R\$5,00 and R\$415,00

The boxplot in figure 4.5 shows the presence of several outliers in this type of data. Because of the constant presence of outliers, it is recommended to not consider the mean of the values in the analysis, but the median. In case, the median value asked for a refund by the customers presented in the sample of 400 complaints is R\$ 67,50 .

Regarding the values refunded by the companies, the histogram in figure 4.6 shows the concentration of values refunded between R\$4,00 and R\$154,00.

As happened when analyzing the values complained, the values refunded by the companies present several outliers, as we identify on the boxplot in 4.7. Taking into count the median, the reference value founded in refunds was R\$89,95. Remembering from table 4.2.2 that in 46%

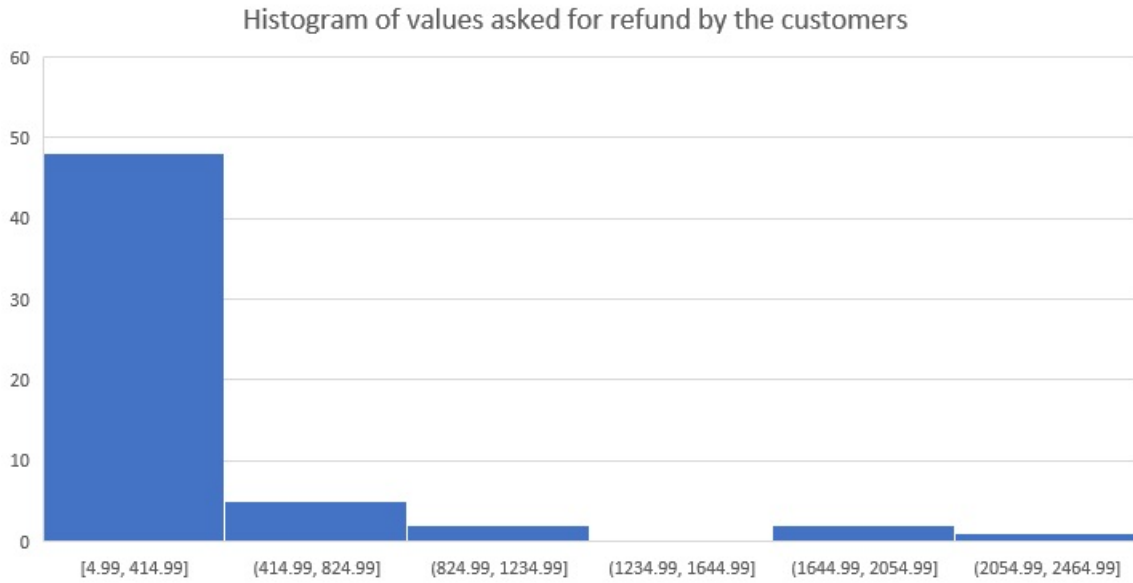


Figure 4.4: Histogram of values complained by customers

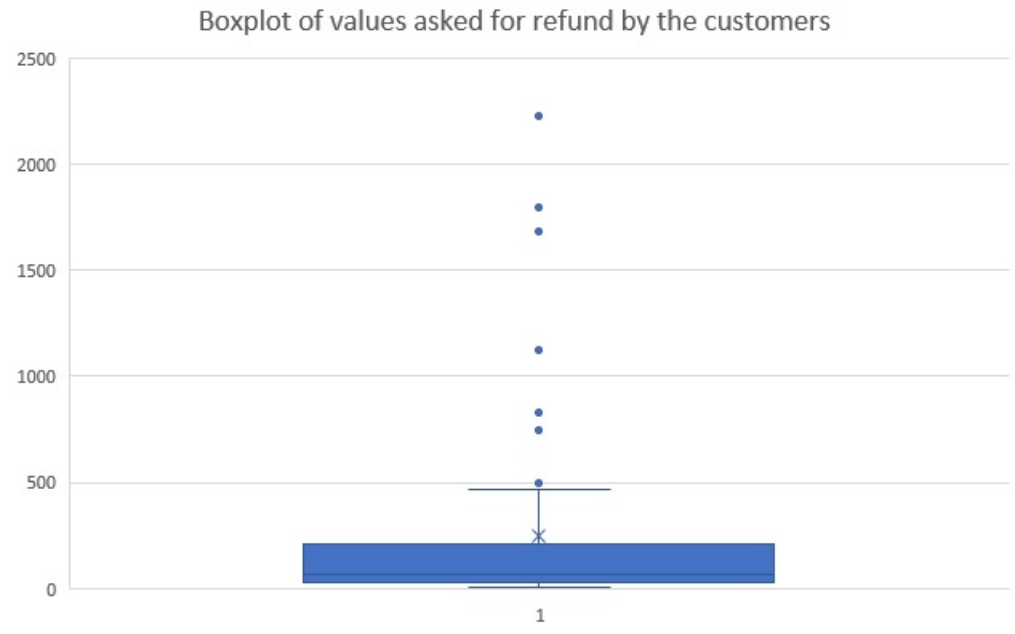


Figure 4.5: Boxplot of values complained by customers

of all cases the companies are refunding customers, and considering the average of 3.425.210 complaints per year (table 1.1), we can estimate that in a year there are refunds in approximately 1.575.596,6 cases. Taking the median refunded value as a reference, **the "Anatel Consumidor" platform supports an annual amount of R\$ 141.724.914** (one hundred and forty-one million, seven hundred and twenty-four thousand, nine hundred and fourteen *reais*) in refunds

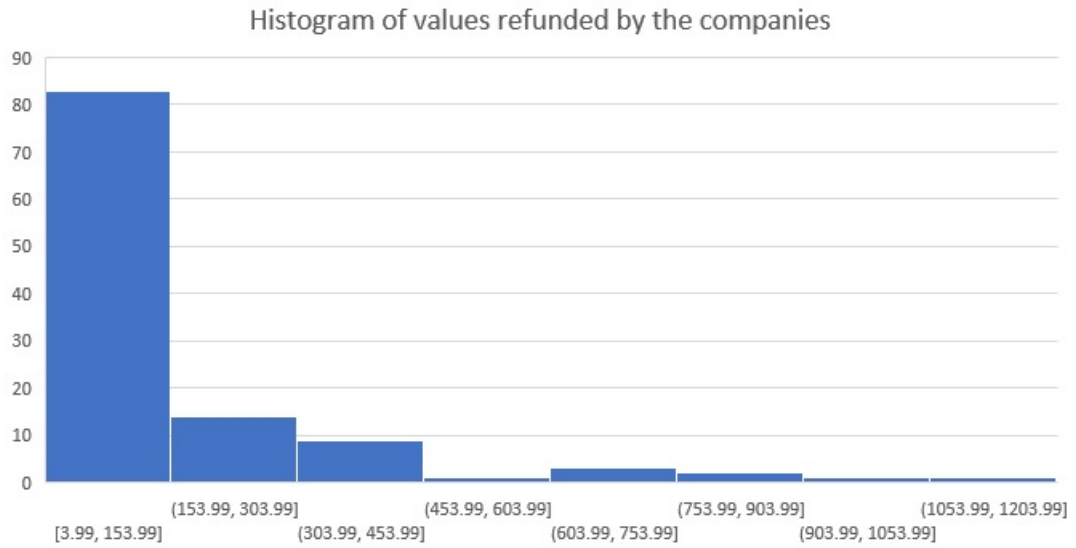


Figure 4.6: Histogram of values refunded by the companies

for telecommunication customers.

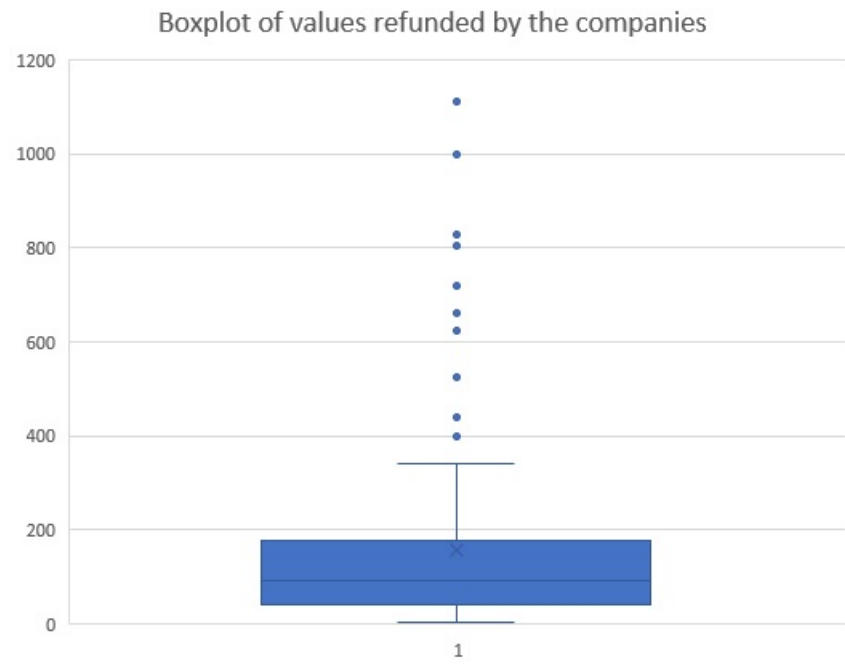


Figure 4.7: Boxplot of values refunded by the companies

Despite the high level of estimation of these calculations over the values complained and the values refunded, they represent relevant findings from this work to Anatel, once they came from a sample collected with statistical significance, and can be an instrument to quantify how much the public institution generates in benefits to society.

Also, the preliminary results represent an important instrument to Anatel understand better the behavior patterns from customers and companies.

4.3 Preprocessing text

Preprocessing in text mining is the step that involves normalizing, cleaning and preparation of the text to be used by the algorithms. In the context of the present work, there are 2 columns containing texts in natural language: the column that contains the customer complaint and the column that contains the company answer. Therefore, the preprocessing step consists of generate 2 new columns containing the result of the below treatments over the original columns.

Figure 4.8 shows the steps present in data preparation:

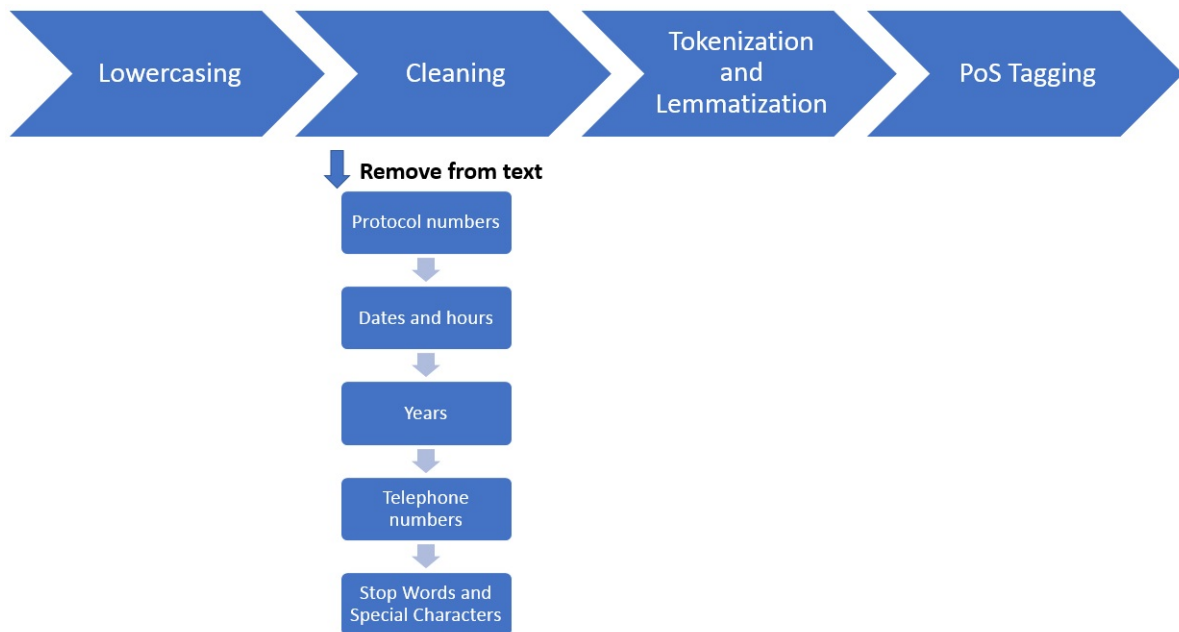


Figure 4.8: Preprocessing Text Tasks

1. *Converting all text to lowercase:* this is a kind of normalization in text, in order to maintain all text in conditions to be compared with other words and avoid mistakes in comparisons.
2. *Text Cleaning:* this task consists of several sub-tasks. The first subtask uses regular expressions to remove text in a pattern that will not be useful to solve the problem. In this context, we remove any word with any of the following patterns:
 - (a) containing 15 numbers, which represents the complaint protocol number;
 - (b) dates in formats 'dd/mm/yyyy' (like 11/07/2020) and 'dd/mm' (like 11/07);
 - (c) hours in formats 'HH:MM' (like 14:35), 'H:MM' (like 09:34), 'HHhMM' (like 14h35) and 'HhMM' (like 9h34);

- (d) telephone number area code, in format '(dd)', like '(11)';
- (e) telephone numbers including area code (cc) in formats 'cc dddddddd' (like 21 985883880), 'cc dddddddd' (like 21 985883880), 'ccdddd-dd' (like 2197033-4392), 'ccdddd-dddd' (like 213321-4392), 'ccdddd dddd' (like 2197033 4392), 'ccdddd dddd' (like 213321 4392), 'cc ddddd-dddd' (like 31 98131-3131), 'cc dddd-dddd' (like 31 3321-3131), 'cc ddddd dddd' like (31 98131 3131), 'cc dddd dddd' (like 31 3321 3131), teleohone numbers with 8-11 digits format (like 11982165069), 'dddd-dddd' (like 99991-0621), format 'dddd-dddd' (like 9991-0621), format 'dddd dddd' (like 9991 0621), format 'dddd dddd'(like 9991 0621);
- (f) the numbers '2019', '2020' and '2021' that appear isolated, because they probably represent a year;

We highlight that removing numbers representing protocols, telephones and dates has a relevant contribution to the steps of detecting the value complained by the customers and the value refunded by the companies, once with this cleaning in the text the numbers that will remain will probably be related with the values complained or refunded.

- (g) Another important cleaning subtask is related to the group of *stop words* and some *special characters*. Stop words are words that are very frequent in a language or in the specific text. Because of its frequency, its presence does more harm than help in analyzing and solving the problem. Therefore, we gain in removing these words from the text. In this context, we define a group of stop words containing the *nltk toolkit stop words set in portuguese language* plus the following terms that are very frequent in complaints but not relevant, like *companies names, conjunctions, months* etc.

The list of stop words is the follow: 'protocolo', 'reclamada', 'prestadora', 'operadora', 'operador', 'prestador', 'empresa', 'oi', 'vivo', 'tim', 'claro', 'embratel', 'sky', 'algar', 'telefonica', 'telefônica', 'net', 'nextel', 'dia', 'dias', 'data', 'ano', 'anos', 'após', 'fez', 'referente', 'consumidor', 'consumidora', 'para', 'pois', 'porém', 'anatel', 'reclama', 'reclamação', 'reclamar', 'providência', 'pede', 'telefônico', 'providências', 'janeiro', 'fevereiro', 'março', 'abril', 'maio', 'junho', 'julho', 'agosto', 'setembro', 'outubro', 'novembro', 'dezembro', 'providencia', 'providencias', 'telefone', 'contrato', 'contratar', 'ser', 'fiz', 'ter', 'sendo', 'diz', 'disse', 'nada', 'nome', 'vir', 'www', 'numero', 'número', 'nr', 'nº', 'vc', 'vcs', 'pq', 'tb', 'ai', 'aí', '(a)', '(es)', '!', '??', 'mês', '(?)', '*', '-', '""', '""', '@', '&', ';', '#'.

We also notice that numbers in general and some punctuation and special characters (',', '.', '\$') can not be removed because they are the base for the Brazilian monetary values, that use '.' as separator of thousands and ',' as separator of decimals, like R\$ 1.215,60 (one thousand two hundred and fifteen reais and sixty cents).

Especially the cleaning task produces a very important result to the next steps in the workflow, once removes from text words that would just make the interpretation more confused. The example in table 4.1 shows the difference in a complaint after the cleaning step, where we identify how the removing of numbers (year, protocol

number and telephone number) results in a text that highlights the monetary value complained.

Table 4.1: Result of Lowercasing and Cleaning subtasks over a complaint

Idiom	Original Customer Complaint
English Version	On April 18, 2020, I opened a protocol with TIM to verify the "disappearance" of approximately R\$7,00 of credit that disappeared from my telephone line (62) 98215-****. In my consumption history on the MEUTIM website there is no date, time or reason for the undue debit. In conversation with the attendant, he was also unable to explain to me the reason for the disappearance of my account balance and opened protocol number 202040396****. Days later, the protocol was concluded without the company making the reimbursement of the amount or even giving me a opinion about what happened. I request the return of my credit or at least an explanation of the fact..
English Version	<p>After Lowercasing and Cleaning</p> <hr/> 18, carry out opening together verify disappearance "about R\$ 7,00 real credit disappear line. History consumption site meutim appear, hour reason for undue debit. Conversation attendant, learn to explain reason disappearance account balance carry out opening. Give finalize company make reimbursement value give opinion respect occurred. order, credit return, minus, explanation of fact.

3. *Tokenization and Lemmatization:* Tokenization is one of the most commons preprocessing subtasks in text mining and consists in split the text into minimum parts, called tokens. In general, each word of the text is a token.

On the other side, Lemmatization is the process of finding the lemma of the word. Lemma represents the entry of a word in the dictionary, and is the origin of any other word. It is important to note that, different of tokenization, the lemmatization is totally dependent on the language. During this work, the tool we adopt to perform this subtask is the *stanza* toolkit.

Therefore, departing from two columns containing original text with the customer complaint and the answers by the companies, we create two new columns containing the cleaned preprocessed text.

4. *PoS Tagging:* Part of Speech Tagging is a process that classifies each word observing the grammatical function it represents in the context of the text. For instance, in certain cases, a word can be a noun, and in others can be a verb. This process is very useful in the second part of the problem, the regression/NER tasks, because with this technique we can identify more easily monetary values in the text.

We must highlight the need to have a specific tagger for the idiom. Because of this, we adopt the *Mac-Morpho Portuguese Corpus* to train the PoS-tagging classifier. Below we present the tagging of a sentence in Portuguese idiom, with the nouns(N), adverbs(ADV), adjectives(ADJ), verbs(V) and articles (A).

Original Sentence: "Vasco da Gama é o maior português de sempre!"

PoS Tagged Sentence: [('vasco', 'N'), ('da', 'N'), ('gama', 'N'), ('é', 'V'), ('o', 'ART'), ('maior', 'ADJ'), ('português', 'ADJ'), ('de', 'PREP'), ('sempre', 'ADV'), ('!', '!')]

The PoS Tagging process involves the concept of n-grams. The "n" represents how many words around the one is being tagging will be observed to define the grammatical function of it. This work adopts 3-grams to train the tagger and 10-grams to the tasks involving detecting the values complained and refunded, which we detail in the modeling section.

For instance, below we exemplify the application of a 3-gram task over a sentence:

"The company has R\$ 50,00 in debit with me!"

('The', 'company', 'has')

('company', 'has', 'R\$')

('has', 'R\$', '50,00')

('R\$', '50,00', 'in')

('50,00', 'in', 'debit')

('in', 'debit', 'with')

('debit', 'with', 'me!')

4.4 Modeling

4.4.1 Structure of the problem

As discussed before, the problem involves two classification tasks and two regression/NER tasks:

a) *Classification Tasks:* to identify if the customer asked for a refund and identifies if the company refunded some money;


b) *Regression/NER Tasks:* to identify the value complained by the customer and identify the value refunded by the company. This problem type has a regression nature because the target is a continuous value. On the other hand, this problem has also a NER nature because techniques of Named Entity Recognition are appropriate to solve it.

Figure 4.9 shows the nature of these 4(four) target values:


Based on figure 4.9, we split the main problem in four, and label the names these problems will be referenced in this work:

- *Task 1:* To identify if the customer complaint contains an asking for refund;
- *Task 2:* Derived from the response to Task 1, to identify how much the customer is asking for as refund;
- *Task 3:* To Identify if the company answer describes a refund for the customer;


CLASSIFICATION SHEET SHAPE						
Id	Refund Asked ?	Complaint	Value Complained	Company answer	Company Refunded ?	Value Refunded
	1- Yes 0- No		value (Ex 20,00) or 0- Not possible to identify		1- Yes 0- No	value (Ex 20,00) or 0- Not possible to identify




Classification
Task n° 1



Regression/NER
Task n° 2



Classification
Task n° 3



Regression/NER
Task n° 4

Figure 4.9: Prediction Task Types

- *Task 4:* Derived from the answer to Task 3, identify how much the company refunded;

Before discussing the models constructed, we highlight some assumptions about the problem nature that are determinant for the choosing of the most appropriate model:

- *Focus on the Precision of the classification models:* Once the results of the predictions have the potential to be used as entries in concrete regulatory actions, *we must avoid false positive* predictions in the classification tasks. Therefore, it is better to have a model with lower general accuracy, but minimum false positives, than a model with high accuracy but several false positives predictions.
- *Focus on the Deviation of the regression and Precision of the NER model:* For similar reasons, the predictions of the asked values for a refund by the customers and refunded values by the companies have the potential to be used in regulatory actions. In this context, we must minimize the errors. In other words, it is preferred a model that avoids predicting the value in dubious circumstances than a model that predicts the values in most situations but with more errors. For the situation the model decides to predict the value, we must evaluate also the less deviation in the prediction task.

Especially for the classification tasks, another important motivation to minimize the false positive predictions is to avoid starting the regression tasks in cases where there is no value asked for a refund or refunded. For instance, if the model has wrong predicted that a company refunded the customer, the regression task, in this case, will start and search for a value in the text that does not represent a value refunded.

The respect for these assumptions allows Anatel to have the following answers with reliability during the predictions:

- *At least* in % of cases, the customer is asking for a refund;
- *At least* in % of cases, the company is refunding;
- The sum of value refunded has *at least* the amount X.

4.4.2 Train and Test datasets

Once the strategy for the classification tasks involves supervised learning, we need to define the datasets for training and test and split according to some ratio.

The definition of the best ratio for this split represents an own study field in literature, that points to different approaches depending on the sample size, the sample cost and the problem nature. (Pawluszek-Filipiak & Borkowski, 2020) and (Rácz, Bajusz, & Héberger, 2021).

However, normally we consider a good start point for general studies ratios between 70%-80% for train and 20%-30% for the test . In this context, we adopt as the training dataset the sample tagged by ourselves, with 1200 complaints and 1200 answers; and as the test dataset, the sample tagged by Anatel specialists, with 400 complaints and 400 companies answers. This configuration represents a split containing 75% of sample for train and 25% for the test.

Another motivation to adopt this division is that the sample tagged by Anatel specialists represents the official interpretation of the institution over the cases. Hence, is appropriate to allocate this sample to test the solution.

4.4.3 Modeling the Classification Tasks

The strategy to treat *Task 1* and *Task 3* involved the adoption of supervised learning, with is an approach where someone that understands the problem domain informs the real target values of the dataset. These values are the labels that are used to train the models and evaluate the algorithms in the prediction activities.

We must remember that, by the problem nature and Anatel point of view, it is important to minimize situations where there is uncertainty in predictions for both classifications and regression tasks. In other words, once the results of the predictions can be used to adopt actions over companies, we must to avoid false-positive situations in the the tasks.

Another relevant aspect is the use of some metadata derived from the complaint form (Figure 1.1 in Introduction chapter). Once the user select the complain type during the input data, this information is used as a feature in the classification tasks.

First wave implementation: own heuristic

Derived from the reading of 1200 customer complaints sample, we identify some words frequently present when the customer is asking for a refund. On the other hand, we figured out words frequently present in situations where the companies are refunding. These words are present, in a lemmatized form, in table 4.4.3. These heuristics did not use any metadata from the problem.

Table 4.4.2: Frequently words in refund cases

Customer Complaints	English version	Companies Answers	English version
ressarcimento, ressarcidar, ressarcir, ressarçar, restituição, restituicao, restituir, reembolso, reembolsar, indenização, indenizar, indenizacao, estorno, estornar, extorno, extornar, extorne, estorne, devolução, devolucao, devolver, crédito, credito, creditar, compensação, compensar, reparação, reparar, reaver, multa, fidelidade, valor, absurdo, desconto, redução, roubo, preço, negociar, debito, juro, caro, contar	refund, refund, refund, refund, refund, refund, refund, indemnification, indemnify, indemnification, refund, refund, refund, refund, refund, refund, credit, credit, credit, compensation, compensate, repair, repair, recover, fine, loyalty, value, absurd, discount, reduction, theft, price, negotiate, debit, interest, expensive, count	lançar, abatimento, isento, isenção, conceder, baixo, acordo, desconto, credito, renovação, ajuste, contestar, lançar, ressarcir, corrigir, diferença, readequação, desconsiderar, concessão, resolver, baixo, cancelamento, providenciar	post, rebate, exempt, exemption, grant, down, agreement, discount, credit, renewal, adjustment, dispute, post, refund, correct, difference, repair, disregard, concession, resolve, down, cancellation, provide

As a consequence of this finding, we develop a heuristic based on words that point to an asking for a refund (by the customer point of view), and words that point to a refund (in the company answers). This heuristic was useful not just to promote features to the classification algorithms, but also to search for complaints in order to enlarge the training dataset. The results are present in the next chapter.

Second wave implementation: Word Embeddings

In the second wave implementation for classification tasks, we adopt *word embeddings*, most specifically *Word2Vec*.

Initially, in order to train the embedding model we use a corpus containing 100,000 (a hundred thousand) customer complaints and 100,000 companies answers. This step is important to prepare the Word2Vec model to identify relationships between words, their meanings and uses in similar contexts with other words. It is important to highlight that we develop this step after the step of preprocessing the text. Therefore, the corpus we use to train the embedding model contains already cleaned data.

In the next step of this task, we submit the blocks of cleaned text to the embedding model in order to create vectors with 300 dimensions that represent the customer complaint or company the answer. The number of 300 dimensions represents a pattern in Word2Vec usual problems.

After, for task 1, we submit as features the 300 dimensions plus the classification code of the complaint (selected by the customer) to different classification algorithms (Decision tree, Gradient Boost etc.), in order to train models to predict if the complaints contain or not an asking for a refund.

Third wave implementation: Heuristics + Word Embeddings

The third wave implementation joins the strategies used in the previous two. Therefore, we use together the 300 features created to the word embedding model, plus the heuristic and metadata, as a proposal to obtains better results.

The figure 4.10 shows the structure of features that are used as input to train the classification algorithms, as well the label tagged manually.

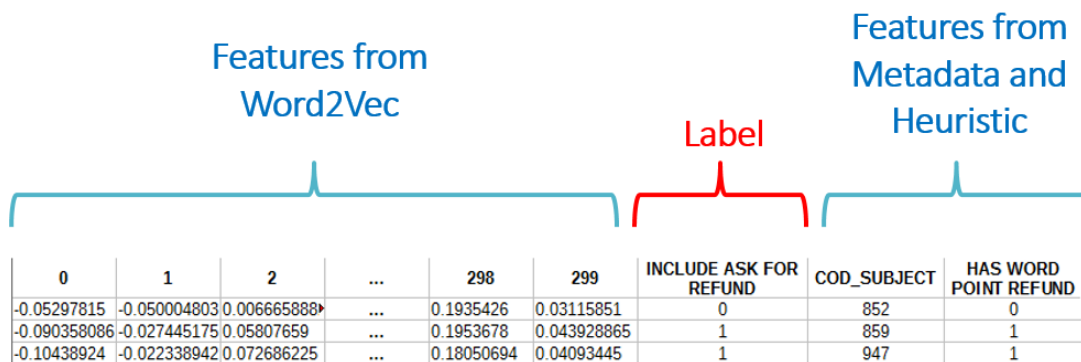


Figure 4.10: Features as inputs to train classification models

4.4.4 Tasks and opportunities to improve the model

The CRISP-DM methodology predicts an incremental and iterative workflow where each iteration shows opportunities to improves the results and to make some adjustments. In this context, from each iteration to the next, various tunings were made, including test different combinations of words to the heuristics, reclassify cases incorrectly classified, uses the prediction of the heuristic as a feature in the others models, change the classifier parameters, and enlarge the corpus to train the models.

Besides these types of improvements, we identify several tasks well addressed in literature in order to improve models already constructed.

Hyperparameter Optimization

The different train algorithms we use in the steps of training the embedding and in the steps of training the classifiers (like Decision Tree and Gradient Boosting) involve several parameters.

During this work, we did manually and just superficially setup of different parameters. However, we underline that optimization can represent a relevant gain in the results. A way to process this tuning automatically is using *Grid Search*, which is a method that makes a complete search over a given subset of the hyperparameters space of the training algorithm (Liashchynskiy & Liashchynskiy, 2019).

Enlarging the corpus

The benefits of use word embeddings are clear in the context of identifying patterns in the text including at the semantic level. However, these gains have a cost: usually, it is necessary to enlarge the corpus to obtain high-quality word embeddings. Roberts (2016)

For example, in this work initially, the embeddings were trained with a corpus of 10.000 customer complaints and 10.000 companies' answers. Once enlarged this corpus to a size 100.000, the gains in the predictions were visible. Despite these gains, the ideal would be to train the embedding with all the corpus of all "Anatel Consumidor". Hence, enlarge this corpus is an opportunity to obtain better results.

Active Learning to enlarge train dataset

Besides the corpus enlargement, enlarge the training dataset also has the potential to generate better results. Once the data was manually tagged, and this human tagging has a cost, an opportunity to enlarge the training dataset would be to use the heuristics in order to search for complaints and answers in the database to compose a larger train dataset, with occurrences tagged by the heuristics.

This idea has support in the literature by the concepts of *active learning*. The principle of active learning is that a learning algorithm can ask an "annotator" to label new instances still during the training step (Settles, 2009). In the context of this work, the annotator could be the heuristics developed.

Review the labels in train and test datasets

Once the definition of the problem involves subjectivity in the interpretation of cases in which customers are asking for a refund and the companies are refunding, we expect to have some mistakes in the human annotations in the train and test dataset. Hence, probably a review of the tagging in both datasets will reduce these errors and promotes gains in the results.

Adoption of more sophisticated representation models

As described in the literature review chapter, BERT (Bidirectional Encoder Representations from Transformers) is considered the State of Art in language representation models nowadays, presenting gains when comparing with Word2Vec models. Therefore, the adoption of BERT has great potential to promote better results.

4.4.5 Modeling the Regression/NER Tasks

As discussed in the literature review chapter, the issue involving identify numbers in a text, discover the meaning of this number in the context and making operations with it is still new a branch of research, especially when applying to non-tagged datasets.

Tasks 2 and 4 of this work involve this complexity. On the one hand, the companies are more explicit when talking about a refund. On the other hand, the customers are much more imprecise to describe the value they are asking for. Remembering the preliminary results, presented in tables 4.2.2, in 58,50% of cases the customers are asking for some refund, but when they do this, we can identify the value in just 25% of cases. On the other hand, from table 4.2.2, in 46% of all cases, the companies refund the customers (even if they did not ask for it), but in just 62% of the refunded cases a human can identify the value.

As evidence of this complexity, in table 4.4.5 we list words or expressions (translated from Portuguese) that point some refund by the companies in the context of the cases studied:

Table 4.4.5: Examples of companies refunds expressions

"we posted R\$15.00 for rebate on next month's bill" / "the termination fine would be exempt, where it will not be charged and we exempt the invoices of R\$69.87 and R\$112.18" / "termination fine will not be generated" / "we grant recharge" / "the invoice was canceled" / "the invoice was written off" / "We clarify that by agreement we apply a discount of BRL 5.00 for 12 months on the claimed line and exempt the customer from the total amount of the invoices" / "the amount of BRL 40.00 of discount was generated" / "on 05/14/2020 the invoices of R\$169.33 due on 04/25/2020 were corrected to the amount of R\$129.71 with maturity extended to 05/28/2020 and the invoice of R \$180.86 maturing on 05/25/2020 was corrected to the amount of R\$91.96 maturing on 05/28/2020" / "discount of 5.00/ the credit in the amount of BRL 56.15 will be generated for the next invoice" / "the exemption from the fine was carried out/ bonus renewal" / "an adjustment in the amount of R\$174.42 was made on the invoice" / "Therefore, the invoice that would be generated for 06/06/2020 was cancelled, that is, the amounts were withdrawn, leaving no future amounts for payment." / "the amount of 44.99 charged on 05/17/2020 was reversed" / "the invoice due 02/20/2020 amounting to BRL 85.99 will be contested" / "value posted as a credit of BRL 143.26 for the next invoice" / "therefore the amount of BRL 12.00 charged for the daily rates were reimbursed" / "we corrected the invoice due 05/25/2020 from R\$145.83 to R\$118.98" / "the discount for the interruption period will be given on the next invoice" / "we provide a balance of R\$ 20.00 in the name of good relationship" / "the adjustment was made in the amount of 292.31" / "we launched the difference of R\$ 66.95 as a discount for the month of July" / "readjustment of values" / "registered a discount of BRL 5.00 for 12 months" / "discount of BRL 34.90 for 12 months" / "values should be disregarded" / "as a concession, the amount of BRL 30.00 was entered" / "demand was resolved" / "cancellation of the fine" / "write-off on the invoice" / "there will be no charge" / "we reached an agreement to keep the amount of BRL 50.00 for the next 5 months" / "we provide the adjustment" / "we provide the change"

As exposed, there are situations where we need to sum values, in other situations subtract values, besides identify which value must be taken in count. Therefore, to the best of my knowl-

edge, this scenario point that this is not a typical regression problem, because it is not the case of having a function that predicts mathematically the target value due to the features.

In this context, we construct the solution inspired by the QSearch prototype (Ho et al., 2020), described in the literature review. Therefore, we develop an heuristic to catch the values asked for a refund and refunded. This heuristic uses concepts of NER, n-grams, and special tokens that point to a refund (in case of customer complaint) or asking for a refund (in case of company answer), as we explain better below:

1. From the output of Task 1, over the Customer Complaints that point an asking for a refund:
 - (a) Using NER, identify monetary values;
 - (b) For each monetary value, search for special lemmas that point an asking for a refund in a distance "r";
 - (c) In case exist some special lemma, include the value in the bag of values asked for refund.
2. From the output of Task 3, over the Companies Answers that point a refund:
 - (a) Using NER, identify monetary values;
 - (b) For each monetary value, search for special lemmas that point a refund in a distance "r";
 - (c) In case exist some special lemma, include the value in the bag of values refunded.

Figure 4.11 shows the steps of application of this heuristic over a simple real example of company answer.

The implementation of this heuristic uses initially the range of 5-grams, and the same sets of words that point to an asking for a refund (for customers complaints) and point a refund (for companies answers) adopted in the classification tasks. During the experiments, we change the range in order to look for better results.

In a second wave implementation, we construct a *second algorithm*, that observes the weight (the semantic) of the values. This allows us to identify if we have to SUM the value, subtract, or use to divide, or to multiply another value.

1. From the output of Task 1, over the Customer Complaints that point an asking for a refund:
 - (a) Using NER, identify monetary values;
 - (b) For each monetary value, search for special lemmas that point an asking for a refund in a distance "r";
 - (c) In case exist some special lemma, search for a neighbor token that point to the weight of the value, and include the result of the value with its weight in the bag of values asked for refund.

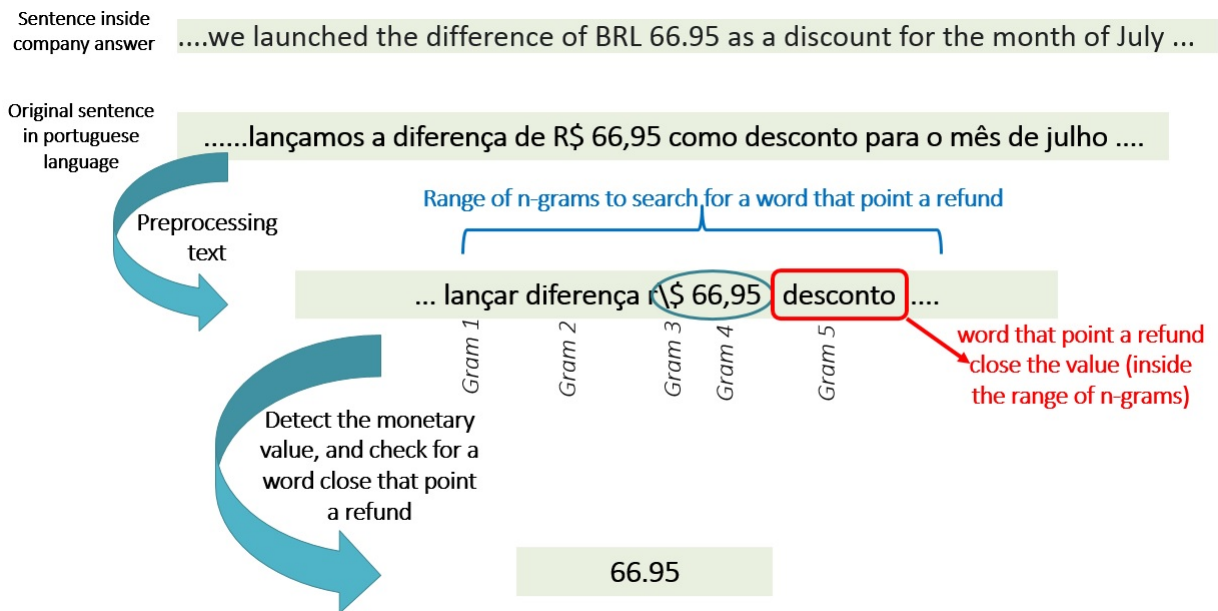


Figure 4.11: Example of the application of the first heuristic to detecting the value refunded

2. From the output of Task 3, over the Companies Answers that point a refund:
 - (a) Using NER, identify monetary values;
 - (b) For each monetary value, search for special lemmas that point a refund in a distance "r";
 - (c) In case exists some special lemma, search for a neighbor token that point to the weight of the value, and include the result of the value with its weight in the bag of values refunded.

Figure 4.12 shows an example of the application of this heuristic.

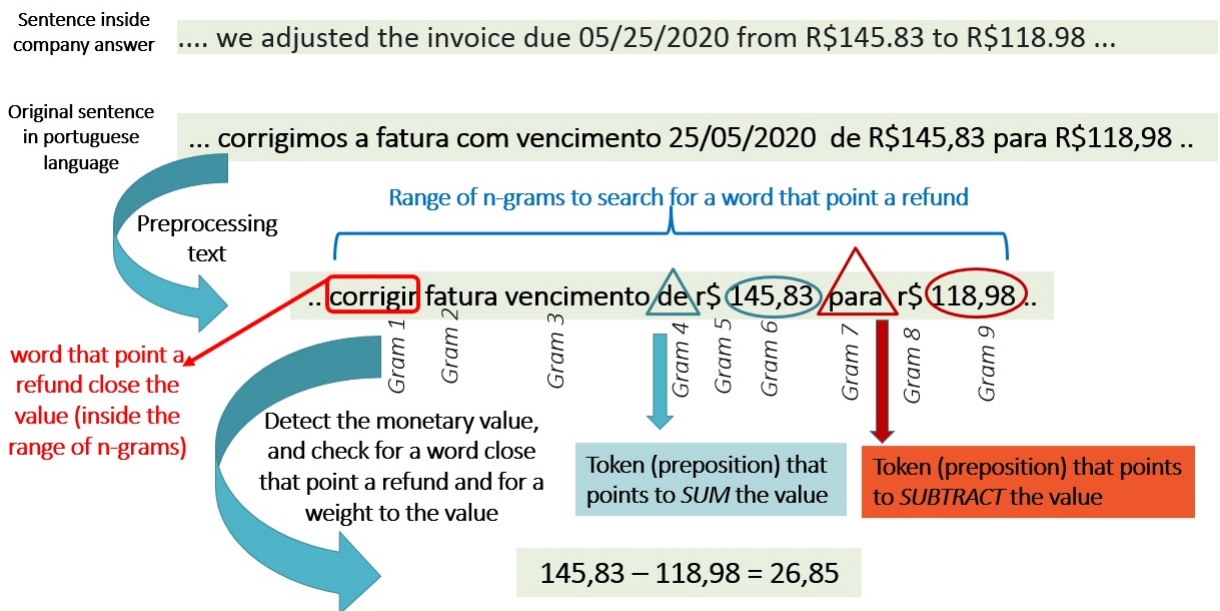


Figure 4.12: Example of application of the second heuristic, with weights to the numbers

Chapter 5

Evaluation and Analysis of Results

In this chapter, we present the criteria of evaluation and the analysis of the scores obtained in different waves of implementation to the four tasks, in order to show the gain or loss in the scores depending on the strategy or algorithm adopted.

Remembering the problem nature and Anatel interest, in the evaluation step, we focus on minimizing false positive predictions in classification tasks and reduce the deviation in regression/NER tasks.

Besides that, in the regression/NER tasks, we have a concern about reducing situations where the algorithm tries to predict values in cases where the real value does not exist in the text.

5.1 Evaluation and Results in Classification Problems - Task 1 and Task 3

5.1.1 Evaluation in Classification Tasks

As discussed in the Literature Review chapter, Precision, Recall and F-Score are the most common measures in classifications tasks. Once the dataset is relatively balanced, and the model must minimize false positive classifications, the measure more appropriate to be observed is the Precision ($\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$). With this measure, the smaller the number of false positives, the greater the precision.

Besides the Precision score, once the corpus is relatively balanced (48% of complaints contains an asking for a refund and 47% companies answers have some refund), the simple Accuracy Score is also a relevant measure. The Accuracy score represents just the percentage of successful classifications in the task.

We present the results for Tasks 1 and Task 3 in three different moments: a) using just a simple heuristic of "special words"; b) using word embedding and the metadata provided by the problem, and c) using word embedding + the heuristic + the metadata.

We remember that *metadata* refers to the complaint subject that the customer selects in the "Anatel Consumidor" form when submitting the complaint (figure 1.1 in the introduction chapter).

Experimental Setup

We perform the experiments related to Task 1 and Task 3 using as train dataset the 1200 cases (complaints and answers) tagged by ourselves, and as test dataset, the 400 cases (complaints and answers) tagged by Anatel specialists. We used the classifier algorithms with standard parameters, and we trained the embedding model (Word2Vec) with a corpus of 100.000 (a hundred thousand) cases of complaints and answers from 2020, with 1001 epochs and creating vectors with 300 dimensions. These values for epochs and dimensions are initial patterns in literature.

Regarding Task 1, we perform the experiment 6 times: once just using the heuristic of "special words", a second adopting Word2Vec + metadata, and 4 more times using Word2Vec + metadata + heuristic for 4 different classifiers.

Regarding Task 3, we adopt the same strategy of Task 1, and perform the experiment 6 times: once just using the heuristic of "special words", a second adopting Word2Vec + metadata, and 4 more times using Word2Vec + metadata + heuristic for 4 different classifiers.

5.1.2 Results using the heuristic

Table 5.1 presents the scores for Task 1 (identify when the customer is asking for a refund) just using the heuristic of "special words". We remember that "special words" are tokens that point to an asking for a refund (in the case of customer complaint) or that point to a refund (in the case of company answer).

Table 5.1: Scores for Task 1 just using the heuristic

Accuracy	0.782
F1	0.802
Recall	0.756
Precision	0.855

Table 5.2 presents the results for task 3 just using the heuristic of "special words". We identify a loss in the scores, comparing with the results obtained for task 1. This can be justified because usually, the texts the companies use to deny a refund are very similar to the expressions they use do agree with a refund, just adding some negative words, like in "*the exemption from the fine is unfounded*". Therefore, in cases where the companies deny the refunds, this simple heuristic tends to generate false positives predictions, which compromise especially the Precision score.

5.1.3 Results using Word2Vec + metadata

In this implementation strategy, we use as inputs for a classifier model the text represented as word embeddings and the metadata provided by the customer complaint form.

Table 5.2: Scores for Task 3 just using the heuristic

Accuracy	0.662
F1	0.705
Recall	0.880
Precision	0.589

Table 5.3 presents the scores after adopting word embeddings plus metadata, case of Word2Vec over Task 1. Here the scores are lower than the obtained with the simple heuristic of "special words". This is an interesting finding, and an evidence that in some cases, for specific domains, simple and *ad hoc* implementations can present better results than more sophisticated approaches.

Table 5.3: Scores for Task 1 using Word2Vec + Metadata

	Ada Boost Classifier
Accuracy	0.700
F1	0.736
Recall	0.717
Precision	0.756

Table 5.4 presents the scores after adopting word embeddings, case of Word2Vec over Task 3. Differently from what we observe in task 1, here the scores are greater than the obtained with the simple heuristic of "special words".

Table 5.4: Scores for Task 3 using Word2Vec + metadata

	Ada Boost Classifier
Accuracy	0.672
F1	0.650
Recall	0.663
Precision	0.638

5.1.4 Results using word2Vec + Heuristic

As part of the third wave of implementation, we combined the concepts of word embeddings with the heuristic of "special words" created.

Table 5.5 presents the scores for Task 1. It is visible a gain in results with the merge of the two models, observing the *Precision* score.

Table 5.5: Scores for Task 1 by classifiers, using Word2Vec + Heuristic + metadata

	Ada Boost	Decision Tree	Gradient Boosting	Random Forest
Accuracy	0.757	0.790	0.787	0.742
F1	0.785	0.808	0.812	0.768
Recall	0.760	0.756	0.786	0.730
Precision	0.812	0.867	0.840	0.810

Table 5.6 contains the scores for Task 3. As happened in Task 1, it is possible to identify the gain obtained with the merge of the two models.

Table 5.6: Scores for Task 3 by classifiers, using Word2Vec + Heuristic + metadata

	Ada Boosting	Decision Tree	Gradient Boosting	Random Forest
Accuracy	0.712	0.657	0.727	0.757
F1	0.679	0.642	0.701	0.731
Recall	0.663	0.668	0.695	0.717
Precision	0.697	0.618	0.707	0.745

5.2 Evaluating and Results in Regression/NER Problems - Task 2 and Task 4

5.2.1 Evaluation in Regression/NER Tasks

As discussed in Literature Review Chapter, some of the most common measures for regression problems are the Mean Squared Error (MSE), the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE), and the Median Absolute Deviation (MAD).

The first one, MSE, computes the average of the squared difference between the real value and the value predicted. The RMSE computes the square root of the average square difference between the real value and the value that was predicted. The MAE computes the absolute difference between the real value and the value predicted. By the end, the MAD is computed based on the median of the values and is more resistant to outliers than the other measures.

Despite MAD be a measure more resistant to outliers, the preliminary results showed that the data provides a situation where a lot of the complaints and answers have not a value to be pointed as a complaint or refund amount. Because of this, there is a huge occurrence of 0 (zero) target values, and therefore to adopt the MAD as the deviation measure is not appropriate.

Once MSE is a measure that computes the squared differences, single observations with large prediction errors influence a lot MSE. Once the problem has the potential to generate some large prediction errors, we do not adopt MSE either the RMSE as the main measures in this work.

On this context, adopt is the MAE as the deviation measure, that is also the most intuitive between the regression measures.

In the analysis of the results, we also consider the *accuracy*, that represent the percentage of "perfect matches" of values in the predictions.

One the other hand, the tasks 2 and 4, dedicated to finding values asked for a refund by customers and refunded by companies, are not regression task in the original concept, despite the measures of regression can be applied. Therefore, besides uses the measures of regression tasks, we also adopted the measures for NER tasks in order to evaluate these tasks.

Adopting also NER/Classification Measures

It is convenient to remember the preliminary results, in order to point that in 58.50% of the complaints the customer is asking for a refund and, among these requests, it is possible to a human identify the value complained in just 25% of cases.

On the other hand, from the point of view of companies answers, in 46% of all complaints the company is describing a refund to the customer and, among these cases, it is possible to a human identify the value refunded in just 62% of cases.

These statistics are an evidence of the problems complexity because in a lot of cases the values complained and refunded are not described in the text. Cases like that point to concern from Anatel, because the algorithms must avoid making predictions in cases in which the monetary values described in the text do not have a relationship with the values complained or refunded.

One approach to this situation would be to include a classification task inside the Regression/NER tasks, as a subtask, in order to observe if it is possible to identify the values related to refund in the text before predict them. But we decide to treat this issue in another way.

In the face of this situation, and inspired by the use of NER measures presented in literature chapter, we adopt the use of classification measures to observe the situations in with the model predicts values despite there are no values related to a refund in the text.

In order to control this behavior in the model, we adopt the following concepts for False Positive (FP), False Negative (FN), True Negative (TN) and True Positive (TP) predictions:

- False Positive (FP): The model predicts a value, but there is no real value in the text;
- False Negative (FN): The model does not predict any value, but there is a real value in the text;
- True Negative (TN): The model does not predict any value, and there is no real value in the text;
- True Positive (TP): The model predicts a value, and there is a real value in the text.

With these conventions in mind, we intend to find a model that generates fewer False Positives predictions, in order to avoid inform monetary values that have no relationship with values complained or refunded.

Therefore, for task 2 and task 4, because of the singularity of the tasks, in evaluation we combine a mix of less deviation (from the point of view of regression measures), accuracy (considering exact matches) and Precision (from the point of view of a NER/Classification task).

Experimental Setup

We perform the experiments relate to Task 2 and Task 4 using a subset of the 400 cases (complaints and answers) tagged by Anatel specialists. We obtain this subset extracting just the customer complaints that contain an asking for a refund, and the company answers that point to a refund. As a result of this extraction, we perform tasks 2 and 4 just over the 234 complaints that contain an asking for a refund (result of Task 1), and just over the 184 answers that describe a refund (result of Task 3).

We implement Task 2 and Task 4 in 2 waves, considering 2 different approaches. In the first one, the algorithm observes just a presence of a word that point an asking for refund (in case of customer text) or a word that point a refund (in case of company text).

In the second approach, the algorithm observes still the words close the values, in order to check if is the case of a *sum* or a *subtraction* of the value. We highlighted that to perform the second approach, we need to reprocess the text, not removing the stop-words, because in this approach the stop-words are important to point to the semantic of the value (if it must be summed, subtracted, multiplied).

Regarding Task 2, we perform the experiment 10 times: 5 adopting the first algorithm for ranges $r = 2$, $r = 3$, $r = 4$, $r = 5$ and $r = 10$, and 5 more times adopting the second algorithm for ranges $r = 2$, $r = 3$, $r = 4$, $r = 5$ and $r = 10$.

We performed Task 4 also 10 times, adopting the same strategy that Task 2.

5.2.2 Results using the first algorithm

Table 5.7 contains the scores for Task 2 considering the first algorithm. Here, r represents the range of grams (tokens) the algorithm will consider to analyze from each side the monetary value. For instance, for an $r = 2$, the algorithm will consider 2-grams from the left and 2-grams from the right.

Table 5.7: Scores for Task 2, considering the first algorithm

	$r = 2$	$r = 3$	$r = 4$	$r = 5$	$r = 10$
Deviation Measures					
MAE	104.327	112.105	127.941	128.856	143.604
MSE	613242.273	611481.319	640851.636	640781.815	815964.263
MAD	0.0	0.0	0.0	0.0	0.0
Accuracy	0.726	0.709	0.700	0.696	0.696
NER/Classification Measures					
F1	0.485	0.540	0.530	0.539	0.595
Recall	0.431	0.517	0.517	0.534	0.620
Precision	0.555	0.566	0.545	0.543	0.571

Table 5.8 presents the scores for Task 4 considering the first algorithm.

Table 5.8: Scores for Task 4, considering the first algorithm

	r = 2	r = 3	r = 4	r = 5	r = 10
Deviation Measures					
MAE	93.476	127.521	135.205	133.953	164.805
MSE	39047.703	374085.557	377107.923	377794.500	398097.728
MAD	26.140	25.000	27.210	25.000	36.680
Accuracy	0.423	0.456	0.456	0.461	0.434
NER/Classification Measures					
F1	0.312	0.532	0.586	0.659	0.753
Recall	0.192	0.394	0.473	0.570	0.736
Precision	0.814	0.818	0.771	0.783	0.770

5.2.3 Results using the second algorithm (considering the "weight" of the value)

Table 5.9 contains the scores for Task 2 considering the second algorithm. Here, r represents the range of grams (tokens) the algorithm will consider to analyze from each side the monetary value. For instance, for an $r=2$, the algorithm will consider 2-grams from the left and 2-grams from the right.

Table 5.9: Scores for Task 2 considering the second algorithm

	r = 2	r = 3	r = 4	r = 5	r = 10
Deviation Measures					
MAE	62.186	105.373	112.799	108.973	134.774
MSE	65943.766	612175.516	617822.201	608452.815	802937.422
MAD	0.0	0.0	0.0	0.0	0.0
Accuracy	0.747	0.726	0.717	0.713	0.696
NER/Classification Measures					
F1	0.033	0.448	0.504	0.527	0.539
Recall	0.017	0.379	0.465	0.500	0.534
Precision	0.500	0.550	0.551	0.557	0.543

On the other hand, table 5.10 shows the scores for Task 4 considering the second algorithm.

Table 5.10: Scores for Task 4, considering the second algorithm

	r = 2	r = 3	r = 4	r = 5	r = 10
Deviation Measures					
MAE	97.057	93.246	93.400	88.057	140.834
MSE	40502.910	38938.251	7800.590	35292.680	379360.710
MAD	30.19	29.990	29.995	25.000	27.210
Accuracy	0.380	0.418	0.407	0.440	0.445
NER/Classification Measures					
F1	nan	0.309	0.351	0.450	0.649
Recall	0.0	0.192	0.228	0.315	0.561
Precision	nan	0.785	0.764	0.782	0.771

5.3 Choosing the Models

As a consequence of the problem nature and Anatel assumptions about the models, the main measure that we consider is the *Precision*, even with a loss of accuracy. In the case of close precision scores, other measures are observed in the choosing of the models.

This decision is a consequence of the assumptions that point to minimize false positives in the classification tasks, and also minimize wrong predictions in the regression/NER tasks.

We justify this strategy because of the potential to use these models to automate real regulatory tasks. In this context, we must avoid false positive predictions for all the tasks.

Therefore, for task 1, we choose the model performed with Word2Vec + Heuristic + Metadata with the Decision Tree classifier, which presents the Precision of 0.867 in the predictions of customer complaints that contain an asking for a refund. This is the model that presents the highest Precision score, and means that the model is right in 86% of the true positive predictions.

For task 2, we choose the model performed with the second algorithm and $r = 3$, which presents the Precision = 0.550, Accuracy = 0.726, and MAE = 105.373 in the predictions of the values asked for a refund in customer complaints. We consider this is the model that contains the most balanced scores (Precision, Accuracy, MAE) compared with the other models. The "exact match" (accuracy) means that the algorithm identifies correctly the value refunded (including value '0') in 72% of cases. The MAE represents the mean absolute error in the values predicted, and the Precision means that the algorithm identifies the cases where there is real value to be predicted in 55% of the true positive predictions.

On the side of companies answers, for task 3, we choose the model performed with Word2Vec + Heuristic + Metadata with the Random Forest classifier, which presents the precision of 0.745 in the predictions of companies answers that contain refund. This is the model that presents the highest Precision score, and means that the model is right in 74% of the true positive predictions.

Finally, for task 4, we choose the model performed with the second algorithm and $r = 5$, which presents the Precision = 0.782, Accuracy = 0.044, and MAE = 88.057 in the predictions of the values asked for a refund in customer complaints. We consider this is the model that contains the most balanced scores (Precision, Accuracy, MAE) compared with the other models. The "exact match" (accuracy) means that the algorithm identifies correctly the value refunded

(including value '0') in 44% of cases. The MAE represents the mean absolute error in the values predicted, and the Precision means that that algorithm identifies the cases where there is real value to be predicted in 78% of the true positive predictions.

Chapter 6

Conclusions, Contributions and Future Works

6.1 Conclusions

Regarding the thesis relevance, the work involves real, sensitive and non-tagging data from Brazilian citizens and telecommunication companies that interact with "Anatel Consumidor". Therefore, the topic has high interest to Anatel, once the organization can use the results of the work to improve its regulatory mission over one of the most strategical Brazilian sectors (telecommunication), taking more accurate decisions with high social and economic impact.

In fact, from Anatel point of view, the results obtained with this work allow identifying behavior patterns in customers complaints and in company answers, as to obtain relevant estimations about this topic, including in monetary aspect.

From the academic point of view, the work is an example of the adoption of Natural Language Processing over a problem involving real data from the Brazilian telecommunication sector.

Also, it is timely to highlight that the work is innovative to Anatel, once offers instruments to perform automatically valuable analysis over a huge quantity of text, in contrast with an activity made nowadays over a minimum fraction of the data and manually.

Therefore, the work aggregates value at the same time to the academic sector, to the governmental sector and to citizens that will be benefited from the results.

Regarding the results, for Task 1, related to predict if the customer is asking for a refund, we obtain a classification model with Precision 0.867. This means that the model is right in 86% of the true positive predictions.

For Task 2, related to detect how much the customer is complaining, we obtain a model with MAE 105.373, Accuracy (Exact Match) of 0.726, and Precision 0.550. The "exact match" means that the algorithm identifies correctly the value refunded (including value '0') in 72% of cases. The MAE represents the mean absolute error in the values predicted, and the Precision means that that algorithm identifies the cases where there is real value to be predicted in 55% of the true positive predictions.

For Task 3, related to predict if the company is refunding the customer, we obtain a classification model with Precision 0.745. This means that the algorithm is right in 74% of the true

positive predictions.

Finally, for Task 4, related to detect how much the company is refunding, we obtain a model with MAE 88.057, Accuracy (Exact Match) of 0.440 and Precision 0.782. The "exact match" means that the algorithm identifies correctly the value refunded (including value '0') in 44% of cases. The MAE represents the mean absolute error in the values predicted, and the Precision means that that model identifies the cases where there is real value to be predicted in 78% of the true positive predictions.

6.2 Contributions

The work presents the following contributions to the Natural Language Processing (NLP) field of study:

1. We present a concrete case study applying *text classification tasks* over sensitive data in real telecommunication complaints involving the following topics:
 - (a) We present an analysis about different ways the customer express their asking for a refund, as well the different ways the companies express their refunding;
 - (b) We present a statistical analysis about customer complaints and companies answers over a sample extracted randomly and classified by Anatel specialists;
 - (c) We construct heuristics for identifying customer complaints involving asking for refund and companies answers involving refunding;
 - (d) We construct an automated process to identify situations where the customers have asked for a refund, and identify the value complained;
 - (e) On the other hand, we construct an automated process to identify situations where the companies have refunded something, and identify the value refunded;
 - (f) We compare the results between simple heuristics versus more sophisticated representation models;
2. We present a concrete case study applying Named Entity Recognition (NER) over sensitive data in real telecommunication complaints involving the following topics:
 - (a) We present heuristics for identifying how much customers have asked for a refund and how much companies have refunded;
 - (b) We construct an automated process to quantify the monetary values the customers complain and the values the companies refund.

6.3 Future Work

Henceforth, the efforts in this study are in order to improve the results based on several aspects.

First, the algorithms used in the steps of training the embedding model and training the classifiers (like Decision Tree and Gradient Boosting) involve several parameters. In this work, we do manually and just superficially the changing of these parameters, but there are instruments to optimize this setup automatically. Is the case of adopting *hyperparameter optimization* in order to obtain automatically the best setup for the learning models.

Another clear opportunity for improvement in the classification tasks is related to *enlarging the corpus* in order to train better the embedding models because these models have the cost of requiring a massive corpus to perform better.

Still related to the classification tasks, the *adoption of Active Learning* represent an opportunity to enlarge the training dataset by the use of this technique to label automatically the complaints.

Once we adopt supervised learning, have correct labels is critical to obtains good performance. Hence, the *reviewing of labels in train and test datasets* to fix errors arising from human tagging represents an opportunity for improvement.

Another clear opportunity is the *adoption of more sophisticated representation models*, like BERT, in order to obtain the benefits of technologies that represent the stat of the art in NLP.

In the same direction, we can invest in the *construction of more sophisticated NER rules* to detect better the result of different values in customer complaints and companies answers.

By the end, we visualize opportunities in *detecting more specifics of each task* in order to fit better the general models to each case. For instance, despite Task 1 and Task 3 have a similar nature, there are particularities of each one that, if observed, have the potential to improve the work.

Bibliography

- Agranonik, M., & Hirakata, V. N. (2011). Cálculo de tamanho de amostra: proporções. *Clinical & Biomedical Research*, 31(3).
- Banerjee, S., Chakrabarti, S., & Ramakrishnan, G. (2009). Learning to rank for quantity consensus queries. In *Proceedings of the 32nd international acm sigir conference on research and development in information retrieval* (pp. 243–250).
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146.
- DA GAMA, J. M. P., FERREIRA, A. C. P. D. L., CARVALHO, D., FACELI, K., LORENA, A. C., & OLIVEIRA, M. (2017). *Extração de conhecimento de dados: data mining*. Sílabo.
- de Aquino Silva, R., da Silva, L., Dutra, M. L., & de Araujo, G. M. (2020). An improved ner methodology to the portuguese language. *Mobile Networks and Applications*, 1–7.
- de Freitas Junior, E. M. (2013). *Análise de fatores críticos em reclamações de clientes de concessionárias de distribuição de energia elétrica* (Master Thesis). Universidade Federal de Pernambuco, <https://repositorio.ufpe.br/handle/123456789/11809>.
- de Telecomunicações, A. N. (2021). *Anatel - reclamações*. Retrieved from <https://www.anatel.gov.br/paineis/consumidor/reclamacoes>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Ferreira, J. D. C. (2019). *Python para pré-processamento e extração de características a partir de texto português* (Master Thesis). Universidade de Coimbra, <http://hdl.handle.net/10316/88030>.
- Grljević, O., & Bošnjak, Z. (2018). Sentiment analysis of customer data. *Strategic Management*, 23(3), 38–49.
- Gualberto, E. S. (2020). *Phishing detection : methods based on natural language processing* (Doctoral dissertation). Universidade de Brasília, <https://repositorio.unb.br/handle/10482/40415>.
- Ho, V. T., Pal, K., Kleer, N., Berberich, K., & Weikum, G. (2020). Entities with quantities: Extraction, search, and ranking. In *Proceedings of the 13th international conference on web search and data mining* (pp. 833–836).
- Hui-nan, M., Jiang-ning, W., & Yan-zhong, D. (2006). Ontology-driven information retrieval system for customers' complaints of mobile communication services. In *2006 international conference on management science and engineering* (pp. 30–35).
- Jurafsky, D. (2000). *Speech & language processing*. Pearson Education India.
- Kaur, G. (2014). Usage of regular expressions in nlp. *International Journal of Research in Engineering and Technology IJERT*, 3(01), 7.

- Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4), 150.
- Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.
- Liashchynskiy, P., & Liashchynskiy, P. (2019). Grid search, random search, genetic algorithm: a big comparison for nas. *arXiv preprint arXiv:1912.06059*.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguistic Investigations*, 30(1), 3–26.
- Nayak, P. (2019). *Understanding searches better than ever before*. Retrieved from <https://blog.google/products/search/search-language-understanding-bert/>
- Pawluszek-Filipiak, K., & Borkowski, A. (2020). On the importance of train–test split ratio of datasets in automatic landslide detection by supervised classification. *Remote Sensing*, 12(18), 3054.
- Pereira, S. L., & de Brito Vieira, T. P. (2006). Estudo da aplicação de um processo gerenciado de produção de software em mpes. *Centro de Ciências Sociais Aplicadas. Universidade Federal da Paraíba. João Pessoa*.
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.
- Rácz, A., Bajusz, D., & Héberger, K. (2021). Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification. *Molecules*, 26(4), 1111.
- Roberts, K. (2016). Assessing the corpus size vs. similarity trade-off for word embeddings in clinical nlp. In *Proceedings of the clinical natural language processing workshop (clinicalnlp)* (pp. 54–63).
- Settles, B. (2009). Active learning literature survey.
- Sousa, G. N. d., Guimarães, I. d. S., Viana, J. A. N., Reinhold, O., Jacob Junior, A. F. L., & Lobato, F. M. F. (2020). Análise do setor de telecomunicação brasileiro: Uma visão sobre reclamações. *RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação*(37), 31–48.
- Thanaki, J. (2017). *Python natural language processing*. Packt Publishing Ltd.
- Union, I. T. (2020). *Statistics*. Retrieved from <https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx>
- Wallace, E., Wang, Y., Li, S., Singh, S., & Gardner, M. (2019). Do nlp models know numbers? probing numeracy in embeddings. *arXiv preprint arXiv:1909.07940*.
- Wirth, R., & Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1).
- Yadav, V., & Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.