LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

MESTRADO

Gestão de Sistemas de Informação

TRABALHO FINAL DE MESTRADO

TRABALHO DE PROJETO

QUALIDADE DOS DADOS & MACHINE LEARNING: UMA NOVA ABORDAGEM AOS CENSOS POPULACIONAIS E HABITACIONAIS

INÊS MARGARIDA SILVA PAZ LOPES

OUTUBRO - 2019

LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

MESTRADO EM

GESTÃO DE SISTEMAS DE INFORMAÇÃO

TRABALHO FINAL DE MESTRADO

TRABALHO DE PROJETO

QUALIDADE DOS DADOS & MACHINE LEARNING: UMA NOVA ABORDAGEM AOS CENSOS POPULACIONAIS E HABITACIONAIS

INÊS MARGARIDA SILVA PAZ LOPES

ORIENTAÇÃO:

PROFESSOR DOUTOR ANTÓNIO MARIA PALMA DOS REIS

OUTUBRO – 2019

Agradecimentos

Durante a execução deste trabalho, foram muitos os que se destacaram pelo seu carinho, apoio e incentivos. Sem dúvida que estes foram os ingredientes principais deste projeto.

Primeiramente, os meus agradecimentos vão para o Professor Doutor António Palma dos Reis, por toda a colaboração e apoio prestado ao longo destes meses, por toda a boa vontade e disponibilidade que sempre demonstrou para comigo e para com o meu projeto. Obrigada pela sua paciência, pelos seus conhecimentos e pelas suas críticas construtivas.

Em segundo lugar, os meus agradecimentos à empresa da qual faço parte, realçando os meus companheiros de projeto com quem partilhei a minha primeira experiência profissional. Ao líder de equipa dirijo um enorme obrigada, pela oportunidade de colocar em papel o nosso trabalho de meses e fazer dele o meu Trabalho Final de Mestrado.

Por fim e de uma forma especial, agradeço à minha família e aos amigos mais próximos, por me motivarem e apoiarem incondicionalmente etapa a etapa, objetivo a objetivo, sonho a sonho.

"Para hacer las cosas bien es necesario: primero, el amor; segundo, la técnica."

António Gaudí

Lista de Siglas e Acrónimos

CAE – Código de Atividade Económica

INE – Instituto Nacional de Estatística

ML – Machine Learning

NLP – Natural Language Processing

OCR – Optical Character Recognition

UNSD - United Nations Statistics Division

Resumo

Os Censos Populacionais e Habitacionais são uma fonte imensa de dados relativos a uma determinada população e país.

O projeto realizado consiste no processo de recolha e preparação de dados manuscritos em papel, da aplicação do inquérito Censo Populacional e Habitacional a uma população de mais de vinte milhões de pessoas.

Este é um tipo de inquérito que se faz à população de um país, tendo como objetivo retirar conclusões a nível geográfico tanto da população, como das suas condições de vida. Os Censos são realizados com alguma frequência, o que permite efetuar comparações e perceber a transformação da sociedade e de um país, ao longo dos anos.

Com o objetivo de tornar os mais de vinte milhões de inquéritos manuscritos em informação útil e de qualidade acerca de um país e de uma população foi necessário dividir o trabalho em três fases, a fase recolha de dados e da sua conversão de imagem para um formato digital onde o texto possa ser editável, a fase de limpeza e tratamento dos dados e, por último, a fase de análise e classificação dos mesmos.

De acordo com cada fase, foram utilizadas diversas metodologias e tecnologias, como é o caso do OCR (*Optical Character Recognition*), NLP (*Natural Language Processing*) e *Machine Learning,* respetivamente. Estas abordagens permitiram uma melhor, mais rápida e mais fiável análise de resultados.

Ao longo do projeto testaram-se diversos cenários e algoritmos de forma a obter os melhores resultados possíveis, tanto a nível da qualidade de dados, como a nível do modelo de classificação de dados. Desta forma, a equipa realizou diversos estudos comparativos, de forma a alcançar o melhor modelo de classificação dos dados possível para o nosso contexto operacional.

A principal conclusão retirada ao longo de todo o projeto foi que quanto maior a qualidade dos dados, mais eficaz e mais eficiente é o modelo. Assim, em futuros projetos semelhantes, devemos focar as nossas atenções no processo inicial, ou seja, no processo de recolha dos dados. Quanto mais organizados e mais percetíveis forem os nossos inputs, mais rápido e fácil é a sua análise, interpretação e classificação.

Palavras-chave: Processamento de Linguagem Natural, Censos, Distância de *Levenshtein, Machine Learning*, Modelo de *Naïve Bayes,* Qualidade dos Dados, Reconhecimento Ótico de Caracteres.

Abstract

The Population and Housing Census is a huge source of data for a particular population and country.

The project undertaken consists on the process of collecting and preparing paper handwritten data obtained from the Population and Housing Census survey applied to a population of over twenty million people.

This type of inquiry done to the population of a country has the purpose of drawing up conclusions and insights on the populations' geographical characteristics, as well as their life conditions. These censuses are done on a frequent basis, which allows for continuous comparisons to be done and thus understand the changes occurring in a given society and country throughout time.

In order to turn more than twenty million handwritten surveys into useful and quality information about a country and a population, it was necessary to divide the work into three phases. The first stage consisted on the collection of data and its conversion into an image in a digital format, where text can be edited, followed by data cleansing and transformation, and finally, the third stage involved the analysis of the data and its respective classification.

In regards to the data analysis, for each sentence there were various methodologies and technologies applied, such as OCR *(Optical Character Recognition), NLP (Natural Language Processing) e Machine Learning.* This approach led to a better, quicker and more reliable analysis of the data.

Throughout the project several scenarios and algorithms were tested in order to obtain the best possible outcomes, concerning data quality, as well as the data classification model. By doing so, the team was able to undergo a

number of diverse comparative studies, allowing the achievement of the most accurate data classification model according to the operational context.

The main conclusion obtained throughout the project was that when the quality of the data is higher, the more effective and efficient the model will be. And so, the message that should be conveyed to other projects in the future is that the focus of our attention must be on the initial stage of the process, being the collection of data. If our inputs are more organised and perceptible, the quicker and easier it will be to analyse, interpret and classify those same inputs.

Key Words: Census, Data Quality, Levenshtein Distance, Machine Learning, Natural Language Processing, Optical Character Recognition, Naïve Bayes Model.

Índice Geral

Índice de Figuras

Índice de Tabelas

1. Introdução

1.1. Enquadramento do tema

Big Data são uma realidade à qual estamos expostos no dia a dia. O ser humano está constantemente a produzir informação, como tal é fundamental criar e encontrar processos que nos permitam analisar grandes quantidades de dados a que estamos sujeitos.

O projeto tem como input mais de vinte milhões de recenseamentos, o que se traduz em dezenas de milhões de dados que requerem tratamento e análise. Tal tarefa iria consumir bastante tempo a nível de trabalho humano, tornando-se numa solução pouco sustentável. Deste modo, tornou-se essencial optar por metodologias baseadas na inteligência artificial, que permitem a produção e disseminação de resultados de qualidade, num curto prazo de tempo.

Realizados de 10 em 10 anos como recomendam as boas práticas internacionais, os Censos Populacionais e Habitacionais de um país referem-se ao processo total de coleta, compilação, avaliação, análise e publicação ou divulgação de dados demográficos, económicos e sociais, referentes a todas as pessoas, num país, numa data específica.

Em países onde os Censos são recolhidos de forma manuscrita em papel, como é o caso do país do projeto em estudo, acresce à análise de dados, todo o processo de recolha, preparação e limpeza dos dados. Desta forma, conclui-se que estamos perante um desafio que envolve conceitos como *Big Data,* Qualidade dos Dados e *Machine Learning*.

O presente projeto está organizado em cinco capítulos. O primeiro capítulo corresponde à introdução, onde é realizado o enquadramento do tema do projeto e é realizada uma descrição dos objetivos do projeto. O segundo capítulo diz respeito à revisão da literatura, no qual são expostos os principais conceitos e metodologias que se aplicam ao contexto do projeto. Posteriormente, o terceiro capítulo corresponde à execução do projeto, onde é descrito todo o processo de desenvolvimento e as metodologias utilizadas. Segue-se o quarto capítulo, a discussão de resultados, onde são discutidos os métodos utilizados e os resultados obtidos com cada um, de forma a justificar o porquê da sua utilização. Por fim, o capítulo quinto corresponde à conclusão e lições aprendidas durante a execução do projeto, de forma a retirar ilações para projetos futuros.

1.2. Descrição e Objetivos do Projeto

O projeto em estudo baseia-se na recolha, limpeza e análise de dados recolhidos de um recenseamento aplicado a uma população de um país africano, os Censos Populacionais e Habitacionais. É um país que conta com mais de vinte milhões de habitantes, como tal foram muitos os inquiridos, contudo o projeto foca apenas uma parte da população, tendo sido recolhida uma amostra aleatória, que representa cerca de 0,6% do total da população.

Este projeto enfrentou diversos desafios de interpretação de dados ao nível das respostas ao questionário aplicado à população. Esses desafios relacionam-se com a escrita, uma vez que existe uma grande variedade de caligrafias, erros ortográficos, incoerência de respostas, terminologias menos adequadas, entre outros erros derivados do tipo de população em estudo. Uma

população caracterizada por ter uma baixa taxa de literacia, de um país pobre, em desenvolvimento, com um baixo nível de desenvolvimento humano, onde se falam mais de 20 dialetos diferentes.

As respostas ao inquérito estão separadas por pergunta, sendo algumas de escolha múltipla, contudo, a grande maioria é de resposta livre em formato quadriculado, isto é, escreve-se uma letra por quadrícula e utilizando letras maiúsculas.

Neste sentido, o primeiro objetivo do projeto foi conseguir recuperar o máximo de respostas possíveis de forma a obter o maior número de dados da população para realizar a análise aos Censos. Quanto mais inputs de respostas aos questionários existirem, mais real e mais exato é o estudo da população e do país em questão.

Posteriormente ao processo de recuperação de expressões, seguiu-se a fase da análise da qualidade dos dados. Tal como foi referido acima, os erros derivam de uma baixa taxa de alfabetização por parte da população em estudo e também, da existência de uma panóplia de dialetos integrados na população, o que leva ao uso de expressões características de difícil compreensão. Foi necessário corrigi-los de forma a facilitar a análise dos mesmos. Desta forma, foi criado um dicionário com uma grande variedade de palavras de forma a cobrir todas as possibilidades de resposta e, assim, deu-se início ao processo de limpeza dos dados.

A correção dos dados através da comparação das respostas com as palavras do dicionário foi um desafio, foi necessário realizar diversos estudos

com algoritmos de medidas de distâncias, de forma a descobrir aquele que melhor desempenho tinha, isto é, o que mais palavras corrigia.

Por fim, após termos as respostas corrigidas e organizadas por pergunta, deu-se início ao objetivo principal do projeto, a análise e disseminação de resultados.

A análise foi efetuada por pergunta, sendo as perguntas de escolha múltipla mais fáceis de analisar do que as perguntas de resposta aberta. Como tal centrámos as nossas atenções nestas últimas, as perguntas de resposta aberta.

Apesar de nesta fase, as respostas estarem alfabeticamente corrigidas, algumas não estavam coerentes com a pergunta, por exemplo, na pergunta relativa à cidade onde mora, muitas vezes, a resposta não era uma cidade, mas sim, uma província. O mesmo acontecia em relação ao distrito e outras perguntas. No caso da pergunta referente à profissão do indivíduo, verificou-se que a população utiliza palavras e expressões diferentes para definir a mesma atividade, por exemplo, associado à profissão “Agricultor”, foram encontradas respostas como “Plantador de batatas”, “Semeio batatas”, “Cultivo terras”, entre muitas outras.

Deste modo, com foi identificada uma nova necessidade – Classificar as respostas à pergunta profissão, de forma a agrupá-las segundo a sua atividade profissional.

Após a identificação da necessidade e do objetivo, chegou-se à conclusão de que estávamos perante um problema de classificação de dados, que poderia

ser resolvido através da implementação de um modelo de classificação através do uso de *Machine Learning*.

2. Revisão da Literatura

Para medir e avaliar a evolução das populações e dos próprios países são realizados Censos Populacionais e Habitacionais de dez em dez anos. Segundo o Instituto Nacional de Estatística, INE, (INE, 2011) os Censos são comparados a uma "fotografia" da população e das suas condições de habitabilidade, resultando em informações acerca do número de indivíduos de um país, nível de escolaridade, condições de vida, atividades profissionais mais predominantes, entre muitos outros indicadores demográficos e socioeconómicos. Estes dados permitem não só uma avaliação da evolução da população do país em questão quando comparados com dados de anos anteriores, mas também permitem a realização de comparações com outros países.

A realização dos Censos Populacionais e Habitacionais exige um grande investimento, não só em termos monetários, como também em esforço humano, uma vez que este inquérito envolve milhões de inquiridos, o que se traduz em dezenas de milhões de dados. De acordo com a Divisão Estatística das Nações Unidas, em inglês United Nations Statistics Division, UNSD, uma instituição com uma longa história em sólidos princípios estatísticos e partilha de conhecimento, é notório que nos últimos anos tem existido um grande investimento na utilização de novas fontes, ferramentas, métodos e tecnologias de dados para recolha e análise de *Big Data* para estatísticas oficiais (UNSD, 2019). Contudo, apesar de estarem marcados muitos simpósios e conferências relacionadas com o tema, ainda não é omnipresente na aplicação de Censos Populacionais e Habitacionais em todos os países, sendo que existem grandes discrepâncias,

destacando o continente Asiático como o mais avançado em termos de tecnologia aplicada a dados estatísticos como recenseamentos.

No continente africano, o continente de detém 33 dos 49 países menos desenvolvidos (Gomes, 2009), já existe tecnologia aplicada aos Censos. Em 2018, a Argélia utilizou *tablets* com tecnologia móvel 3G, como ferramenta de comunicação, de forma a recolher as respostas aos Censos Populacionais e Habitacionais e a inseri-las, automaticamente, numa base de dados própria. Em comparação com os questionários em papel, o uso da tecnologia, neste caso associada aos *tablets*, permite ter acesso aos dados em tempo real, pois as respostas vão instantaneamente para uma base de dados; permite uma maior precisão e uma menor taxa de erros, dado que o formato digital consegue efetuar uma correção das respostas de forma automática; e, ainda, proporciona uma maior perceção das necessidades e dificuldades dos inquiridos (Bourezgue, 2017).

De acordo com Dekker (2001), pequenas melhorias a nível da tecnologia em operações de Censos, podem representar grandes ganhos a nível de qualidade e de custos totais, contudo exige uma análise sobre certos aspetos que podem influenciar o desempenho da tecnologia, como por exemplo:

- Decidir qual a tecnologia mais apropriada;
- Manter a integridade de sistemas estatísticos;
- Realizar *outsourcing* de algumas tarefas;
- Mantar a confidencialidade dos dados.

Estes são pontos que carecem de análise prévia à realização dos Censos e dependem da capacidade tecnológica e monetária de cada país, da magnitude

do projeto, experiências anteriores e tempo de preparação prévias, por exemplo. O planeamento da realização de um projeto de Censos deve ser um misto de inovador com conservador, pois, a solução utilizada, seja mais ou menos tecnológica, tem que resultar à primeira.

Apesar das diversas vantagens apresentadas em cima, a tecnologia ainda não chega a todos os cantos do mundo. Muitos países continuam a utilizar questionários em papel, para realizar os seus Censos por possuírem menos opções tecnológicas, no entanto existem outras metodologias e soluções tecnológicas que se podem aplicar aos recenseamentos manuscritos em papel, quer seja para facilitar no (2.1) reconhecimento de caracteres, para a (2.2) recuperação e limpeza de dados das respostas ao Censos e, por fim, para auxiliar na (2.3) análise dos dados (Dekker, 2001).

2.1. Reconhecimento de Caracteres

Censos realizados através de questionários manuscritos em papel com perguntas de resposta aberta são um grande desafio, dado que permitem uma vasta utilização de vocabulário, o uso de qualquer material de escrita, lápis ou canetas e uma grande variedade de estilos de escrita e de erros (Cao et al., 2012).

Entre 2000 e 2001, vários países como a Estónia, Tailândia, Filipinas, Macau, Indonésia e Aruba realizaram censos manuscritos em papel, e utilizaram o Reconhecimento de Caracteres Inteligente, em Inglês, *Intelligent Character Recognition* (IRC), uma tecnologia estendida do Reconhecimento Ótico de Caracteres (OCR) (Dekker, 2001). OCR é o processo de conversão de

documentos digitalizados escrito à mão para um formato editável (Alotaibi et al., 2017), que não considera informação temporal, uma vez que é um processo *offline,* isto pressupõe que os erros não podem ser corrigidos *a posteriori* (Keysers et al., 2017).

Uhliarik (2013) e Alotaibi et al. (2017) defendem que um sistema de ORC passa pelas seguintes etapas: 1) pré-processamento da imagem que contém o documento manuscrito, esta fase inclui a remoção de ruído e normalização das imagens através do corte e dimensionamento, de forma a deixar apenas a informação importante; 2) reconhecimento de caracteres utilizando a segmentação algorítmica de palavras e extração de caracteres, separando-os para posterior identificação individual; 3) classificação dos caracteres por comparação a uma base de dados.

2.2. Limpeza de dados

O *output* do OCR provindo de documentos manuscritos digitalizados exige uma correção linguística uma vez que existe uma grande probabilidade de não conseguir reconhecer todas as palavras corretamente.

Segundo Chen et al. (2019), o Processamento de Linguagem Natural, NLP, permite realizar uma análise a um elevado número de dados, estruturando a informação, deixando-a apta para posteriores análises. Xie et al. (2019) e Alotaibi et al. (2017) afirmam que o NLP inclui três processos principais, análise lexical, onde se inclui a segmentação de palavras e o reconhecimento de classes identificadas; análise semântica, relacionada com o significado e contexto; e, por fim, análise sintática, ou seja, análise da sintaxe e da gramática. Estes processos

permitem ao NLP aumentar a eficiência do modelo proposto, uma vez que o seu procedimento natural passa por realizar uma correção ao texto, através da correspondia entre o resultado do OCR e um dicionário, mediante a semelhança entre as palavras (Kukish, 1992).

Gomaa et al. (2013) defende que as palavras são semelhantes no seu léxico quando existe semelhança na sequência de caracteres que as constituem. Para medir essa mesma semelhança, os autores definem medidas "*String-Based*", medidas que medem a distância de semelhança ou diferença entre palavras que estejam a ser comparadas. Deste modo, para fins de correspondência de texto, existem vários algoritmos, sendo que para o contexto do projeto foram estudados os algoritmos em baixo apresentados.

2.2.1. Distância de Levenshtein

Segundo Haldar et al. (2001) a Distância de Levenshtein é utilizada para medir a distância entre duas palavras similares, sendo que o valor da distância é o número total de edições necessárias para transformar a palavra (x) na sua correção (y). As operações de edição podem ser de diferentes tipos: eliminações, inserções ou substituições de caracteres. Um estudo de Damerau (1964), considera ainda outro tipo de edição, a transposição, ou seja, a permuta de posição entre dois caracteres (Hicham et al. 2012). Por exemplo, para as palavras "apto" e "pato" a Distância de Levenshtein é dois, ocorrem duas substituições, "a" por "p" e "p" por "a", e a Distância de Damerau-Levenshtein é um, ocorre apenas uma transposição das letras "a" e "p".

2.2.2. Similaridade de Jaccard

Esta métrica consiste em dividir o tamanho da interseção das palavras pelo tamanho da união das mesmas, ou seja, é o rácio entre o número de caracteres semelhantes sobre o número total de caracteres das palavras. O valor da semelhança de Jaccard varia em 0 e 1, quanto mais perto de 1 mais semelhantes são as palavras, e quanto mais perto de 0 mais diferentes são (Gomaa et al., 2013; Sarkar et al., 2015).

2.2.3. Distância de Jaro-Winkler

A Distância de Jaro-Winkler é uma associação do algoritmo de Jaro com o algoritmo de Winkler. Jaro (1989) criou uma medida semelhança de caracteres que tem em conta inserções, exclusões e transposições de caracteres, incluindo erros ortográficos. Uns anos depois, Winkler (1994) modificou esse algoritmo apoiando a ideia de que diferenças de caracteres no início da palavra são mais significativas do que diferenças perto do final, como tal as primeiras têm mais peso (Gomaa et al., 2013).

2.2.4. Distância de Needleman-Wunsch

Needleman-Wunsch (1970) é um algoritmo de programação dinâmica utilizado no alinhamento global em pares de duas sequências biológicas, demonstrando-se mais adequado quando as duas sequências têm um grau significativo de semelhança. O algoritmo atribui pontuação aos caracteres de

acordo com uma matriz de similaridade. Somando a pontuação final obtém-se a Distância de Needleman-Wunsch (Nanni et al., 2008; Gomaa et al., 2013).

2.2.5. Distância de Smith-Waterman

O algoritmo de Smith-Waterman (1981) é outro exemplo de algoritmo de programação dinâmica. Neste caso, o algoritmo procura o melhor alinhamento, construindo uma tabela de soluções. É útil para sequencias com um grau de diferença maior, pois procura possibilidades de semelhança entre elas (Poole et al., 2010; Gomaa et al., 2013).

O contributo do NLP juntamente com as medidas de distância na correção contribuem de forma positiva para a correção de dados, contudo podem existir problemas de ambiguidade na correção das palavras. De um modo geral, quanto mais incerto ou aleatório for um evento, mais informação ele irá conter, defende Shannon (1948) ao desenvolver a Teoria da Informação. Assim, o autor do artigo "*A Mathematical Theory of Communication*" definiu o conceito de “Entropia”, como uma medida da quantidade de informação existente num evento.

Considerando a entropia como mais um critério de correção de palavras, devem ser definidos os seguintes elementos:

A B C T

- A – Número de soluções encontradas à menor distância possível;
- B – Número de soluções encontradas entre a menor distância possível e a menor distância possível mais 10% do tamanho da palavra;

- C – Número de soluções encontradas entre menor distância possível mais 10% do tamanho da palavra e a menor distância possível mais 20% do tamanho da palavra;
- T – Total de soluções encontradas (soma de A com B e C).

Ao realizar um estudo do valor de cada um dos elementos é possível definir qual o valor máximo que queremos que seja devolvido. Por exemplo, para uma palavra P que está a ser comparada com um dicionário D, podemos obter os seguintes valores: A2 B0 C0 T2, isto significa que foram encontradas duas possíveis soluções para a correção de P à menor distância de caracteres possível. Com base nestes valores, é possível definir o modelo de modo a que apenas seja efetuada a correção quando o valor de A for igual a um, deste modo, temos uma certeza absoluta de que a correção está correta.

Tal como Kukish (1992) defende, devem ser aliados, em paralelo ao NLP outras regras de processamento da informação e limpeza de dados com o intuito de resolver ambiguidades e de apresentar resultados mais certos.

2.3. Análise de dados

Posterior à fase de recuperação e limpeza dos dados, segue-se a fase de análise dos dados. *Machine Learning*, ML, é um ramo da inteligência artificial que, de acordo Larsson & Segerås (2016), se baseia no conhecimento que as máquinas podem adquirir sobre um contexto específico e, mais tarde, usar esse mesmo conhecimento para efetuar tomadas de decisões sem que seja necessária intervenção humana.

Kulkarni (2012) divide os problemas de ML em dois tipos, dependendo se necessitam ou não de dados de treino. Caso sejam utilizados dados de treino, isto é, caso haja processamento dos dados por parte da máquina de forma a rotulá-los, estamos perante o tipo de ML, Aprendizagem Supervisionada; caso não existam dados de treino, nem dados rotulados, o problema de ML insere-se no tipo Aprendizagem Não Supervisionada.

A classificação de texto é uma tarefa importante e comum quando se fala em análise de dados e é um dos tipos de Aprendizagem Supervisionada mais frequentes (Bužić & Dobša, 2018). Existem diferentes métodos para resolver este tipo de problema de ML, sendo que a cada método está a associado um algoritmo diferente.

Ao realizar a escolha do tipo de algoritmo para criar o modelo de treino, é importante avaliar qual o objetivo a alcançar, de forma a obter os melhores resultados possíveis. Para problemas de classificação, Rish (2001) afirma que o algoritmo de Naïve Bayes assume que as classes e os atributos são independentes, isto pressupõe que os parâmetros para cada atributo podem ser aprendidos separadamente, o que simplifica bastante o modelo, especialmente quando o número de atributos é elevado. Para o caso específico do projeto em questão, a classificação de documentos é um domínio com um grande número de atributos, uma vez que os atributos são as diferentes palavras existentes. O modelo Multinomial de Naïve Bayes tem em conta o número de ocorrências de cada palavra, registando-as num documento, e assume que um atributo, ou evento, num documento é independente do contexto e da posição da palavra no

documento (McCallum, 1998). Estes pressupostos baseiam-se no uso simples da regra de Bayes (Bužić & Dobša, 2018):

$$P(c|d) = \frac{P(d|c)\,P(c)}{P(d)}$$

Onde,

- c – classe;
- d – documento;
- P(c) – probabilidade da classe;
- P(d) – probabilidade do documento;
- P(d|c) – probabilidade condicional da classe c para um dado documento d;
- P(c|d) – probabilidade condicional do documento d pertencer a uma dada classe c.

O algoritmo Multinomial de Naïve Bayes, apesar de ser um modelo simples que atinge bons resultados, não é muito eficiente para classes minoritárias, ou seja, neste caso para palavras que ocorrem menos vezes. Este acontecimento deve-se ao facto de existirem poucos dados de treino para certas classes, o que dificulta a aprendizagem do classificador e, seguidamente, resulta numa classificação incorreta dos exemplos dessas classes (Pérez-Ortiz et al., 2015). Desta forma, sabendo que existem classes com grandes discrepâncias a nível de número de ocorrências é necessário utilizar métodos de amostragem para realizar o balanceamento das classes, como é o caso do O*versampling*. Este método adiciona exemplos às classes, durante a fase de treino do modelo, de

forma aleatória, equilibrando o número de ocorrências de todas as classes, não existindo assim discrepâncias entre elas (Dattagupta, 2017).

Após a criação de um modelo de ML, é necessário medir o seu desempenho e nível de performance, para tal, o primeiro passo para realizar uma análise de resultados ao modelo é criar uma Matriz de Confusão (Tabela I) uma matriz que apresenta as quatro combinações de valores seguintes (Sunasra, 2017):

- Verdadeiros Positivos, TP – Valores para os quais tanto a previsão como o output final estão corretos, ou seja, assumem ambos o valor X;
- Verdadeiros Negativos, TN – Quando a previsão diz que o valor não é X e realmente não é;
- Falsos Positivos, FP – Casos onde a previsão é diferente do output final, isto é, a previsão é X e o verdadeiro output é Y;
- Falsos Negativos, FN – Quando a previsão não é X e, na verdade, é X.

Tabela 1 - Confusion Matrix.

		Actual	
		Positives	Negatives
Predicted	Positives	**TP**	**FP**
	Negatives	**FN**	**TN**

Após serem calculados estes valores, podem ser aplicadas as medidas de performance ao modelo de classificação, anteriormente criado. Segundo Feigenbaum (2016) e Bužić & Dobša (2018), os procedimentos de correspondência devem aspirar os seguintes critérios de desempenho:

- Eficiência – Medida que representa a taxa de valores verdadeiramente positivos (TPR), ou seja, matematicamente é a divisão dos valores verdadeiramente positivos sobre o total de positivos previstos, os verdadeiros e falsos positivos. Esta é uma boa medida para avaliar o custo de existirem muitos falsos positivos, uma vez que a sua existência significa que estão a ser identificadas respostas erradas do sistema como corretas.

$$TPR = \frac{TP}{(TP + FP)}$$

- Exatidão – A exatidão do modelo, em inglês *Accuracy*, é a proporção de todos os casos classificados como corretos (verdadeiros positivos e verdadeiros negativos) sobre todos os casos possíveis (verdadeiros positivos, verdadeiros negativos, falsos negativos e falsos positivos). Quando mais perto do valor 1, mais exato é o modelo.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

- *Recall* – Esta medida de desempenho representa a taxa de valores realmente positivos, ou seja, é a proporção de casos positivos que foram corretamente reconhecidos como positivos sobre todos os casos reais. Esta métrica deve ser utilizada quando existe um alto custo associado aos valores falsos negativos, os valores corretos previstos como errados.

$$Positivos\ Reais = \frac{TP}{(TP + FN)}$$

Bužić & Dobša (2018) salientam, ainda, que as medidas não devem ser utilizadas individualmente, mas sim em conjunto pois são complementares e proporcionam um maior nível de informação para decidir se o modelo é satisfatório, se pode sofrer melhorias, ou, até mesmo, se deve ser descartado (Alotaibi et al., 2017).

3. Execução do Projeto

O projeto iniciou-se com a receção dos questionários em papel, num escritório, no país em estudo. Foi contratada uma equipa com o objetivo de retirar uma amostra do universo total dos inquéritos e, posteriormente, digitalizá-los de modo a que chegassem aos nossos escritórios, mais propriamente à equipa responsável pelo projeto, na qual eu me insero.

Para começarmos a trabalhar na análise das respostas, era necessário tê-las em formato digital. Neste sentido, a primeira técnica utilizada foi o Reconhecimento Ótico de Caracteres, do Inglês, *Optical Character Recognition*, OCR, (Kurzweil, 1970), um sistema que reconhece caracteres a partir de texto manualmente escrito e que permite a identificação de diversas fontes e estilos de escrita.

O processo de reconhecimento de caracteres, através das técnicas de OCR, passou pelas seguintes etapas:

1. Digitalização das respostas dos questionários, escritas em formato um caracter por quadrícula, o que facilita o reconhecimento ótico dos caracteres, através de scanner, este processo é denominado “Aquisição de Imagem”;
2. Conversão de todas as imagens das palavras no mesmo formato, .jpg;
3. Binarização das imagens, isto é, o processo no qual os pixéis da imagem são separados em dois grupos, o fundo a branco e o caracter a preto. Este processo é uma maneira simples de separar o texto do fundo, ou seja, retira as quadrículas, deixando apenas os caracteres de texto.

O output final após o processo do OCR, foram as respostas ao questionário, em formato texto, .txt, separadas por “pipes”, “|”, de forma a facilitar a passagem para formato Excel. Seguidamente, foi efetuada uma divisão dos dados, sendo estes agrupados por pergunta do questionário, para posterior análise.

Na fase de análise de dados, começámos por criar um dicionário com a máxima variedade possível de palavras do idioma falado no país onde foi realizado o projeto, e com algumas palavras características do país em questão. O dicionário foi criado com base em dicionários online, de forma a ficar o mais completo possível, uma vez que o seu objetivo era a recuperação das respostas aos questionários. Desta forma, iniciou-se o processo de limpeza de dados, que teve duas fases, na primeira realizou-se uma análise automática e, na segunda fase, uma análise manual.

Na primeira fase, iniciou-se o processo de classificação de texto, recorrendo-se ao NLP e à aplicação em simultâneo de algoritmos para a validação de palavras. Aplicaram-se regras de validação automática para a comparação de palavras e expressões, a distância de *Levenshtein* (Levenshtein, 1966) e a similaridade de *Jaccard* (Jaccard, 1901), indicadores de semelhança entre palavras. Tal como estes dois indicadores, existem muitos outros, contudo após uma análise à comparação de resultados obtidos com cada indicador, estes demonstraram um melhor e mais elevado desempenho (ver Tabela 2, discussão de resultados).

Após várias tentativas de correção de texto através da semelhança entre palavras com os indicadores escolhidos, chegou-se à conclusão que para

palavras pequenas a distância de *Levenshtein* não alcança resultados muito positivos, por exemplo, definir um valor máximo de duas edições entre palavras, pode ser bom para palavras de tamanho superior a seis caracteres, contudo pode ser um valor muito elevado caso a palavra tenha apenas três ou quatro caracteres. Verificou-se assim, que não era possível chegar de imediato a um consenso para definir o valor máximo de edições para uma mais elevada percentagem de palavras corretas. Posteriormente, surgiu a ideia de tornar a distância de *Levenshtein* em valor percentual, ou seja, ao invés de utilizar um número inteiro como valor máximo de edições, utilizou-se uma percentagem baseada no tamanho da palavra. Desta forma, adequou-se a distância entre as palavras ao tamanho das mesmas.

Repetiu-se o processo diversas vezes, de forma a tentar perceber qual o valor percentual ideal, baseado no número de caracteres das palavras, que mais se adequava e com o qual se conseguia obter melhores resultados. Chegou-se à conclusão de que esse valor seria de 0,2, ou seja, 20% do número de caracteres total que constitui a palavra. Por exemplo, tendo em conta a palavra “agricultor”, cujo número de caracteres perfaz um total de 10, o número de edições máximas possíveis para esta palavra, tendo em conta a distância de *Levenshtein* percentual de 20%, é de dois caracteres. Atentando na palavra “agricultores”, adição de “es” à palavra “agricultor”, diz-se que “agricultores” se encontra a uma distância de *Levenshtein* de 2, de “agricultor”.

A similaridade de *Jaccard*, veio acrescentar mais certezas à correção das palavras. Através do estudo comparativo do desempenho de algoritmos verificou-se que esta era a medida com tempo mais rápido de processamento,

acrescentava apenas 2,3 minutos ao tempo de execução do código, como tal adicionou-se ao modelo como segunda medida utilizada para medir distância entre palavras, para casos onde a distância de *Levenshtein* não fosse suficiente para a correção das palavras.

Após a análise realizada através de dicionário utilizando medidas de distância, na qual foram recuperadas as palavras que apresentaram maiores índices de semelhança e menor distância entre as mesmas e as palavras do dicionário, deparámo-nos com uma grande quantidade de dados ainda por corrigir. Parte das respostas não foram recuperadas através dos métodos anteriormente explicados, por apresentarem distâncias entre as palavras maiores que as definidas para a correção das mesmas.

Posto isto, utilizou-se a entropia (Shannon, 1948) como uma métrica de qualidade dos dados para auxiliar na tomada de decisão de qual a expressão que deve ser devolvida pelo modelo de NLP. Ao introduzir a função no modelo, para casos onde a correção da palavra resultasse em mais do que uma opção, a função não retornava nenhuma das respostas consideradas corretas, isto é, se a correção da palavra não fosse única e clara de acordo com as medidas de distância em cima estipuladas era devolvido o resultado “NAN”, representativo de “Não existe correção”. Em suma, esta metodologia, por um lado, apresenta elevados níveis de exatidão, porém deixa sem correção uma vasta percentagem das respostas, mas traduzindo-se numa não contaminação dos dados, evitando assim o enviesamento destes.

Numa fase seguinte, procedeu-se a uma análise manual das respostas, de forma a tentar recuperar o máximo de respostas possíveis.

Para a análise manual, foi realizada uma partição dos dados, ou seja, as respostas que não se corrigiram automaticamente através do modelo de NLP, foram divididas em ficheiros de formato .csv de acordo com o número da pergunta a que pertenciam, e manualmente sofreram uma correção linguística. No entanto, nem todas as respostas foram recuperadas, algumas por serem impercetíveis, outras por não terem sentido, nem contexto e, ainda, por problemas do OCR, que pode ter falhado no reconhecimento dos caracteres. Tudo isto originou a perda de uma elevada quantidade de dados para sempre irrecuperáveis.

Finalizando a fase de correção dos dados, estes foram divididos por pergunta do questionário em formato de texto e iniciou-se a fase de controlo de qualidade. Nesta etapa verificaram-se alguns constrangimentos, a quantidade elevada de outputs provindos do NLP para análise, o facto da validação dos dados ser apenas focada numa amostra aleatória e não na população total e, ainda, o elevado consumo de tempo que leva a realizar um controlo de qualidade.

Após termos os dados corrigidos e organizados por pergunta, seguiu-se a fase de análise dos dados e a sua classificação. Nesta fase, surgiu a necessidade de criar um modelo de classificação de *Machine Learning* de forma a categorizar certas respostas. Modelo esse que seguiu os seguintes passos:

1. Identificação do problema – (1) Classificar as respostas à pergunta referente à atividade profissional, de forma, a agrupá-las por CAE (Código de Atividade Económica); (2) Atribuição do CAE às empresas ou às entidades patronais de cada inquirido. O primeiro problema

surgiu, devido à variedade de respostas encontradas para definir a mesma profissão, já o segundo surgiu uma vez que muitas vezes o CAE da empresa não é o mesmo do funcionário. Por exemplo, uma empresa de consultoria informática, com CAE = X, emprega também seguranças, cujo CAE = Y, e funcionários de limpeza, com CAE = Z, como tal existiu, também, essa necessidade de classificação.

2. Identificação do tipo de problema de *Machine Learning* – Tal como a necessidade indica estamos perante um problema de classificação, que se enquadra na Aprendizagem Supervisionada de ML;
3. Selecionar o algoritmo mais indicado para o tipo problema em questão – Este foi considerado o passo mais complicado da construção do modelo, uma vez que não é fácil decidir qual o melhor algoritmo, tudo depende do tipo de dados e do tipo de problema. Ao consultarmos a biblioteca *Scikit-Learn*, uma biblioteca de código desenvolvido em *Python* para a criação de modelos de *Machine Learning*, encontrámos um esquema que serviu como guia para ajudar na decisão de qual o melhor algoritmo de acordo com o nosso tipo de dados.

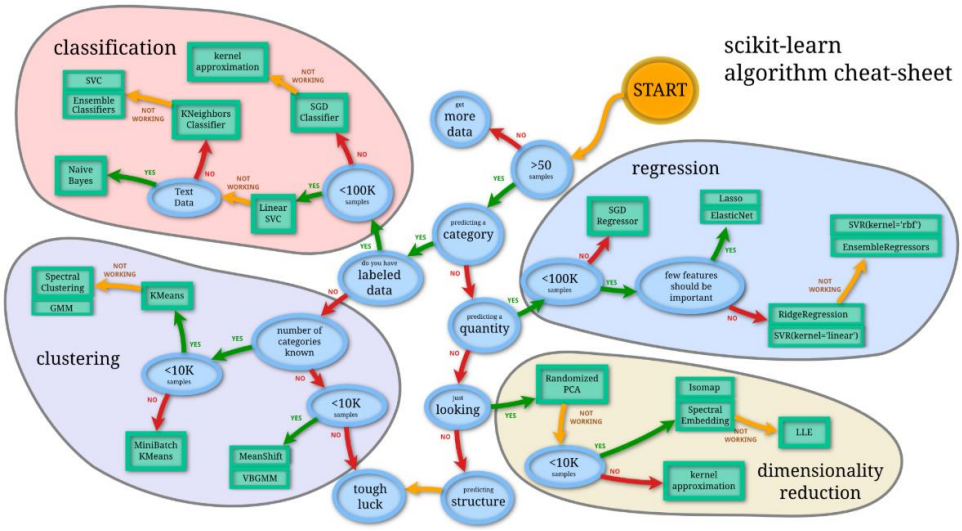

Figura 1 - Escolha do algoritmo

Seguindo o esquema, quando a dimensão da amostra é superior a 50 observações, preveem-se categorias com dados legendados e, posteriormente, é novamente avaliado o número de observações e sendo este inferior a 100 000, o primeiro algoritmo sugerido é o Linear SVC, porém, sendo dados de texto, o algoritmo de Naïve Bayes é o mais indicado. Foi, também, realizado um estudo comparativo dos algoritmos de classificação, e, de facto, verificou-se que o algoritmo de Naïve Bayes era o que alcançava melhores resultados.

4. Criação do Modelo – Para efetuar o treino do modelo, os dados da amostra foram divididos em 70% para treino e 30% para teste, foi, também, necessário garantir que as amostras do conjunto de teste seriam diferentes do conjunto de treino, definindo um *data frame* temporário com os dados de treino. Ainda no treino do modelo, foram contadas o número total de ocorrências de cada classe, de forma a saber qual a probabilidade de cada uma. Como as probabilidades

eram muito díspares, realizou-se o processo de *oversampling*, uma técnica utilizada para equilibrar as probabilidades das classes, ou seja, de ajustar a sua distribuição replicando as amostras. Tendo como base a classe com o maior número de ocorrências, o ajuste é feito em todas as outras classes ficando assim com o mesmo número de ocorrências da classe tida como base. Deste modo, o modelo não foi influenciado pelo número de ocorrências, dado que todas as classes apresentavam a mesmo probabilidade. Em seguida, definiram-se as entradas, ou seja, os atributos convertidos em valores numéricos através da vectorização e invocou-se a função Multinomial de Naïve Bayes.

Em suma, o modelo foi criado com os dados de treino, 70% dos dados, e a sua performance foi posteriormente avaliada com os dados de teste, os restantes 30% dos dados.

5. Avaliação de resultados – Depois do modelo criado, foi necessário medir o seu desempenho e avaliar os resultados com ele obtidos. Para tal, foram extraídas duas tabelas do modelo, uma de resultados corretos e outra de resultados incorretos, ou seja, na primeira tabela a previsão corresponde à correção e na segunda o contrário. Com estes dados, foram aplicadas medidas de performance de forma a indicarem-nos se o modelo estava a obter resultados satisfatórios, ou se necessitava de melhorias.

   Para a aplicação das medidas de performance, foi necessário criar a matriz de confusão, que contem as decisões tomadas pelo classificador. A matriz apresenta quatro tipos de resultados:

- Verdadeiros Positivos (TP);
- Verdadeiros Negativos (TN);
- Falsos Positivos (FP);
- Falsos Negativos (FN).

Após obtermos estes resultados, aplicámos então as medidas de desempenho, eficiência, exatidão e *recall* ao modelo.

6. Predição – Esta foi a última fase, dado que o modelo de teste apresentou uma boa performance, com uma eficiência de cerca de 95%, utilizámo-lo para prever os CAE tanto das profissões dos inquiridos, como das empresas das quais faziam parte e, desta forma, conseguimos classificar os dados que necessitávamos.

Após a realização do modelo de *Machine Learning*, atingimos o nosso objetivo final, ter dados disponíveis, classificados e de qualidade para a realização de análises estatísticas e disseminação de resultados, avaliando indicadores referentes tanto ao nível de vida da população como às condições do país.

4. Discussão de Resultados

Neste capítulo, procedemos à discussão dos resultados obtidos de acordo com as metodologias utilizadas, para tal, foi desenhado um esquema que representa todas as fases do projeto.

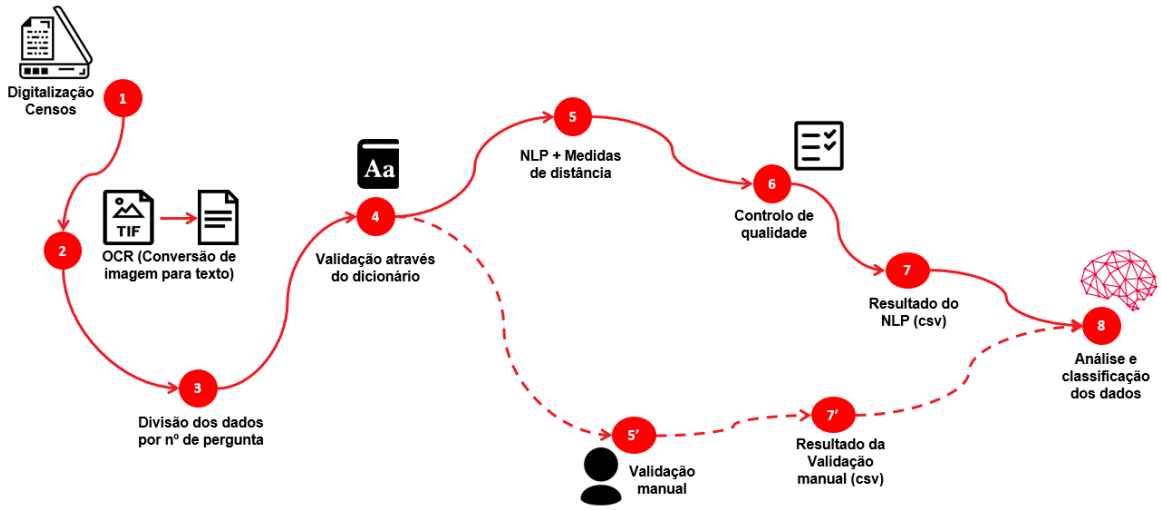

Figura 2 - Fluxo do projeto.

A fase 1 do esquema representa o processo de digitalização dos inquéritos populacionais. Sendo mais de vinte milhões de inquéritos, não seria viável digitalizá-los na totalidade, como tal extraiu-se uma amostra de cerca de 0,6% da população. Apesar de a amostra ser aleatória, perdem-se muitos dados ao realizar Censos manuscritos, em vez de Censos em formato digital, pois a análise cobre um parte muito pequena da população.

Na segunda fase, realizou-se a recolha de dados e utilizou-se uma tecnologia de reconhecimento de caracteres manuscritos a partir de um arquivo de imagem, *Optical Caracter Recognition*, OCR. De uma forma sucinta, o OCR

permitiu realizar o processo de aquisição de imagem, ou seja, permitiu o reconhecimento das respostas presentes no inquérito e apresentou-as em formato texto, separadas por *pipes*, o que permitiu, facilmente, o seu transporte para formato Excel de modo a serem divididas por número de pergunta. Importa realçar que, mais uma vez, muitos dados foram perdidos neste processo, devido a erros de leitura do OCR e má caligrafia do inquirido.

Na fase 4 e 5 procedeu-se à limpeza dos dados e recuperação de expressões, através da criação de um dicionário e da aplicação do Processamento de Linguagem Natural, NLP, e, em simultâneo, de algoritmos de medida de distância entre palavras.

Para o processo de classificação de texto e recuperação de expressões, existem diferentes tipos de abordagens para a comparação de palavras e expressões, como tal, foi necessário realizar um estudo dos algoritmos de medida de similaridade entre palavras de forma a perceber aquele que melhor se adaptava ao caso, isto é, o algoritmo que melhor conciliava o número de respostas corrigidas com o tempo de processamento. Para tal, foi selecionada uma parte da base de dados que continha as respostas aos censos e, para a mesma amostra, adicionou-se ao modelo de NLP, o algoritmo de cada medida de distância em separado, de forma a realizar, posteriormente, uma comparação de desempenho e tempo de processamento.

Deste modo, obtiveram-se os seguintes resultados:

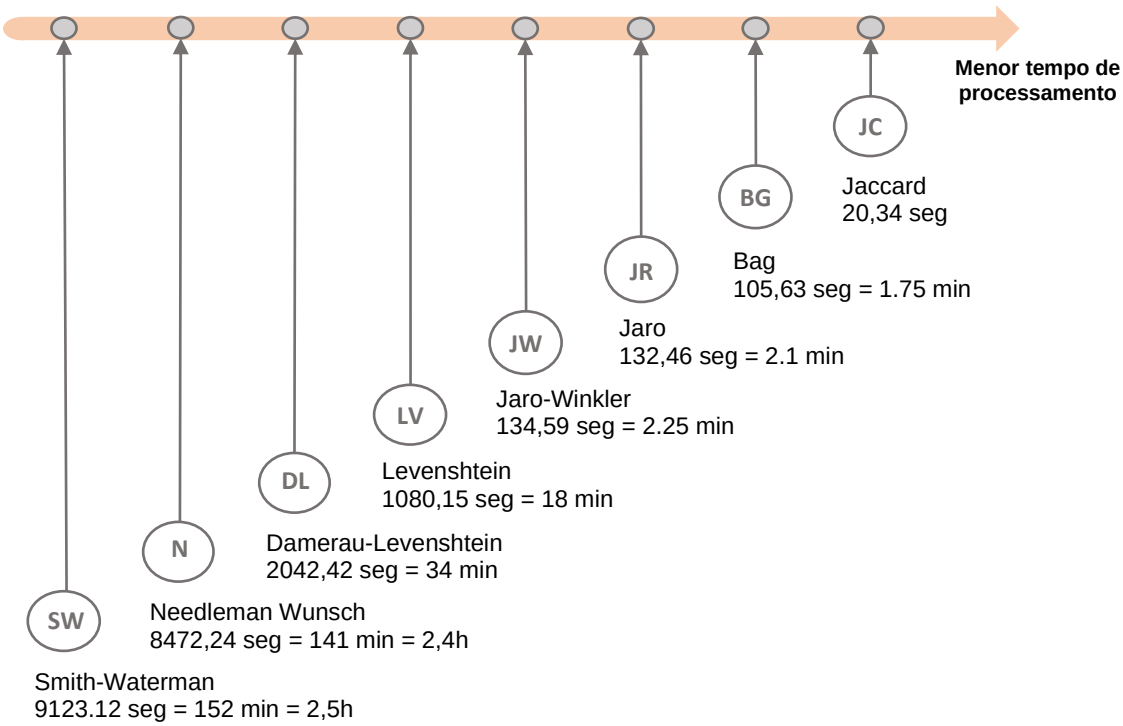

Figura 3 - Tempos de Processamento dos Algoritmos de Distância

Tabela 2 - Comparação das distâncias de correção algorítmica.

OCR Capture	Levenshtein		Damerau-Levenshtein		Needleman-Wunsch		Smith-Waterman		Jaro-Winkler		Jaro		Bag		Jaccard	
1. CERA LICA	SERRA LEOA	4,00	SERRA LEOA	4,00	SRI LANCA	0,33	ARGELIA	0,21	CONGOLESA	6,00	AFRICA	5,00	COSTA RICA	5,00	SRI LANCA	2,00
2. ?WANDA	UGANDA	2,00	UGANDA	2,00	CANADA	0,50	TSWANA	0,22	RUANDA	2,00	LUANDA	2,00	LUANDA	2,00	CANADA	2,00
3. B?RR?N?E	BURUNDI	5,00	BURUNDI	5,00	BENIN	0,50	BURUNDES	0,22	BURKINA	5,00	ARMENIA	6,00	ARMENIA	6,00	BURUNDES	4,00
4. ?ONAO	CONGO	2,00	CONGO	2,00	ANGOLA	0,50	MONACO	0,18	CONGOLESA	2,00	CONGO	2,00	CONGO	2,00	CONGO	2,00
5. AUAN?A	RUANDA	2,00	RUANDA	2,00	RUANDA	0,50	TAIUANESA	0,20	RUANDA	2,00	LUANDA	2,00	LUANDA	2,00	CANADA	2,00
6. EST?DO LNIIJ?	ESTADOS UNIDOS	6,00	ESTADOS UNIDOS	6,00	ESTADOS UNIDOS	0,38	ESTONIA	0,23	ESTADOS UNIDOS	8,00	ESTADOS UNIDOS	7,00	ESTADOS UNIDOS	7,00	TIMOR LESTE	6,00

Figura 4 - Exemplos imagem OCR.

Com base nos critérios anteriormente definidos e nos resultados obtidos de desempenho (tabela II) e tempo de processamento (figura 3), definiu-se que se incorporaria no modelo a distância de *Levenshtein* e a distância de *Jaccard*. A primeira porque é a medida com a qual se obteve melhores resultados em termos de desempenho, e a segunda, uma vez que foi a que apresentou um tempo de processamento mais baixo. Apesar de a distância de *Jaccard* ser a medida que apresenta valores de desempenho mais baixos, destaca-se pelo seu baixo tempo de processamento. Tomou-se a decisão de acrescentá-la ao modelo, dado que não interferiu com o tempo de processamento e, demonstrou ser uma mais valia na correção de palavras e expressões, caso a distância de *Levenshtein* não resultasse.

Estas duas fases foram essenciais para tornar os dados disponíveis em informação válida e de qualidade, contudo, mais uma vez, houve uma grande perda de dados, pois não foi possível encontrar uma correção exata de muitas das respostas.

Se por um lado o controlo de qualidade da correção dos dados foi uma tarefa fácil de realizar, uma vez que os dados apenas eram corrigidos caso a distância entre as respostas e o dicionário fosse muito diminuta, por outro lado, restaram muitas respostas para correção manual, fase 5', um processo que exigiu um grande esforço devido à quantidade de respostas que ficaram por corrigir e ao tempo o que demora a fazê-lo.

Numa última etapa, fase 8, na qual as respostas já se encontravam corrigidas e agrupadas por pergunta, em formato .cvs, procedeu-se à fase de análise. Ao iniciarmos esta fase, verificou-se que para analisar certos

indicadores, como o número de população empregada, os setores de atividade e a evolução do emprego no país, seria necessário efetuar uma classificação dessas mesmas respostas. Para tal, criou-se um modelo de classificação de *Machine Learning*, baseado no algoritmo de Naïve Bayes, que permitiu agrupar as respostas tanto da profissão do inquirido, como do setor da empresa para a qual trabalhava, por CAE, de forma a facilitar a análise.

Existem vários algoritmos que dão resposta a problemas de classificação de ML, como tal foi, novamente, necessário efetuar um estudo comparativo de forma a decidir qual o melhor algoritmo para o tipo de dados do projeto. A estratégia utilizada foi a mesma dos algoritmos de distância de similaridade de palavras, efetuaram-se testes ao modelo com os diferentes algoritmos, de forma a perceber qual obtinha melhores resultados e melhor performance a nível de tempo.

Como previsto pela biblioteca *Sickit-Learn*, o algoritmo de *Naïve Bayes* foi o que obteve melhores resultados, tanto a nível de desempenho, com uma percentagem de eficiência de 95%, como a nível de performance temporal, demorando cerca de uma hora ser executado.

Esta foi uma tarefa que apesar de demorada na sua realização, tornou possível uma melhor e mais eficaz análise da população.

5. Conclusões, Limitações e Trabalhos Futuros

Tendo como objeto de estudo o inquérito dos Censos Populacionais e Habitacionais aplicados à população de um determinado país, o objetivo deste projeto consistiu na preparação de dados e, posterior, disseminação de resultados. Desta forma, o foco deste Trabalho Final de Mestrado está nas metodologias aplicadas ao longo de todo o processo de análise e tratamento de dados, uma vez que foram utilizadas técnicas de reconhecimento, correção e classificação de dados inovadoras que constituíram assim, uma nova abordagem ao estudo dos Censos Populacionais e Habitacionais.

Após o término do projeto, é importante retirar conclusões e ilações dos constrangimentos encontrados, de forma a que futuramente as dificuldades atuais passem a ser enfrentadas com naturalidade e não sejam um entrave durante a realização do trabalho.

Deste modo, é de realçar a importância da fase de recolha dos dados. Esta é das etapas mais relevantes de um projeto de análise de dados, é a fase na qual recebemos os inputs para o nosso trabalho, devendo estes ser o mais adequados e úteis possível, de forma a retirar o máximo proveito deles e obter os melhores outputs. Neste sentido, teria sido proveitoso apostar na recolha de dados em formato digital ao invés de documentos manuscritos em papel, pois como foi percetível neste projeto, estes últimos são difíceis de interpretar e carecem de uma maior análise, devido às variações da caligrafia de cada individuo, dos erros ortográficos que podem vir a ser cometidos ou, até mesmo, de erros de repostas que podem ter que vir a ser riscadas e reescritas. Tudo isto

são motivos que podem estar por de trás de uma grande perda de dados e de informação, que mais tarde, pode vir a ser refletida negativamente no resultado final.

Para além das dificuldades na interpretação dos dados manuscritos recolhidos através dos inquéritos Censos, grande parte do tempo despendido neste projeto foi a recuperar e a corrigir as respostas, sendo que este não era o objetivo principal do projeto, mas sim a análise dos indicadores sociais através das respostas da população. Caso as respostas fossem recolhidas em formato digital, esta seria uma tarefa que facilmente se resolveria, efetuando uma correção em tempo real dos dados.

No decorrer do projeto surgiu a necessidade de aplicar algoritmos de distância entre palavras, de forma a facilitar a recuperação das respostas, tendo sido utilizada distância de Levenshtein e a medida de similaridade de Jaccard, uma vez que após testes de comparação com outros algoritmos estes foram o que apresentaram melhores performance, a nível de correção e tempo de processamento. Esta foi uma forma de acelerar o processo de recuperação e correção de respostas, contudo ao primarmos pela exatidão de resultados, muitos ficaram para posterior correção manual.

Em relação ao modelo de *Machine Learning* criado para classificar o CAE, o algoritmo utilizado foi o de Naïve Bayes, porque mostrou ser o mais indicado para prever categorias com dados de texto legendados.

Em futuros inquéritos, para além do uso de formato digital, através de tablets e computadores, por exemplo, devem-se evitar as perguntas de resposta

aberta e optar por utilizar perguntas com respostas pré-definidas para seleção. Esta medida para além de reduzir a taxa de erro, minimiza o tempo de resposta.

Acredita-se que este projeto contribui para apresentar uma nova abordagem de como dados manuscritos podem ser trabalhados e analisados através de diversas técnicas atuais de reconhecimento de caracteres, recuperação e correção de palavras/expressões e, ainda, de classificação automática de respostas, tudo de forma a facilitar a análise e interpretação de dados. Contudo, é ainda mais, uma lição para o futuro e a prova que as novas tecnologias facilitam, e muito, o trabalho humano.

6. Bibliografia

Alotaibi, F., Abdullah, M., Abdullah, R., Rahmat, R., Hashem, I. & Sangaiah, A. (2017). Optical Character Recognition for Quranic Image Similarity Matching. *IEEE Access* 6, 554 – 562.

Bourezgue, T. (2017). Using tablets for the 2018 Algerian Census: Census data management and quality assessment. *Statistical Journal of the IAOS* 33, 777–784.

Bužić, D. & Dobša, J. (2018). Lyrics Classification using Naive Bayes. 2018 *41st International Convention on Information and Communication Technology, Electronics and Microelectronics*. Croácia.

Cao, H., Subramanian, K., Peng, X., Chen, J., Prasad, R. & Natarajan, P. (2012). Extracting Information from Handwritten Content in Census Forms. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012),* 306 – 309

Chen, P., Liu, Q., Wei, L., Zhao, B., Jia, Y., LV, H. & Fei, X. (2019). Automatically Structuring on Chinese Ultrasound Report of Cerebrovascular Diseases via Natural Language Processing. *IEEE Access* 7, 89043 - 89050.

Damerau, F. (1964). A technique for computer detection and correction of spelling errors. *Magazine Communications of the ACM* 7 (3), 171-176.

Dattagupta, S. (2017). A performance comparison of oversampling methods for data generation in imbalanced learning tasks. Universidade Nova de Lisboa. Portugal.

Dekker, A. (2001). Adapting new technologies to censos operations. Symposium on Global Review of 2000 Round of Population and Housing Censuses: *Mid-Decade Assessment and Future Prospects*, New York: OECD Glossary.

Feigenbaum, J. (2016). A Machine Learning Approach to Census Record Linking. *Scholar at Harvard.*

Gomaa, W. & Fahmy, A. (2013). A Survey of Text Similarity Approaches. *International Journal of Computer Applications* 68 (13), 13-18.

Gomes, A. (2009). Dados das Nações Unidas: Quais são os países menos desenvolvidos?. *Jornal Público Online.*

Haldar, R. & Mukhopadhyay, D. (2001). Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach. *ArXiv Cornell University.*

Hicham, G., Abdallah, Y. & Mustapha, B. (2012). Introduction of the weight edition errors in the Levenshtein distance. *International Journal of Advanced Research in Artificial Intelligence* 1 (5), 30-32.

Instituto Nacional de Estatística (2011). Para que servem?, 2014. Portugal. Disponível em: https://censos.ine.pt/xportal/xmain?xpid=CENSOS&xpgid=censos_pqservem [Acesso em: 2019/3/3].

Keysers, D., Deselaers, T., Rowley, H., Wang, L. & Carbune, V. (2017). Multi-Language Online Handwriting Recognition. *IEEE Transactions on pattern analysis and Machine Intelligence* 39 (6), 1180-1194.

Kukich, K. (1992). Techniques For Automatically Correcting Words In Text. *Journal ACM Computing Surveys (CSUR)* 24 (4), 377-439.

Kulkarni, P. (2012). Reinforcement and Systemic Machine Learning for Decision Making: Wiley IEEE Press.

Larsson, A. & Segerås, T. (2016). Automated invoice handling with machine learning and OCR. Estocolmo.

McCallum, A. & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. *Work Learn Text Categ* 752.

Nanni, L. & Lumini. A. (2008). Generalized Needleman–Wunsch algorithm for the recognition of T-cell epitopes. *Expert Systems with Applications* 35, 1463–1467.

Pérez-Ortiz, M., Gutiérrez, P., Tino, P. & Hervás-Martínez, C. (2015). Over-sampling the minority class in the feature space. *IEEE Transactions on Neural Networks and Learning Systems* 27 (9).

Poole, S., Khajeh-Saeeda, A. & Perrot, J. (2010). Acceleration of the Smith–Waterman algorithm using single and multiple graphics processors. *Journal of Computational Physics* 229, 4247–4258.

Rish, I (2001). An empirical study of the naive Bayes classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence, 3, 41-46. IBM New York.

Sarkar, S., Pakray, P., Das, D., Saha, S., Bentham, J., Gelbukh, A. (2015). Language Independent Paraphrases Detection. *Journal CEUR Workshop Proceedings* 1737, 256-259.

Shannon, C. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal* 27, 379–423, 623–656.

Sunasra, M. (2017). Performance Metrics for Classification problems in Machine Learning. Disponível em: https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b [Acesso em: 2019/9/5].

Uhliarik, I. (2013). Handwritten character recognition using machine learning methods. Bratislava.

United Nations Statistics Division (2019). International Symposium on the use of Big Data for official statistics. Disponível em: https://unstats.un.org/bigdata/events/2019/hangzhou/default.asp [Acesso em: 2019/3/3].

Winkler, W. (1994). Advanced Methods of Record Linkage. *American Statistical Association, Proceedings of the Section of Survey Research Methods*, 467-472.

Xie, Q., Zhou, X., Wang, J., Gao, X., Chen, X. & Liu, C. (2019). Matching Real-World Facilities to Building Information Modeling Data Using Natural Language Processing. *IEEE Access* 7, 119465 – 119475.