INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Análise de prospetos e regulamentos de gestão de produtos de investimento

Tiago de Sousa Alves

Mestrado em Engenharia Informática

Orientador:
Doutor Luís Miguel Martins Nunes, Professor Associado,
Iscte – Instituto Universitário de Lisboa

Co-Orientador:
Doutor Ricardo Daniel Santos Faro Marques Ribeiro, Professor Associado, Iscte – Instituto Universitário de Lisboa

Outubro, 2023

TECHNOLOGY
AND ARCHITECTURE

Departamento de Ciências e Tecnologias da Informação

Análise de prospetos e regulamentos de gestão de produtos de investimento

Tiago de Sousa Alves

Mestrado em Engenharia Informática

Orientador:
Doutor Luís Miguel Martins Nunes, Professor Associado,
Iscte – Instituto Universitário de Lisboa

Co-Orientador:
Doutor Ricardo Daniel Santos Faro Marques Ribeiro, Professor Associado, Iscte – Instituto Universitário de Lisboa

Outubro, 2023

Agradecimentos

Primeiro quero agradecer aos meus orientadores, Professor Luís Nunes e Professor Ricardo Ribeiro por todo o apoio e ajuda. Obrigado por me guiarem nesta jornada com toda a paciência e dedicação.

Em segundo quero agradecer à CMVM - Comissão do Mercado de Valores Mobiliários por me ter apoiado em todas as etapas do processo e por se terem dedicado tanto a este trabalho. A vossa ajuda foi fundamental para a progressão do mesmo.

Quero agradecer também aos meus colegas do ISCTE - Instituto Universitário de Lisboa, em especial ao David Rodrigues que foi uma ajuda muito preciosa durante a elaboração deste trabalho.

Por fim, um obrigado muito grande à minha família. Quero agradecer aos meus pais e à minha irmã por todas as palavras de apoio e força que me deram e um agradecimento muito especial à minha namorada, Ana, que foi incansável durante todo este processo e que muita força me deu para continuar. A todos um muito obrigado.

Tiago de Sousa Alves

Financiamento

Este trabalho foi parcialmente apoiado pela Fundação para a Ciência e a Tecnologia, I.P. (FCT) [Projectos ISTAR: UIDB/04466/2020 e UIDP/04466/2020].

Resumo

A análise de documentos com recurso a técnicas de PLN (Processamento de Língua Natural) é um tema em grande desenvolvimento. Neste trabalho desenvolvemos um protótipo que recorre a algumas técnicas de PLN com o objetivo de ajudar a CMVM - Comissão do Mercado de Valores Mobiliários a validar prospetos financeiros que são documentos jurídicos emitidos por empresas destinados aos investidores sobre os valores mobiliários que estas emitem e sobre si próprias.

O trabalho consiste em primeiro extrair a informação dos documentos mantendo a estrutura dos mesmos. Para isso recorremos ao índice que nos dá toda a informação necessária dos capítulos e respetivas páginas.

De seguida foram desenvolvidas algumas regras com base no Regulamento (UE) 2017 1129 para serem aplicadas à secção do sumário. Estas regras estão divididas em três grupos: REGEX's, Reconhecimento de Entidades Mencionadas e cálculo de similaridade de textos. Os resultados obtidos foram positivos ainda que tenham sido detetadas algumas melhorias preciosas que se podem aplicar futuramente.

Por fim são aplicas as regras desenvolvidas às devidas secções e avaliados os resultados obtidos através de um conjunto de documentos previamente escolhidos por nós, onde é feita uma comparação entre a validação automática e a validação manual. Nesta fase testamos desde a extração do índice, passando pela identificação e extração do sumário e de seguida aplicação das regras ao mesmo. Os resultados são animadores, contudo percebemos que o facto de não haver uma normalização dos documentos cria algumas dificuldades no processo de validação automática.

Palavras Chave: *PLN*,*Documentos Legais*, *Extração de informação NER*

Abstract

The analysis of documents using NLP (Natural Language Processing) techniques is a hot topic. In this work we have developed a prototype that uses some NLP techniques with the aim of helping the CMVM - Comissão do Mercado de Valores Mobiliários to validate financial prospectuses, which are legal documents issued by companies to investors about the securities they issue and about themselves.

The work consists of first extracting the information from the documents while maintaining their structure. To do this, we used the table of contents, which gives us all the information we need about the chapters and their respective pages.

Next, we developed some rules based on Regulation (UE) 2017 1129 to be applied to the summary section. These rules are divided into three groups: REGEX's, Named Entety Recognition and text similarity calculation. The results obtained were positive, although some valuable improvements were detected that could be applied in the future.

Finally, the rules developed are applied to the appropriate sections and the results obtained are evaluated using a set of documents previously chosen by us, where a comparison is made between automatic validation and manual validation. At this stage we tested everything from extracting the table of contents to identifying and extracting the summary and then applying the rules to it. The results are encouraging, but we can see that the fact that the documents are not standardized creates some difficulties in the automatic validation process.

KEYWORDS: *NLP*, *Legal documents*, *Information extraction*, *NER*

Conteúdo

Agradecimentos i

Financiamento iii

Resumo v

Abstract vii

Lista de Figuras xi

Lista de Tabelas xiii

Capítulo 1. Introdução 1
1.1. Descrição do problema 1
1.2. Motivação 2
1.3. Questões de investigação 3
1.4. Estrutura 3

Capítulo 2. Revisão da literatura 5
2.1. Revisão sistemática 5
2.2. Extração de informação de textos legais 6
2.3. Utilização de ontologias na extração de informação 9
2.4. Sumário 11

Capítulo 3. Análise dos dados 13
3.1. Dados cedidos pela CMVM 13
3.2. Regulamento (UE) 2017/1129 15
3.3. Regras analisadas 15

Capítulo 4. Metodologia proposta 17
4.1. Extração da estrutura do documento 18
4.1.1. Deteção do índice 18
4.1.2. Extração do índice 20
4.1.3. Validação dos títulos extraídos do índice 21
4.1.4. Extração da estrutura do sumário 24
4.2. Criação das regras com base no regulamento 25
4.2.1. Reconhecimento de padrões através de expressões regulares 26
4.2.2. Reconhecimento de entidades nomeadas 27
4.2.3. Localização de excertos no texto através de similaridade 28

4.2.4. Regras que não obtiveram resultados positivos 30
4.3. Aplicação das regras às devidas secções 30

Capítulo 5. Avaliação da validação do prospeto 33
5.1. Avaliação das regras desenvolvidas 33
5.2. Avaliação da extração da estrutura de um prospeto 35

Capítulo 6. Conclusões e trabalho futuro 37

Bibliografia 39

Appendix A. Anexos 43

Lista de Figuras

2.1 Distribuição dos métodos de investigação em IE de documentos legais segundo [3] 7
2.2 Distribuição das tarefas de investigação em IE de documentos legais segundo [3] 8
2.3 Métricas da extração de entidades nomeadas consigas em [7] 8
2.4 Arquitetura do ICIX [9] 9
2.5 Ontologia proposta por [13] 10

3.1 Quantidade de prospetos emitidos em cada língua do longo dos diversos anos 14
3.2 Quantidade de páginas por prospeto 14

4.1 Esquema dos processos desenvolvidos nas diferentes etapas da metodologia 17

A.1Exemplo de índice com subsecções 43
A.2Exemplo de índice apenas com secções principais 44
A.3Exemplo de índice sem separador entre título e página correspondente 45
A.4Exemplo sumário em tabela com secções identificadas 46
A.5Exemplo sumário em modo texto 47

Lista de Tabelas

2.1 Filtros utilizados na pesquisa de trabalhos relacionados 6
2.2 Quantidade de documentos após processo de filtragem do trabalho relacionado 6
2.3 Sumário dos artigos revistos nesta secção 12

3.1 Quantidade de palavras por prospeto contabilizando todas as ocurrências e apenas uma ocurrencia de cada palavra 15

4.1 Deteção do índice: comparação entre as páginas esperadas e páginas detetadas pelo algoritmo 19
4.2 Comparação entre o número de pares [Título,pág] extraídos e o número real para cada documento 22
4.3 Comparação entre validação manual e automática se o título se encontra na página indicada pelo índice para cada documento 23

5.1 Métricas obtidas através do teste das regras 33
5.2 Resultado da avaliação da extração da estrutura de um prospeto 35

CAPÍTULO 1

Introdução

A inteligência artificial (IA) tem cada vez mais impacto no dia a dia das pessoas, e é cada vez mais utilizada, mesmo onde muitas vezes não se sabe que a estamos a utilizar. Se por um lado é relativamente intuitivo entender que um carro que conduz sozinho recorre a IA, quando queremos que o GPS encontre o trajeto mais rápido ou, quando aceitamos as sugestões de palavras do teclado do telefone para escrever uma mensagem já não é tão evidente darmos conta de que a mesma está presente.

A IA é uma excelente ajuda principalmente na otimização de processos, pois tirando partido da grande capacidade de processamento dos computadores, consegue processar e analisar dados muitas vezes mais rápido que um humano, quer sejam dados estruturados, por exemplo guardados numa numa base de dados, quer sejam dados não estruturados como por exemplo textos.

Muitas empresas têm vindo a beneficiar imenso com os avanços da IA no domínio textual, pois fica cada vez mais fácil, por exemplo, analisar a opinião e o sentimento que um certo produto causa no público em geral, ou perceber se a opinião pública em relação a um empresa tende a ser mais positiva ou negativa.

Uma área que pode beneficiar muito com o desenvolvimento da IA no domínio do processamento computacional da língua é a área legal, pois normalmente os documentos legais são muito extensos e demoram muito tempo a serem analisados, por isso, otimizar este processo pode ser bastante benéfico para reduzir esforços e custos.

Este trabalho irá focar-se no campo do processamento de língua natural (PLN), uma área da IA que estuda diferentes abordagens para computadores processarem e interpretarem linguagem falada e escrita por humanos.

1.1. Descrição do problema

O objetivo deste trabalho incidirá na análise de um tipo específico de documentos legais, os prospetos para valores mobiliários, que são documentos emitidos pelas empresas, destinados aos potenciais investidores, com as informações sobre si mesmas e sobre os valores mobiliários que as mesmas emitem. Os prospetos antes de serem emitidos precisam de ser validados pela entidade responsável por tal, e é com o intuito de estudar como dar suporte a essa validação que este trabalho foi desenvolvido.

A análise destes documentos divide-se em três fases:

Análise formal: Verificar se os prospetos cumprem com os requisitos estabelecidos no Regulamento (UE) 2017/1129, requisitos esses que podem ser relativos ao formato do prospeto, às informações específicas ou mínimas a incluir no documento universal, entre outros.

Análise de substância: Verificar se a informação contida nos prospetos está de acordo com a informação que se encontra em fontes públicas como ESMA e CMVM.

Análise de usabilidade: Verificar se a informação é clara, tendo em conta o público-alvo.

Apesar das três etapas de análise de um prospeto, o foco deste trabalho será essencialmente a primeira fase, que é a mais trabalhosa e que consome mais tempo, pois requer que se leia o documento inteiro e se verifique que todos os pontos do regulamento se encontram em conformidade.

Como se pode observar na secção 3 existem vários tipos de regras para serem aplicadas. Neste trabalho, o foco foram as regras que devem ser aplicadas ao sumário, Artigo 7º, uma vez que este tem bastante diversidade de regras e assim se consegue perceber se é possível criar e aplicar regras de diferentes naturezas, observar se os resultados obtidos são positivos e com isso ter uma ideia das dificuldades de validar um prospeto automaticamente.

1.2. Motivação

A análise textual tem sido uma área de grande interesse ao longo dos últimos anos. Todos os dias são trocados milhões de e-mails, enviadas mensagens, celebrados contratos, muitas das vezes de forma textual. Por isso, tem existido um grande investimento por parte das empresas e das universidades para estudarem este tema.

Pôr em prática os conhecimentos adquiridos na área da IA é sempre um desafio, porém é uma mais-valia para as empresas ou instituições que vêem os seus processos a serem otimizados, e contam com uma nova ajuda na execução de tarefas muitas vezes complexas e aborrecidas.

Este projeto foi desenvolvido através de uma parceria entre o Iscte - Instituto Universitário de Lisboa e a CMVM (Comissão do Mercado de Valores Mobiliários), tendo esta última como missão supervisionar e regular os mercados de instrumentos financeiros, assim como os agentes que neles atuam, promovendo a proteção dos investidores.[1]

Neste trabalho, o objetivo é perceber se é possível desenvolver um processo que ajude a diminuir o tempo que a CMVM demora a avaliar um prospeto para valores mobiliários, documento esse que chega a ter por vezes mais de 300 páginas, como se pode observar no Capítulo 3, e onde é preciso analisar cada ponto do Regulamento em vigor. Este processo é demorado e os prazos de validação são curtos, o que dificulta o trabalho de quem fica responsável pelo mesmo, podendo dar origem a erros. Com ajuda da IA é possível que o tempo de análise de um prospeto seja reduzido, e que a quantidade de iterações necessárias para correção de erros diminua, dado que o potencial de disponibilizar aos proponentes

[1] `https://www.cmvm.pt/pt/CMVM/Apresentacao/Pages/Apresentacao-o-que-e-a-CMVM.aspx`

um sistema prévio de verificação automática poderá reduzir a quantidade de erros na primeira versão.

1.3. Questões de investigação

O objetivo desta investigação é perceber e analisar os problemas que ocorrem durante a validação de um prospeto, e a grande questão que envolve este projeto é tentar perceber se é possível criar um processo de validação automática, com o objetivo de diminuir o tempo que demora avaliar um prospeto e diminuir os erros inerentes ao processo de validação.

Neste caso, devido à pequena quantidade de dados disponíveis para desenvolvimento do projeto, a abordagem que pareceu mais adequada foi mapear as condições de aprovação de prospetos e posteriormente avaliar essa implementação.

Para desenvolver o um protótipo que vá ao encontro das expectativas, iremos abordar algumas questões que nos parecem fundamentais:

- Quais as dificuldades na sistematização e codificação das regras indicadas no regulamento que rege a validação de prospetos mobiliários?
- De que forma se consegue extrair o texto dos prospetos mantendo a sua estrutura?
- É possível mapear as regras desenvolvidas para aplicar às devidas secções?
- Qual a eficácia de um sistema de validação de prospetos construído deste modo?

1.4. Estrutura

Este trabalho encontra-se dividido em quatro partes, sendo que na primeira parte, no **Capítulo 2**, é feita uma análise de alguns trabalhos realizados anteriormente que possam trazer algumas dicas úteis para o desenvolvimento deste.

No **Capítulo 3** é feita uma análise pormenorizada ao conjunto de dados (documentos que foram cedidos pela CMVM), onde se tenta prever alguns problemas e dificuldades que os mesmos possam trazer, e também algumas semelhanças com os conjunto de dados utilizados nos trabalhos analisados no capítulo anterior.

Posto isto, no **Capítulo 4** encontra-se uma descrição ao pormenor do que foi desenvolvido ao longo da execução deste trabalho, assim como a explicação de algumas decisões tomadas e qual o objetivo que se pretende atingir com as mesmas.

De seguida, no **Capítulo 5**, os resultados obtidos são avaliados e é feita uma análise crítica dos mesmos, acrescentando algumas melhorias que se acredita que possam ser desenvolvidas no futuro para melhorar ou complementar os resultados, ou colmatar algumas limitações que o mesmo possa ter.

Por fim, no **Capítulo 6**, é feito um resumo do trabalho desenvolvido e dos resultados obtidos. É abordado também algum trabalho futuro com vista a melhorar o que foi desenvolvido, assim como são vistas algumas ideias que não foram desenvolvidas mas que podem acrescentar muito valor. Ainda neste capítulo são respondidas as questões de investigação propostas no início do trabalho.

CAPÍTULO 2

Revisão da literatura

Este trabalho tem como objetivo a validação de vários requisitos estabelecidos no Regulamento (UE) 2017/1129, que visam tanto o formato do prospeto como a informação que deve estar contido no mesmo. Para fazer esta validação é primeiro necessário sistematizar e codificar os requisitos do regulamento em regras para serem aplicadas e de seguida é necessário identificar as partes do texto às quais se devem aplicar as regras e por fim, mapear que regras devem ser aplicadas a cada fragmento do texto.

Numa primeira fase pensou-se em utilizar algoritmos de aprendizagem automática, treinando-os para conseguirem prever se um prospeto estaria ou não de acordo com a legislação dos mesmos. Para isso iriam ser utilizados prospetos já validados pela CMVM como exemplos positivos, e versões mais antigas dos mesmos, que não tinham sido aprovadas como exemplo negativos. Contudo, como a quantidade de dados disponíveis para a elaboração deste trabalho é relativamente pequena, optou-se por alterar a abordagem, passando assim a pensar-se num sistema baseado em regras desenvolvidas com base no regulamento em vigor.

2.1. Revisão sistemática

Para fazer a revisão da literatura o método utilizado foi baseado no PRISMA `http://www.prisma-statement.org/`. O mesmo consistiu em primeiro formular as questões de investigação às quais se pretende responder, que podem ser vistas na secção 1.3, de seguida foram estabelecidos os critérios de inclusão/exclusão dos artigos que viessem a ser encontrados pela pesquisa. Estes critérios devem ser escolhidos antes de fazer a pesquisa para evitar enviesamentos, apesar dos mesmos poderem ser afinados para se adaptar aos métodos aplicados. Para fazer a pesquisa para este trabalho foi utilizado o Scopus `https://www.scopus.com/home.uri` que é uma base de dados bibliográfica e de resumos académicos. Após a pesquisa, deve ser feita uma refinação dos artigos que vão ser utilizados, primeiro através do seu título e resumo, e numa segunda passagem através da introdução e conclusão, onde devem ser escolhidos os artigos que mais contribuem para ajudar a encontrar as respostas às perguntas iniciais. Por fim, deve ser feita uma síntese de todos os artigos que foram previamente escolhidos e analisados, mostrando o contributo dos mesmos para a investigação.

Para encontrar os artigos para este trabalho as palavras-chaves procuradas na base de dados Scopus foram *"Information extraction", "Information Validation", "Legal Documents" e "Natural Language Processing Systems"*. Os critérios de inclusão/exclusão podem ser encontrados na Tabela 2.1.

Tabela 2.1. Filtros utilizados na pesquisa de trabalhos relacionados

Inclusão	Exclusão
Artigos em revista ou atas de conferências	Trabalhos pagos
Trabalhos finalizado	Capítulos de livros
Escritos em inglês	
De 2012 a 2022	
Na área da Computação ou Engenharia	

Esta pesquisa retornou 77 artigos, contudo, como a palavra-chave *"Natural Language Processing System"* é muito genérica, grande parte dos artigos encontrados não estavam bem enquadrados no âmbito deste trabalho. Então, através do título e resumo, foi feita uma seleção de 34 artigos que melhor se enquadram com o tema da dissertação. A partir de referências presentes nos mesmos, foram incluídos mais 9 artigos. O resumo pode ser visto na Tabela 2.1.

Tabela 2.2. Quantidade de documentos após processo de filtragem do trabalho relacionado

Etapas	Total de artigos
Resultado da pesquisa inicial	77
Após analisar titulo e resumo	34
Após inclusão de artigos provenientes de referências dos anteriores	43

2.2. Extração de informação de textos legais

No que toca a extração de informação de documentos legais os objetivos são variados. Desde a entrada em vigor do Regulamento Geral de Proteção de Dados (RGPD) que o estudo na área se tem intensificado e, por exemplo, a procura da melhor abordagem para anonimização de documentos tem sido uma das questões em estudo como pode ser observado em [1]. Outro propósito também muito estudado nesta área é a de sumarização de textos legais, essencialmente por três motivos: o facto de os documentos serem muito extensos, a estrutura dos mesmos ser bastante diversificada e o vocabulário utilizado ser muito específico [2]. Para ambos os exemplos dados anteriormente, é preciso extrair a informação, no primeiro caso para identificar as entidades que se querem anonimizar, no segundo caso o objetivo é perceber qual é a informação mais relevante que deve fazer parte do sumário.

Existem várias técnicas utilizadas para extrair informação de documentos legais, contudo a mais utilizada é baseada em regras como fica evidente em [3]. Neste artigo os autores demonstram que com o passar do tempo se tentam adotar novos métodos como SVM's (Support Vector Machines) ou CRF's (Conditional Random Fields), como pode ser observado no artigo , mas que ainda assim a abordagem baseada em regras continua a ser a mais utilizada de longe como demonstra a Figura 2.2. Em [4] de 2019, foi criada uma *Framework* que deteta e extrai as premissas, queixas e as suas relações a partir de

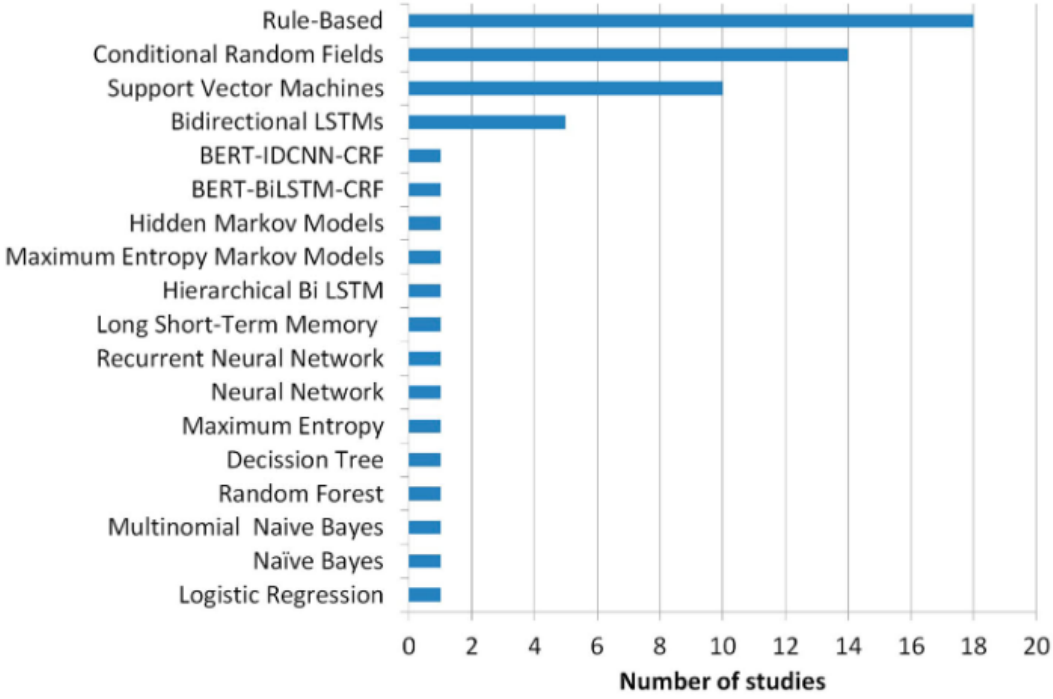


FIGURA 2.1. Distribuição dos métodos de investigação em IE de documentos legais segundo [3]

textos legais não estruturados, processando-os e estruturando os mesmos para posteriormente serem processados por modelos previamente treinados para dados deste tipo. Este trabalho foi desenvolvido com um conjunto de dados de textos legais em Árabe, onde os mesmos só conseguiram 90 documentos, contudo ainda assim usaram algoritmos supervisionados (SVM's e B-Naive Bayes) e obtiveram resultados a rondar os 86% de taxa de acerto. Ainda em [3] os autores mostram que a tarefa mais executada na extração de informação de textos legais é a tarefa de reconhecimento e classificação de entidades nomeadas tem como objetivo identificar e reconhecer entidades como pessoas, organizações, locais, etc. [5]. Um sistema de reconhecimento de entidades nomeadas tem algumas restrições, não só impostas pelo domínio das entidades que foram utilizadas para treinar o mesmo, como também restrições devido à língua. A maior parte dos sistemas de reconhecimento de entidades nomeadas foram treinados para a língua inglesa, se bem que sistemas em Chinês também aparecem muitas vezes na literatura. Em [6] são utilizados modelos de *Embedings* para fazer o reconhecimento de entidades nomeadas em Português. Apesar dos dados utilizados para treino serem grande parte em Português do Brasil, a avaliação é feita utilizando um conjunto de dados proveniente do HAREM que é um concurso de reconhecimento de entidades nomeadas em Português de Portugal [7]. Com esta abordagem os resultados foram bastante positivos tendo conseguido alcançar perto dos 75% de precisão no geral das entidades que pertencem ao HAREM, o que é uma melhoria de cerca de 4% do estado da arte.

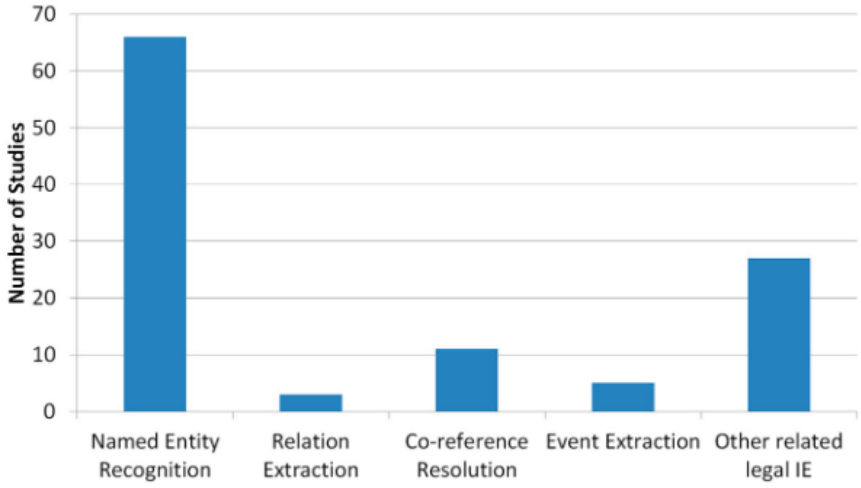


FIGURA 2.2. Distribuição das tarefas de investigação em IE de documentos legais segundo [3]

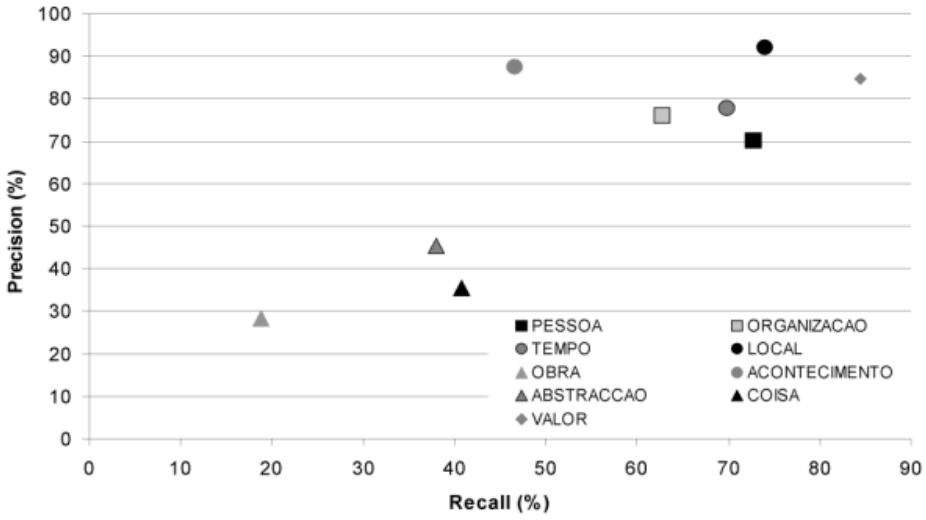


FIGURA 2.3. Métricas da extração de entidades nomeadas consigas em [7]

Existem já alguns sistemas de extração de informação de documentos legais e regulatórios como é o caso do LexNLP, uma biblioteca em Python que visa ajudar nas pesquisas académicas e aplicações industriais recorrendo a técnicas de processamento de língua natural e de aprendizagem automática [8]. Em [9] podemos conhecer com pormenor o *ICIX: A Semantic Information Extraction Architecture*, que é uma arquitetura para extrair informação estruturada a partir de documentos de texto, e que espera obter informação específica sobre um determinado domínio. Este sistema é composto por uma base de dados de documentos, uma base de conhecimento e alguns processos que os autores chamam de serviços tais como o "Pre-Process Service" ou o "Data Curation Service",

entre outros, que funcionam em conjunto para extrair informação, algo que poderá ser adaptado para o nosso problema.

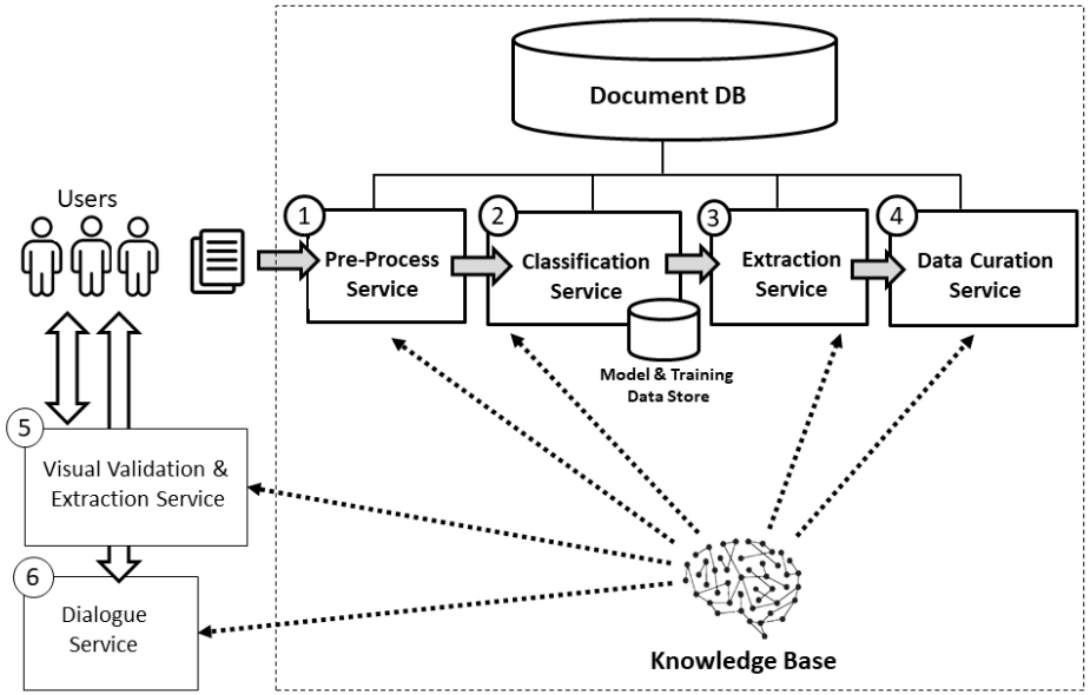


FIGURA 2.4. Arquitetura do ICIX [9]

Outra ferramenta muito interessante e que merece algum destaque é o GATE [10] (General Architecture for Text Engineering), ferramenta esta que foi utilizada para extrair informação a partir de documentos regulatórios em [11]. Esta ferramenta, apesar de já ser relativamente antiga, continua a ser utilizada por alguns investigadores para extrair informação a partir de documentos, pois é bastante completa. Para além de aceitar uma grande variedade de formatos tais como TXT, HTML, PDF, entre outros, é composta por módulos como por exemplo o ANNIE (A Nearly-NEW Information Extraction System) ou o JAPE (JAVA Annotation Pattern Engine) que tornam esta ferramenta muito versátil.

2.3. Utilização de ontologias na extração de informação

Já em 1999 se falava sobre ontologias na Inteligência Artificial, e se defendia o uso das mesmas. Segundo os autores de [12] ontologias são "conteúdos teóricos sobre os tipos de objeto, propriedades dos objetos e relações entre objetos que são possíveis num domínio específico do conhecimento. As mesmas fornecem potenciais termos para descrevermos o nosso conhecimento sobre um domínio". Analisando o problema que se pretende resolver com este trabalho, a provável utilização de ontologias servirá essencialmente o propósito de criar relações entre que regras se devem aplicar a que tipo de prospeto.

No artigo [13] é mostrado como foi criada uma *framework* para extração de informação artigos científicos (OFFIE) a partir de diversas base de dados de artigos. Neste trabalho a extração de informação é feita a partir de regras de fuzzy e de documentos PDF que os

autores defendem que o melhor é converter para XML pois é mais fácil aplicar as regras. A ontologia define a informação que é suposto retirar de cada artigo, como título, resumo, referências, entre outras. Esta abordagem pode ser muito interessante no contexto do nosso trabalho, uma vez que parece possível criar uma ontologia com base na legislação dos prospetos, e usá-la para definir que informação deve ser retirada de cada prospeto dependendo do tipo do mesmo.



FIGURA 2.5. Ontologia proposta por [13]

Em [14] a extração de informação é baseada em ontologias, onde a abordagem utilizada para extrair informação foi dividida em duas fases, uma linguística e uma computacional. Na primeira fase, recorreram a especialistas para gerarem duas ontologias, uma baseada em regras linguísticas e outra no domínio em questão. Já na fase computacional recorrem a um sistema chamado SAURON para criar a base de conhecimento de acordo com a inferência das regras criadas na fase anterior. Há alguns cuidados que se devem ter aquando da criação de uma ontologia a partir de extração de informação. No contexto de documentos legais, sendo um domínio tão específico, o uso de técnicas de aprendizagem automática não é recomendado segundos os autores de [15] sendo que o método que os mesmos consideram mais adequado é recorrer a expressões regulares (REGEX) [16] para identificar padrões, o que ainda assim continua a ter alguns problemas, pois documentos legais contém muitas páginas e muita informação parecida, o que não ajuda. O trabalho dos mesmos está divido em três partes: uma parte de processamento de texto onde aplicam OCR (Optical Character Recognition) pois têm

muitos ficheiros PDF que são digitalizações de documentos, seguido de um sistema de verificação e correção de erros ortográficos, na segunda parte, fragmentam o texto em secções de acordo com as secções presentes na ontologia, por fim, na terceira parte fazem a extração da informação de acordo com a secção que foi atribuído ao pedaço de texto em questão no passo anterior. Este trabalho é ainda envolvido numa experiência de melhoria do mesmo num projeto intitulado uAIS [17], onde os autores se propõem a utilizar o sistema distribuído para melhorar o desempenho e a escalabilidade do mesmo. Em contrapartida, em [18] são utilizados modelos mais sofisticados do que REGEX para conseguirem extrair palavras chaves e definições de documentos legais, utilizando o módulo de sumarização do Tensorflow `https://www.tensorflow.org/` e de seguida o modelo LDA (*Latend Dirichlet Allocation*) para popularem uma ontologia. Usam também o *Word2Vec* do Tensorflow `https://www.tensorflow.org/text/tutorials/word2vec` para conseguirem encontrar palavras com semântica semelhante, e enriquecerem a ontologia, que posteriormente servirá para terem uma base de conhecimento que consiga responder a questões sobre o tema colocadas pelos utilizadores. Criar sistemas que ajudem os utilizadores que não estão familiarizados com o domínio legal a extrair informação de forma mais intuitiva e tornar a mesma mais acessível todos, também tem sido uma prática comum como se pode observar em [19] ou em [20] apesar de nestes trabalhos não serem utilizadas ontologias como no primeiro caso. O uso de ontologias é benéfico, pois recorrendo a estas podem ser feitas inferências e consegue-se compreender melhor as relações que existem entre os temas, algo que não é tão acessível nos métodos tradicionais como bases de dados relacionais [12].

2.4. Sumário

Na Tabela 2.4 apresentam-se os artigos mais relevantes para o trabalho e é feito um sumário dos mesmos. Estes artigos servem como base para abordagem pensada para a elaboração deste trabalho pois têm algumas características em comum e chegaram a resultados animadores través das suas abordagens.

Tabela 2.3. Sumário dos artigos revistos nesta secção

Artigo	**Resumo**
[9]	Apresentam uma proposta de uma arquitetura para extrair informação a partir de documentos legais recorrendo a uma ontologia. Essa ontologia guarda as entidades que devem ser extraídas, as suas relações e como devem ser processadas. Para esse processamento têm vários módulos como pré-processamento, classificação e extração, que estão todos ligados à base de conhecimento.
[4]	Demonstram que conseguiram detetar a informação pretendida com *taxa de acerto* de 86,3% utilizando SVM's e B-Naive Bayes, mesmo tendo um conjunto de dados relativamente pequeno. Interessante tentar esta abordagem no nosso trabalho, tentando assim ultrapassar alguns problemas que as REGEX podem ter em documentos tão extensos, com um domínio muito específico e não estruturados.
[6]	Reconhecimento de entidades nomeadas utilizando modelos de embedings num conjunto de dados em Português, tendo obtido mais de 83% de *precisão* e mais de 81% de *cobertura* com o modelo FlairBBP, algo que poderá ser utilizado numa fase mais avançada do projeto, visto que muitos dos documentos disponíveis para a elaboração deste trabalho encontram-se redigidos em língua portuguesa, utilizando este método para validar se as entidades que queremos encontrar estão presentes no documento.
[13]	Framework para ajudar a extrair informação de artigos científicos com recurso a uma ontologia. Este trabalho tem muitas parecenças ao que se pretende desenvolver, mas precisa de ser adaptado para documentos legais. Será utilizada a ontologia do mesmo como inspiração.

CAPÍTULO 3

Análise dos dados

Neste capítulo redige-se uma breve contextualização dos dados necessários para melhor se entender no que consiste a primeira de três fases da análise destes documentos. A análise formal, como foi visto na secção 1.1, consiste em validar se um conjunto de regras, descritas no Regulamento (UE) 2017/1129 se encontram aplicadas e em conformidade num determinado prospeto. Este é um regulamento muito extenso e redigido a nível europeu, sendo que pode/deve ser complementado com legislação nacional, o que torna esta tarefa muito extensa e meticulosa.

As regras do regulamento são complexas e o mesmo conta com 49 artigos, sendo que, para esta fase, as que devem ser consideradas são as que se encontram no capítulo III do regulamento, que se referem ao conteúdo e formato do prospeto. Para o desenvolvimento do protótipo para avaliar o sumário de um prospeto, apenas foram consideradas as regras presentes no artigo 7º.

Já os prospetos em si, são também documentos muito longos, que devem conter toda a informação considerada necessária para se tomar uma decisão de investimento, o que faz com que os mesmos contenham muita informação crucial ao longo das várias páginas como pode ser observado mais a frente.

3.1. Dados cedidos pela CMVM

Foram cedidos 303 prospetos com o objetivo de serem analisados e darem a conhecer os tipos de documentos utilizados para desenvolver este trabalho. Estes documentos foram avaliados entre 2001 e 2022. 67% destes documentos estão redigidos em língua portuguesa, enquanto os restantes estão em inglês distribuídos ao longo dos diversos anos. Fazendo uma breve análise dos mesmos é possível observar que até 2009 a maioria dos documentos eram redigidos em português, sendo que só a partir daí começou a haver mais prospetos em inglês, tendo mesmo, chegado a haver mais prospetos em inglês do que em português em alguns anos a partir de 2015, como se pode observar na Figura 3.1. Isto é um ponto importante, pois demonstra que a solução desenvolvida deve suportar os dois idiomas.

Um prospeto deverá estar dividido em diversas secções e capítulos ditados pelo Regulamento (UE) 2017/1129, sendo que no conjunto de prospetos que foram cedidos para a elaboração deste projeto, os documentos têm, em média, 171 páginas. Os prospetos em inglês tendem a ser maiores como pode ser visto na Figura 3.1. Os prospetos em português contam com uma média de 64 027 palavras, enquanto que, nos que foram redigidos na língua inglesa esse número aumenta para 83 763. Estes números,que podem ser melhor analisados na Tabela 3.1, demonstram a dimensão de cada documento, tornando-se

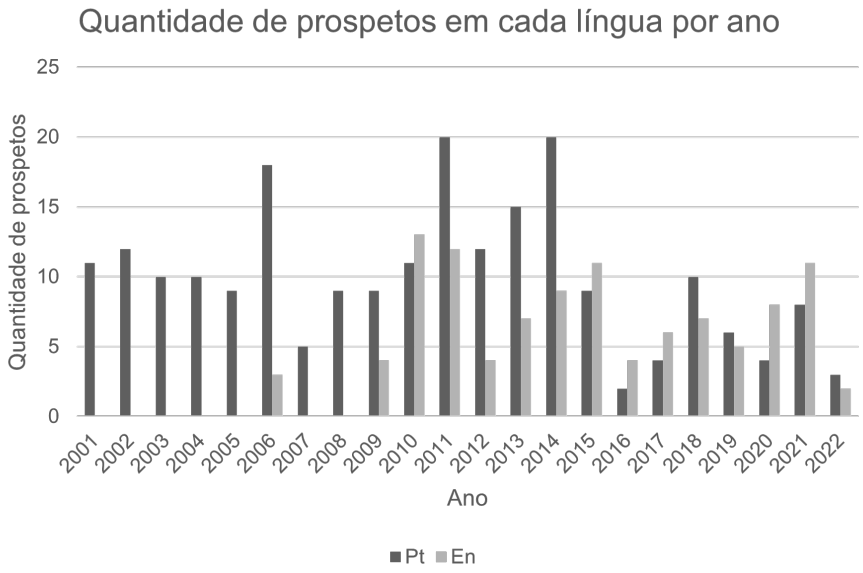


FIGURA 3.1. Quantidade de prospetos emitidos em cada língua do longo dos diversos anos

percetível o demorado tempo de análise, assim como a suscetibilidade a erros por parte dos avaliadores.

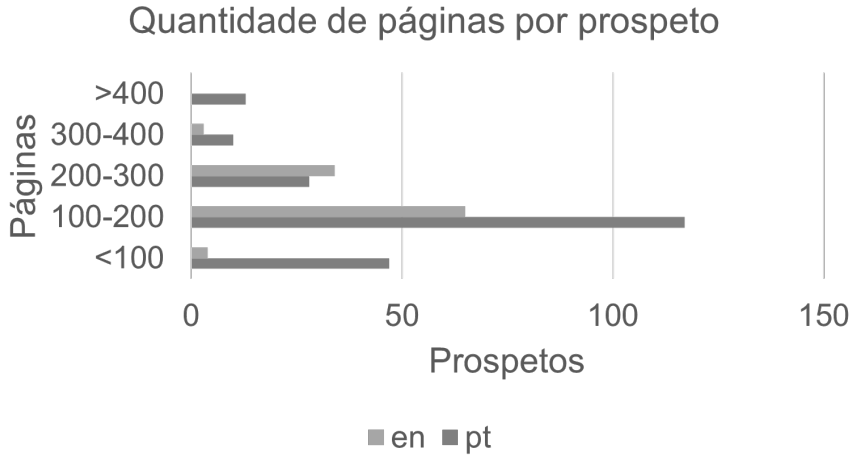


FIGURA 3.2. Quantidade de páginas por prospeto

Alguns destes documentos, contêm ligações para páginas externas ou ficheiros externos, e são representados no documentos através de *URL's*, que devem ser verificados, a fim de avaliar se os mesmos estão ativos. Em média, cada prospeto do conjunto que foi cedido para a elaboração deste trabalho contém 5 links, sendo que o máximo encontrado num só prospeto foram 28 *URL's*. No total foram avaliados 1798 onde 70% dos mesmos (1266) se encontram válidos enquanto os restantes estão inválidos. Uma possível razão para uma quantidade tão elevada de links inválidos é o facto de muitos prospetos já terem mais de 15 anos.

TABELA 3.1. Quantidade de palavras por prospeto contabilizando todas as ocurrências e apenas uma ocurrencia de cada palavra

	Pt	En
Média	64 027	83 763
Máximo	271 369	180 391
Mínimo	3 169	1 503
	Pt	En
Média	7 088	5 890
Máximo	29 700	12 015
Mínimo	887	556

3.2. Regulamento (UE) 2017/1129

Para se entender melhor o problema, é feita uma análise um pouco mais profunda do regulamento, mostrando qual o seu objetivo e o tipo de regras que podem ser extraídas do mesmo.

O Regulamento (EU) 2017/1129 do Parlamento Europeu e do Conselho estabelece as regras e requisitos para a oferta pública de valores mobiliários na União Europeia (UE) e regula o funcionamento do mercado de capitais europeu, tendo como principais objetivos aumentar a integridade e transparência dos mercados e garantir uma proteção adequada aos investidores.

Deste regulamento podem ser extraídas as informações obrigatórias que devem constar num prospeto, sendo algumas delas, as informações sobre o emitente, os valores mobiliários oferecidos, os riscos associados à oferta, os direitos dos titulares dos valores mobiliários, as condições da oferta e qualquer informação relevante para ajudar os investidores a tomar decisões informadas. Essas informações devem estar dispostas numa estrutura clara e organizada, permitindo aos investidores encontrarem facilmente a informação necessária. A estrutura pode variar de acordo com o tipo de oferta, contudo tem um conjunto de secções obrigatórias tais como: o sumário, as advertências, as informações sobre o emitente, as informações sobre os valores mobiliários, os fatores de risco, entre outras.

3.3. Regras analisadas

Existem vários tipos de regras que podem ser extraídas do regulamento para serem aplicadas a um prospeto. Algumas delas são subjetivas como é o caso do ponto 2 do artigo 7º que diz:

"2. O sumário deve ser redigido em linguagem e estilo que facilitem a compreensão da informação, nomeadamente linguagem clara, não técnica, concisa e compreensível para os investidores". Este tipo de regras não será tida em conta nesta fase do trabalho. Já as regras que têm que ver com a forma e estrutura de um prospeto são mais adequadas para esta fase, como é o caso do ponto 4 do artigo 7º do regulamento:

"4. O sumário deve ser constituído pelas seguintes quatro secções:

a) Uma introdução que contenha as advertências;
b) Informação fundamental sobre o emitente;

c) Informação fundamental sobre os valores mobiliários;
d) Informação fundamental sobre a oferta de valores mobiliários ao público e/ou a admissão à negociação num mercado regulamentado."

Ao se analisar esta regra fica-se com uma ideia da estrutura que o sumário deve ter, mas ainda assim, não fica totalmente evidente o que é considerado, por exemplo, "Informação fundamental sobre o emitente". Para se perceber melhor, tem que se continuar a analisar o regulamento.No ponto 6 do mesmo artigo, na alínea a) é possível retirar mais informações: "6. Da secção a que se refere o nº 4, alínea b), devem constar as seguintes informações:

a) Numa subsecção intitulada «Quem é o emitente dos valores mobiliários?», uma breve descrição do emitente dos valores mobiliários, incluindo pelo menos:
   i) o seu domicílio e forma jurídica, o seu LEI, a legislação ao abrigo da qual exerce a sua atividade e o seu país de registo,
   ii) as suas principais atividades,
   iii) os seus principais acionistas, inclusive se o emitente é direta ou indiretamente detido ou controlado e por quem,
   iv) a identidade dos seus principais administradores,
   v) a identidade dos seus revisores oficiais de contas".

Desta regra é possível extrair que na secção "Informação fundamental sobre o emitente" deve existir uma subsecção intitulada "Quem é o emitente dos valores mobiliários?", onde devem constar informações como o domicílio e forma jurídica, o seu LEI (Legal Entity Identifier), país de origem entre outros. Esta analise é feita a todas as regras presentes no artigo 7º, uma vez que são as regras que são necessárias para validar o sumário do prospeto.

CAPÍTULO 4

Metodologia proposta

Neste capítulo é explicado o processo desenvolvido para ajudar na validação automática de um prospeto, das vantagens do mesmo e dos desafios que surgiram durante o desenvolvimento. Este capítulo divide-se em três fases principais:

(1) Extração da estrutura do documento;
(2) Criação das regras com base no regulamento;
(3) Aplicação das regras às devidas secções.

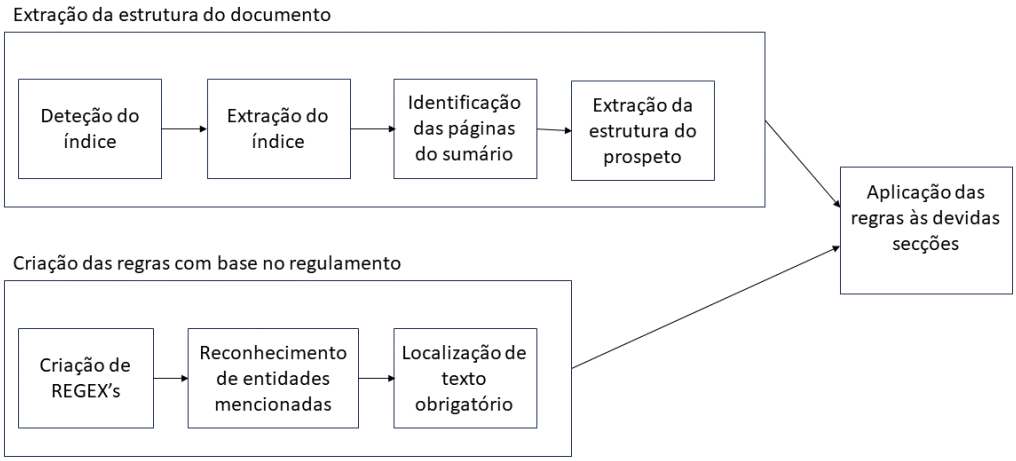


FIGURA 4.1. Esquema dos processos desenvolvidos nas diferentes etapas da metodologia

As duas primeiras fases são independentes, sendo que a terceira depende de ambas. Na primeira fase é descrito todo o desenvolvimento necessário para identificar e extrair a estrutura de um prospeto. Para atingir esse objetivo, recorre-se ao índice que é parte obrigatória destes documentos. Na segunda fase, explica-se como foram desenvolvidas algumas regras recorrendo a técnicas de PLN (Processamento de língua Natural) para identificar alguns pontos exigidos no regulamento analisado anteriormente. Por fim, na terceira fase, são aplicadas as regras às devidas secções e é feita uma avaliação às abordagens optadas a fim de alcançarem os resultados desejados.

Em nenhuma destas fases se recorre a ontologias pois as relações entre objetos e as suas propriedades, que as ontologias descrevem e que permitem a utilização de ferramentas de inferência, não se mostraram relevantes na representação destes documentos. Não se

encontraram nestes documentos conceitos cuja complexidade exigisse a criação de uma ontologia.

4.1. Extração da estrutura do documento

Extrair a estrutura de um documento tornou-se fundamental para se conseguir validar o mesmo. Após a análise do regulamento percebe-se que existem regras diferentes para cada secção do documento, e por isso, para serem aplicadas as regras certas às devidas secções, é necessário encontrar uma solução que permita obter a estrutura dos documentos, uma vez que não existe uma padronização dos mesmos.

O índice é uma secção presente em todos os prospetos, e que dá informação muito útil acerca do mesmo. Através do índice consegue perceber-se que secções estão presentes no documento e em que páginas podem ser encontradas. Com isto, obtém-se uma estrutura do documento que permite navegar diretamente para as páginas das secções que se pretende.

4.1.1. Deteção do índice

A primeira abordagem pensada para detetar o índice foi procurar pela expressão "índice" nas primeiras páginas do documento. Rapidamente se tornou óbvio que não é uma abordagem viável, por alguns fatores:

(1) Nos documentos onde o índice está identificado com um título, existem diversas formas de esse título estar representado, podendo variar ente "índice", "Content", "Table of contents" entre outros. Isto podia ser facilmente contornado, criando uma lista de todas as variações que fossem encontradas, contudo não seria escalável caso no futuro aparecesse uma nova variação.

(2) Muitos dos índices são compostos por mais do que uma página, contudo apenas a primeira é identificada.

(3) Muitos dos documentos não têm qualquer identificação na página do índice, apresentando apenas a Tabela.

Estes motivos levaram a pensar numa solução mais robusta e escalável. O índice é uma secção do documento com um formato muito próprio que normalmente segue o seguinte padrão (o anexo 3 mostra um exemplo que não respeita este padrão):

`[Título da secção] [Caracter/es delimidador/es] [Número da página]`

com o caracter delimitador podendo ser, por exemplo, '`.`' ou '`-`' como se pode observar nos anexos 1 e 2. Esta estrutura é única e não é encontrada em mais nenhuma página do documento. Foi utilizado o modelo *Doc2Vec* [21] para obter uma representação do texto que seja compatível como input requerido pelo modelo de classificação utilizado no passo seguinte.

Foi então criado um modelo de classificação, que recebe uma página do documento e classifica a mesma como índice ou não índice. Para atingir este objetivo, foi treinado um modelo de classificação Logistic Regression. Para treinar o modelo, foram categorizadas manualmente 217 páginas, sendo que um terço das mesmas representam a categoria índice,

TABELA 4.1. Deteção do índice: comparação entre as páginas esperadas e páginas detetadas pelo algoritmo

Doc	Páginas expectáveis	Páginas detetadas	Precisão	Cobertua	F1-Score
1	[2,3,4]	[2]	100%	33%	50%
2	[2,3,4]	[2]	100%	33%	50%
3	[4]	[4]	100%	100%	100%
4	[2]	[2]	100%	100%	100%
5	[2,3,4,5,6,7]	[1,3,4,6,7]	80%	67%	73%
6	[3,4,5,6]	[2,3,5,6]	75%	75%	75%
7	[8,9,10]	[8,9,10]	100%	100%	100%
8	[2,3,4]	[1,2,4]	67%	67%	67%
9	[2,3,4,5]	[3,4,5]	100%	75%	86%
10	[2,3,4]	[4]	100%	33%	50%
11	[3]	[3]	100%	100%	100%
12	[5]	[5]	100%	100%	100%
13	[2,3,4]	[3,4]	100%	67%	80%
14	[2,3,4]	[3]	100%	33%	50%
15	[3]	[3]	100%	100%	100%
16	[3]	[3,14]	50%	100%	67%
17	[3]	[3]	100%	100%	100%
18	[3]	[2,3,5]	33%	100%	50%
19	[3]	[3]	100%	100%	100%
20	[2]	[2]	100%	100%	100%
21	[2]	[1,2]	50%	100%	67%
22	[3]	[3]	100%	100%	100%
23	[2,3,4,5]	[4,5]	100%	50%	67%
24	[3]	[3]	100%	100%	100%
25	[2]	[1,2]	50%	100%	67%
26	[3]	[3]	100%	100%	100%
27	[3]	[ ]	0%	0%	0%
		Média	85%	79%	78%

e os restantes dois terços representam diferentes páginas do documento, garantindo uma adequada diversidade nos dados.

Para gerar representações das páginas, o texto de cada página é extraído linha a linha. Cada linha é de seguida transformada num vetor utilizando o modelo *Doc2Vec* da biblioteca *gensim* [22] para fazer corresponder com o input exigido pelo *Logistic Regression*. O modelo *Doc2Vec* é baseado no modelo *Word2Vec* [23], e tem como objetivo transformar frases (documentos) em vetores, independentemente do seu tamanho.

Para se avaliar esta deteção, foram marcadas manualmente todas as páginas dos índices de 27 documentos e comparadas com a deteção feita pelo algoritmo. Os resultados desta avaliação podem ser observados na Tabela 4.1, e demonstram que o modelo precisa de algumas melhorias.

Para avaliar a prestação do modelo foi gerada a matriz de confusão resultante da classificação. Foram considerados como:

- Verdadeiros Positivos (VP) as páginas corretamente classificadas;

- Falso Positivos (FP) as páginas que o modelo classificou como índice mas que não o são;
- Falsos Negativos (FN) as páginas que o modelo não classificou como índice mas que na verdade o são;e
- Verdadeiros Negativos (VN) as restantes

e de seguida recorreu-se a três métricas para avaliar a classificação:

- **Precisão:** Indica de entre todas as páginas que são da categoria índice, quantas o modelo acertou. É calculado com a formula seguinte:

$$\frac{VP}{VP + FP}$$

;

- **Cobertura:** De todas as páginas que o modelo classificou como índice, quantas estão corretas. É calculado com a formula seguinte:

$$\frac{VP}{VP + FN}$$

;

- **F1-score:**Média harmónica entre precisão e cobertura. É calculado com a formula seguinte:

$$2 * \frac{Preciso * Cobertura}{Preciso + Cobertura}$$

Através da análise da Tabela 4.1 verifica-se que a métrica de *precisão* foi ligeiramente superior ao *cobertura*. São mais preocupantes os cenários em que o modelo não classifica como índice uma página que o devia ser, pois na etapa seguinte, aquando da extração, é necessário que todas as páginas tenham sido detetadas para se conseguir extrair com maior exatidão. Por outro lado, páginas terem sido classificadas como índice não o sendo, não são um problema muito grande, pois na fase seguinte é possível detetar as mesmas graças à sua estrutura referida anteriormente.

Fazendo uma análise mais detalhada às páginas que foram incorretamente classificadas, pode observar-se que alguns casos são páginas que não são inteiramente dedicadas ao índice, começando a secção seguinte ainda na mesma página. Outro cenário de falha são páginas que contém Tabelas, o que pode ser confundível com o índice. Contudo em ambos os cenários, acredita-se que treinando o modelo com mais dados e ainda maior diversidade nos mesmos deve ajudar a melhorar os resultados obtidos nesta fase.

4.1.2. Extração do índice

Uma vez que que as páginas que pertencem ao índice são detetadas, passa-se à fase de extração da informação das mesmas. A extração tem como objetivo encontrar um par [Título,Pág] para cada secção do documento. Para esse efeito começa-se por ler linha a linha de cada página. Primeiro deve ser verificado se essa linha se refere ao título da

página ou não, uma vez que nem todos os índices contém título. De seguida, para cada linha, é feita uma verificação se o último caracter da mesma é um dígito recorrendo à função *isdigit()* da biblioteca *string* do *Python*.

- Caso seja um dígito, são lidos da direita para a esquerda todos os carateres até deixar de ser dígito. Desta forma é obtido o número da página.
- Caso não seja um dígito, está-se perante um título longo que continua na linha seguinte, por isso junta-se a linha seguinte com a atual e trabalha-se como se de uma só se tratasse.

Depois de serem lidos os dígitos, continua-se a leitura de cada caracter, até se encontrar uma letra entre (a-z) ou digito recorrendo à função *isalpha()* também da biblioteca *string* e *isdigit()* respetivamente. Uma vez encontrado, faz-se a leitura de todos os caracteres desde o início até à posição do caractere respetivo, e desta forma obtém-se o título da secção sem ter interferência dos cacteres delimitadores.

Por fim, é necessário verificar se existe offset entre as páginas extraídas do índice e as páginas reais. Este offset pode ser introduzido por diversos fatores, sendo os mais comuns

- Contabilização da página da capa como página 1 do documento
- Introdução de página em branco após a capa e não contabilizar com a mesma para a conetiquetaem das páginas.

Isto introduz incongruências nas páginas extraídas do índice com a conetiquetaem real, o que de seguida dificulta o acesso às secções através da numeração extraída do índice. Para calcular este offset, assume-se que o primeiro par [Título, pág] extraído se encontra correto. Verifica-se então a existência desse mesmo título no intervalo de páginas [pág-2;pág+2]. O offset é calculado subtraindo o número da Pág. onde o título foi encontrado pelo número da Pág. que foi extraída do índice. Por fim, faz-se uma atualização da Tabela do índice, somando o offset a todas as páginas extraídas.

A extração do par [Título, pág] obteve resultados muito positivos como se pode observar na Tabela 4.2, onde se verifica que foram extraídos corretamente 96% dos pares esperados.

Um desafio encontrado durante este processo foi o facto de algumas páginas do índice conterem mais informação como título da página, cabeçalhos ou paginação, o que por vezes pode introduzir alguns erros pontuais. Na Tabela 4.2 pode observar-se também que houve um documento onde a extração falhou muitos dos pares esperados, e após uma avaliação manual, percebe-se que o problema é que o documento em questão não segue a estrutura habitual dos restantes documentos, não tendo caractere delimitador entre o título e o número da página, o que resultou em erros inesperados nessa extração.

4.1.3. Validação dos títulos extraídos do índice

Uma das tarefas que consome muito tempo na validação de um prospeto, é verificar se o índice está de acordo com o documento. Para isso, é preciso validar se cada par [Título, pág] está correto, verificando se na página em questão se encontra o respetivo título da

Tabela 4.2. Comparação entre o número de pares [Título,pág] extraídos e o número real para cada documento

Doc	Real	Extraído	Taxa e acerto
1	26	26	100%
2	38	38	100%
3	21	21	100%
4	102	102	100%
5	143	142	99%
6	68	68	100%
7	102	102	100
8	71	70	99%
9	88	88	100%
10	22	22	100%
11	23	22	96%
12	27	27	100%
13	68	68	100%
14	-	-	-
15	24	24	100%
16	21	8	38%
17	19	19	100%
18	21	20	100%
19	19	19	100%
20	10	10	100%
21	20	20	100%
22	18	16	89%
23	37	36	97%
24	22	22	100%
25	19	17	89%
26	22	22	100%
27	-	-	-
		Média	96%

secção. Para se verificar o título, tem que se ter em conta alguns erros ortográficos ou pequenas variações na escrita do mesmo, por isso a utilização do cálculo de similaridade de *strings* é fundamental para validar se os dois títulos avaliados são os mesmos. Esta tarefa pode ser feita automaticamente como é demonstrado de seguida.

Uma vez que foram obtidos os pares extraídos na secção anterior, para cada um:

(1) navega-se até à página em questão;
(2) extrai-se linha a linha dessa página;
(3) verifica-se se está perante um título longo que ocupe duas linhas e, se sim, extrai-se a linha seguinte e junta-se à atual, continuando o processo normalmente de seguida;
(4) calcula-se a similaridade entre as duas strings (Título previamente extraído e linha atual) recorrendo para isso à classe *SequenceMatcher* da biblioteca *difflib* do *Python*;

TABELA 4.3. Comparação entre validação manual e automática se o título se encontra na página indicada pelo índice para cada documento

Doc	Total	Deteção Automática Correto	Incorreto	Deteção Manual Correto	Incorreto	Taxa de acerto
1	88	86	0	86	0	100%
2	98	98	0	98	0	100%
3	21	21	0	21	0	100%
4	12	12	0	12	0	100%
5	204	201	**3**	204	0	99%
6	90	1	**89**	90	0	1%
7	102	100	2	102	0	98%
8	113	113	0	113	0	100%
9	135	135	0	135	0	100%
10	74	50	**24**	74	0	68%
11	23	23	0	23	0	100%
12	27	27	0	27	0	100%
13	114	114	0	114	0	100%
14	-	-	-	-	-	-
15	24	23	**1**	24	0	96%
16	21	5	16	5	16	100%
17	19	19	0	19	0	100%
18	21	6	15	6	15	100%
19	19	19	0	19	0	100%
20	10	8	2	8	2	100%
21	20	18	**2**	20	0	90%
22	17	16	1	16	1	100%
23	105	103	**2**	105	0	98%
24	22	22	0	22	0	100%
25	19	15	**4**	19	0	79%
26	22	22	0	22	0	100%
27	22	22	0	22	0	100%
					Média	93%

(5) só é considerado positivo, caso a similaridade seja superior a 90% e o título seja a única informação na linha em questão, para assim não haver o risco de se encontrar palavras no meio do texto que sejam confundidos com o título que pretendemos encontrar.

Para se conseguir avaliar esta secção do trabalho, verificou-se manualmente em 27 documentos, se os índice estavam de acordo com o resto do documento, apontando quantos títulos existem no documento, e quantos se encontram ou não na página indicada no índice. Depois foi feita uma comparação dessa validação manual com os resultados obtido pelo algoritmo.

Essa comparação pode ser observada na Tabela 4.3, onde se constata a **negrito** os casos onde o algoritmo não obteve os resultados esperados pela validação manual. Na maioria dos casos, os resultados obtidos pelo algoritmo foram os esperados, incluindo nos

documentos 16,18,20 e 22 onde alguns dos títulos não se encontravam na página indicada pelo índice.

4.1.4. Extração da estrutura do sumário

Através da análise dos prospetos cedidos para a elaboração deste trabalho ficou evidente que, apesar de todos os documentos precisarem de conter a mesma informação, nem todos a tinham, havendo informação em falta em alguns prospetos. O facto de não existir uma estrutura padronizada como pode ser observado nos anexos 3 e 4, seguida por todos os emitentes dificulta muito o processo de validação automática, uma vez que o local no prospeto onde se pode encontrar uma determinada informação varia de documento para documento.

Recorrendo a uma biblioteca do *Python* para ler PDF's denominada *PyMuPDF*, é possível extrair o tamanho e o tipo de letra em que o texto do documento se encontra redigido. Com isto, tentou-se identificar títulos de secções, subsecções e parágrafos, pela diferença dos tamanhos e tipo de letra.

O processo passa por analisar o texto, perceber quais os tipos de letra e tamanhos que existem, criando etiquetas como por exemplo "<h1>" no caso do título principal ou "<p>" no caso dos parágrafos. Depois, aplicar as etiquetas criadas ao texto extraído do documento, e de seguida criar a estrutura a partir dos títulos e subtítulos, tendo como base o trabalho desenvolvido em [24]

4.1.4.1. ***Identificação das etiquetas e marcação do texto*** Nesta subsecção é feita uma descrição do processo de identificação das etiquetas e do processo de marcação do texto:

(1) são contabilizadas quantas palavras existem para cada par [tamanho, tipo de letra]
(2) considera-se que o par que obtiver maior quantidade de palavras é o paragrafo, atribuindo a etiqueta <p> ao mesmo.
(3) são ordenados os restantes pares por ordem decrescente do tamanho.
(4) comparam-se 2 a 2, e caso tenham o mesmo tamanho mas um dos tipo de letra tiver a marcação **"Bold"** então esse passa para a frente na ordem de importância.
(5) atribui-se as etiquetas do tipo *"header"* sendo que o <h1> é o primeiro par, até <hn> que ficará com o par mais pequeno, com n=quantidade de pares existentes.
(6) por último percorre-se o texto novamente, e atribui-se as etiquetas aos pedaços de texto com base no seu tamanho e tipo de letra

4.1.4.2. ***Identificações das secções principais*** No ponto 4 do artigo 7º do regulamento, são indicadas quais as secções principais que devem compor o sumário, sendo elas:

- uma introdução que contenha as advertências;
- informação fundamental sobre o emitente;
- informação fundamental sobre os valores mobiliários;

- informação fundamental sobre a oferta de valores mobiliários ao público e/ou a admissão à negociação num mercado regulamentado.

Graças ao reconhecimento dos títulos feito no passo anterior, é possível agora tentar identificar as quatro secções principais do sumário. Esta identificação é feita através de similaridade de strings, comparando cada título existente no documento com os títulos extraídos do regulamento. O título que obtiver maior similaridade com o suposto é considerado o título da secção principal, obtendo assim as quatro secções desejadas.

Esta abordagem funciona melhor para as secções principais do que para as subsecções, uma vez que os emitentes tendem a seguir o regulamento no que toca aos títulos das secções principais. No que toca às subsecções, existe um maior grau de liberdade, e por vezes encontram-se sinónimos para representar o título que é exigido no regulamento.

4.1.4.3. ***Criação da estrutura do sumário*** A partir das quatro secções principais identificadas no passo anterior, e com recurso às etiquetas atribuídas ao texto, é criada a estrutura do sumário:

- primeiramente é criada uma secção, que pode ser composta por uma lista de secções ou por um parágrafo;
- todos os blocos de texto identificados com uma etiqueta iniciada por "<h" dão origem a novas secções;
- os blocos de texto identificados por <p> são parágrafos da última secção criada;
- sempre que é criada uma nova secção, a mesma é adicionada à lista de secções da última secção com uma etiqueta superior a esta.

4.2. Criação das regras com base no regulamento

Durante a análise ao regulamento, percebe-se que existem algumas regras mais subjetivas. Um exemplo de uma regra desse tipo é no artigo 7º, ponto 6, alínea a) ponto ii) em que a informação que necessita de ser extraída são "As principais atividades do emitente dos valores mobiliários". Sendo uma informação muito subjetiva, não foi possível criar um processo que conseguisse identificar este tipo de informações, contudo, é possível tentar identificar se existe alguma secção do sumário cujo o nome remeta para as atividades. Ainda assim existem regras um pouco mais simples de avaliar, como é o caso por exemplo do artigo 7º, ponto três que diz que o sumário deve ter um máximo de sete páginas. Para avaliar esta regra, basta contar o número de páginas do sumário e verificar se não ultrapassa o máximo exigido.

Foram consideradas regras de aplicação direta, as regras onde o objetivo é somente verificar, por exemplo por meio de técnicas de Processamento de Língua natural (PLN), se uma dada informação se encontra na devida secção. Este tipo de regras são as que melhor se conseguem avaliar, pois não dependem de interpretação para se validar se a informação necessária se encontra presente.

4.2.1. Reconhecimento de padrões através de expressões regulares

Para algumas destas regras foram criadas e aplicadas expressões regulares (*REGEX*) [16] com o objetivo de identificar essas informações num determinado texto. Uma expressão regular é uma cadeia de caracteres que contem uma combinação de caracteres normais e metacaracteres ou metasequencias especiais. Caracteres normais fazem correspondência com eles próprios, enquanto meta-caracteres ou meta-sequencias são caracteres ou sequencias de caracteres que representam quantidades, locais ou tipos de caracteres [25]. Nesta secção são explicadas as *REGEX's* que foram criadas para serem aplicadas ao sumário do prospeto.

ISIN: o *ISIN* é um código de identificação internacional atribuído aos valores mobiliários para facilitar transações bolsistas. É um código alfanumérico de 12 caracteres, sendo que os dois primeiros identificam o país onde o título foi emitido.

[A-Z][A-Z](\ )0,1(\ d|[A-Z])10

- **[A-Z][A-Z]** dois caracteres para código do país
- **(\ )0,1** Espaço em branco 0 ou 1 vez
- **(|[A-Z])10]** 10 dígitos ou carateres

LEI: o *Legal Entity Identifier (LEI)* é um código que identifica univocamente pessoas coletivas e outros intervenientes no mercado. É um código alfanumérico com 20 caracteres.

[0-9A-Za-z]{20}

- **[0-9A-Za-z]** qualquer dígito ou letra de [a-z] maiúscula ou minúscula
- **{20}** tamanho exato de 20 carateres

Dados de contacto: os dados de contacto considerados foram o *email* e *Telefone*.

- **[Email]** [_a-z0-9-]+.[_a-z0-9-]+@[a-z0-9-]+.[a-z0-9-]+.[a-z]{2,3}
  - **[_a-z0-9-]+.[_a-z0-9-]+** qualquer conjunto de palavra ou dígito repetido 1 ou mais vezes podendo conter um ponto final no meio. ex: pessoa, pessoa.pessoa, pessoa123, pessoa.123...
  - **@[a-z0-9-]+.[a-z0-9-]+** qualquer conjunto de palavras ou dígito repetidos 1 ou mais vezes antecedido do caractere @ ex: @hotmail, @exemplo.2, @exemplo12.exemplo
  - **.[a-z]{2,3}** conjunto de letras de [a-z] com 2 ou 3 caracteres antecedido de um ponto final ex: .com, .pt
- **[Telefone]** (\+351|351|0351)0,1\ *(9[1236]{1}|2[0-9]{1})[0-9]{1}(\ |-){0,1}[0-9]{3}(\ |-){0,1}[0-9]{3}
  - **(\+351|351|0351)0,1** identificador de número português
  - **(9[1236]{1}|2[0-9]{1})[0-9]{1}** combinações possíveis para os dois primeiros caracteres podendo começar por 9 seguido de [1,2,3,6] no caso de um número móvel ou começando por 2 seguindo de qualquer dígito no caso de ser um número fixo
  - **[0-9]{1}** qualquer dígito representado 1 vez

- **[Note:]** A regra da forma que foi desenvolvida permite que o número de telefone seja escrito com os 9 dígitos seguidos, ou em grupos de 3 separados por (" ") ou ("-")

Data: A data de aprovação do prospeto pode ser escrita de diversas formas, contudo a mais comum nos diversos prospetos que foram verificados é d̈d de mês de yyyy: ([0-9]{1,2}\s?d?e?\s?[a-zA-Z]{4,9}\s?d?e?\s?[0-9]{4})

- **[0-9]{1,2}** um ou dois digitos de 0 a 9
- **\s?d?e?\s?** tenta encontrar o padrão " de " podendo ou não conter o espaço antes e depois da palavra
- **[a-zA-Z]{4,9}** tenta detetar um conjunto de 4 a 9 caracteres, que representam o nome dos meses por extenso
- **[0-9]{4}** 4 dígitos que representam o ano

4.2.2. Reconhecimento de entidades nomeadas

Através da abordagem anterior não foi possível aplicar todas as regras necessárias, pois algumas informações podem tomar diferentes formas como é o caso de nomes de pessoas/empresas, moradas, entre outras, e por esse motivo as REGEX's não são a melhor abordagem para esses casos. Tentou-se então aplicar outra técnica de PLN, o Reconhecimento de entidades Nomeadas (NER).

Uma entidade nomeada é um termo que representa um objeto no mundo real como pessoas, locais ou organizações. É designada como um elemento rígido ou um elemento atómico ou membro de uma classe semântica que pode variar consoante o domínio de interesse. No domínio geral, uma pessoa, um local, uma organização, etc. são entidades importantes; por exemplo no domínio da medicina, os nomes dos medicamentos e os nomes das doenças são reconhecidos como entidades [5].

Reconhecimento de entidades nomeadas é uma sub tarefa importante de extração de informação [26] e tem como objetivo reconhecer entidades nomeadas em textos não estruturados e escritos em língua natural, identificando a categoria a que cada entidade pertence, como pessoas, locais, etc. Existem várias abordagens para desenvolver processos de NER, como modelos baseado em regras, modelos estatísticos ou modelos que recorrem a *deep learning*.

Para aplicar esta técnica recorreu-se ao *Spacy* [27] que é uma biblioteca de processamento de língua natural (PLN) avançada para *Python*. Esta biblioteca tem vários módulos e disponibiliza uma grande diversidade de *pipelines* de PLN em diferentes idiomas. Neste trabalho tirou-se partido do pipeline *NER* tanto em português ("pt_core_news_md") como em inglês ("en_core_web_lg"). Isto permitiu validar se algumas informações que não eram possíveis de se identificar com REGEX's se encontram no prospeto ou não.

No artigo 7º, ponto 5, alínea b) é pedido que se encontre a identidade, os dados de contacto e o LEI do emitente. Se para os últimos dois foi possível identificar através de aplicação de REGEX's, a identidade do emitente é mais difícil encontrar através de um padrão. Recorrendo à classe *"ORG"* das entidades que o *Spacy* consegue identificar, é

possível validar se num texto se encontra uma organização, que neste exemplo poderá ser o emitente.

Outra regra onde o NER foi importante é no artigo 7º, ponto 6, alinea a) ponto iii) onde é suposto encontrar os acionistas. Os acionistas podem ser tanto pessoas como organizações. Ambas são classes de entidades reconhecidas pelo *Spacy*. Tal como nos exemplos anteriores, o pipeline permite identificar a informação necessária.

A última regra para a qual se recorre ao NER, é a identificação do país de registo de uma organização. Através da etiqueta *"LOC"* é possível identificar localizações, sendo os países identificados com esta etiqueta. A desvanetiquetaem é que outras localizações como cidades ou monumentos como "Pequim" ou "Torre Eiffel" respetivamente também são identificados neste grupo. Para restringir apenas a países é possível criar uma lista com vários países e considerar apenas os resultados do NER que estejam contidos nessa lista. Ainda assim, se estiver escrito mais do que um país no texto, não é possível identificar qual deles corresponde ao país de registo, apesar de ficarem registado no relatório final todas as ocorrências.

Os resultados obtidos com o *Spacy* foram importantes para perceber que este método poderia resultar, contudo, o *Spacy* tendia a captar muitos *tokens* que não devia, como por exemplo a palavra "Emitentes" ou "CodVM" quando a procura era por uma organização. Então decidiu-se optar por um pipeline de NER que apresenta resultados melhores recorrendo a *Transformers* [28] chamado *"wikineural-multilingual-ner"* [29] que é um modelo que também recorre a redes neuronais e suporta nove idiomas.Este modelo tem a vanetiquetaem de não ser preciso identificar o idioma do prospeto ao contrário do que acontecia no *Spacy*.

4.2.3. Localização de excertos no texto através de similaridade

Por fim, o último grupo de regras desenvolvidas tem como objetivo encontrar textos que devem estar escritos no prospeto, mas que podem sofrer pequenas alterações.

Um exemplo desta regra no sumário é o caso das advertências, onde no Regulamento diz:

> " Da referida secção devem constar as seguintes advertências:
> a) O sumário deverá ser entendido como uma introdução ao prospeto;
> b) Qualquer decisão de investimento nos valores mobiliários deverá basear-se numa análise do prospeto no seu conjunto pelo investidor;
> c) Se aplicável, uma advertência de que o investidor pode perder a totalidade ou parte do capital investido e, caso a responsabilidade do investidor não se limite ao montante do investimento, uma advertência de que o investidor pode perder mais do que o capital investido e a extensão de tal perda potencial;
> d) Caso seja apresentada em tribunal uma queixa relativa à informação contida num prospeto, o investidor queixoso poderá, nos termos do direito nacional,

ter de suportar os custos de tradução do prospeto antes do início do processo judicial;

e) Só pode ser assacada responsabilidade civil às pessoas que tenham elaborado o sumário, incluindo qualquer tradução deste, mas apenas caso o mesmo, quando lido em conjunto com as outras partes do prospeto, contenha menções enganosas, inexatas ou incoerentes ou não preste a informação fundamental para ajudar os investidores a decidirem se devem investir nesses valores mobiliários;

f) Se aplicável, a advertência relativa à compreensão exigida nos termos do artigo 8.o, n.o 3, alínea b), do Regulamento (UE) n.o 1286/2014."

Neste caso, as advertências devem poder ser encontradas na introdução, contudo, as mesmas podem sofrer algumas alterações ou serem escritas de outras formas. Isto levou à criação de uma solução onde se calcula similaridade entre frases, e, caso duas frases obtenham uma similaridade igual ou superior a 65%, são consideradas a mesma. No entanto, duas das advertências, apenas devem constar no caso de "serem aplicáveis", algo que não nos é possível verificar.

O processo de deteção das advertências é feito do seguinte modo:

(1) É criada uma lista com as frases que representam as advertências que devem ser encontradas, na maioria mantendo a forma como esta escrita no regulamento. No entanto a alinea c) e f) sofrem alterações, ficando "O investidor pode perder a totalidade ou parte do capital investido" e "Está prestes a adquirir um produto que não é simples e cuja compreensão poderá ser difícil." respetivamente.
(2) Ao texto que se quer encontrar as advertências aplica-se a função *sent_tokenize* da biblioteca *nltk* [30] para se obter cada frase que se encontra no texto.
(3) Para cada advertência, é calculada a similaridade com cada frase.
(4) É guardada a frase com similaridade maior.
(5) Se essa frase obtiver uma similaridade superior a 60% considera-se que se encontrou a advertência em questão.
(6) Caso contrário, retorna-se uma mensagem que avisa sobre a ausência dessa mesma advertência.

Outra regra criada que pertence a este grupo é a deteção da moeda no qual o valor mobiliário é transacionado. Tendo em conta que existe uma lista de moedas finita, é possível detetar que moeda está presento no documento, comparando cada palavra do texto com as moedas existentes. É preciso ter cuidado com o facto de a mesma moeda poder estar escrita de diferentes formas, por exemplo "Euro", "Euros" ou "EUR". Se para o caso especifico do Euro era possivel detetar recorrendo a técnicas de PLN como *stemming* ou *lemmatization*, para o caso das variações do Dólar já não seria possível. Para contornar este problema, foi criado um dicionário onde a chave é o nome da moeda, e o valor é uma lista das diferentes nomenclaturas que podem ocorrer.

4.2.4. Regras que não obtiveram resultados positivos

Algumas regras demonstraram ser mais complicas de se desenvolver por diversos motivos. O domicílio é um exemplo disso mesmo, uma vez que é uma morada, e as moradas podem ser escritas de diferentes formas, por exemplo começar por "Rua", "Avenida", "Praceta" ou abreviações, ter ou não código postal, entre outras diferenças, o que dificulta o processo de criação de uma regex para encontrar um padrão. Tentou-se utilizar o NER recorrendo à etiqueta *"LOC"*, contudo ainda assim não foi possível identificar moradas com exatidão.

O valor nominal de um ativo financeiro, aparentemente é algo muito simples de identificar, uma vez que é composto apenas por alguns dígitos podendo ser seguido ou não pelo símbolo ou da nomenclatura da moeda no qual é transitado, por exemplo "5,00 €" ou "3,50 euros". No entanto, num contexto de documentos legais para instrumentos financeiros, estas representações encontram-se muitas vezes no documento, por isso, não é possível afirmar qual dos valores encontrados representa o valor nominal do ativo. O mesmo acontece para o número de valores mobiliários emitidos.

Para se detetar a categoria dos valores mobiliários, é criada uma lista com as possíveis categorias, uma vez que também existe uma quantidade finita das mesmas. Contudo, os resultados não foram tão bem sucedidos como seria expectável, uma vez que alguns dos nomes das categorias mais utilizadas são palavras comuns neste tipo de documento. É o caso de "Ações" ou de "Obrigações", onde ambas as palavras podem ser encontradas ao longo dos prospetos, com significados completamente diferentes, e por isso, esta deteção não foi incluída.

4.3. Aplicação das regras às devidas secções

Para se poderem aplicar as regras desenvolvidas às secções corretas, á feita uma identificação manual de que regras devem ser aplicadas a cada secção, ainda que algumas regras se repitam em diferentes secções.

Para esta fase é essencial que todas as regras sejam desenvolvidas respeitando a mesma estrutura de *input* e de *output*, pois recorreu-se a *higher order functions* para identificar para cada secção que regras deviam ser aplicadas. Para cada secção é criado um dicionário com a seguinte informação como se pode observar um exemplo no código representado abaixo:

```
    . Estrutura do dicionário de identificação de que regras devem ser aplicadas
    a uma secção
SectionB = {
    'sectionName': "Section B",
    'subsection': ["Section B1","Section B2", "Section B3"],
    'rules': {
        'info1':funcaoInfo1,
        'info2':funcaoInfo2,
```

```
        'info3e4':[funcaoInfo3,funcaoInfo4],
        'info5':funcaoInfo5,
    }
}
```

(1) sectionName: indica o nome da secção principal à qual as regras devem ser aplicadas.
(2) subsections: uma lista, que pode ser vazia, com o nome das subsecções que a respetiva secção principal deve conter.
(3) rules: um dicionário, onde a *key* representa o nome da regra e o *value* é uma lista de funções que compõe uma regra.

É criado um dicionário para cada secção, que serve como input para uma função que trata da validação tanto das subsecções como das regras de um documento. Para validar as subsecções, tira-se partido da estruturação explicada anteriormente, e recorrendo novamente à similaridade de strings compara-se os títulos de cada subsecção do documento com os títulos provenientes do dicionário. Para validar as regras, aplica-se cada função proveniente do dicionário ao texto obtido através da estruturação.

Esta abordagem funciona bem caso os documentos obedeçam à mesma estrutura, ou seja, se for possível identificar as secções principais requeridas no Regulamento. Um problema desta abordagem é, caso o Regulamento sofra modificações, nomeadamente ao nível do título das secções ou nas regras que devem ser aplicadas, as mesmas têm que ser feitas manualmente nestes dicionários, o que pode trazer alguns constrangimentos.

CAPÍTULO 5

Avaliação da validação do prospeto

A validação de um prospeto passa por duas fases que se completam mas que funcionam em separado. Primeiro foram desenvolvidas algumas regras, que posteriormente são aplicadas às secções detetadas na extração da estrutura do documento como se pode observar no Capítulo 4. A avaliação desta fase do projeto também ocorre em duas partes distintas. Primeiro são avaliadas as regras que foram desenvolvidas. Para isso, cada regra é avaliada individualmente através de métricas como *taxa de acerto*, *precisão* e *cobertura*. De seguida é feita uma avaliação à extração da estrutura do documento, onde se tenta perceber se todas as secções foram bem extraídas e identificadas.

5.1. Avaliação das regras desenvolvidas

Para cada regra foi criado um conjunto de 50 parágrafos à exceção da regra das advertências que apenas teve 10. No conjunto dos parágrafos, alguns contêm a informação que é necessário detetar, outros não. Para além disso, a informação que pretendemos detetar pode estar presente mais do que uma vez em alguns parágrafos.

Na Tabela 5.1 fica evidente que no primeiro grupo de regras, *REGEX*, todas as regras obtiveram resultados ótimos. Isto era expectável, uma vez que as informações que procuram seguem padrões bem definidos, por isso são sempre bem detetadas quando estão presentes e não são confundidas com outras informações.

Tabela 5.1. Métricas obtidas através do teste das regras

		Parágrafos	Expectável	Encontrado	Accuracy	Precision	Recall	F1-score
Regex	ISIN	50	32	32	100%	100%	100%	100%
	LEI	50	30	30	100%	100%	100%	100%
	Email	50	30	30	100%	100%	100%	100%
	Telefone	50	25	25	100%	100%	100%	100%
	Data	50	25	25	100%	100%	100%	100%
NER	Pais Registo	50	16	33	48%	48%	100%	65%
	Organizacoes	50	33	22	67%	100%	67%	80%
	Pessoas	50	30	22	63%	91%	67%	77%
Deteção de texto	Adv1	10	10	7	70%	100%	70%	82%
	Adv2	10	10	9	90%	100%	90%	95%
	Adv3	10	3	1	33%	100%	33%	50%
	Adv4	10	10	7	70%	100%	70%	82%
	Adv5	10	10	2	20%	100%	83%	41%
	Adv6	10	0	0	0%	0%	0%	0%
	Moedas	50	30	22	73%	100%	58%	25%

O segundo grupo de regras, as que recorrem ao processo de *NER*, obteve resultados um pouco piores. No caso da deteção do país de registo, ao utilizarmos a etiqueta "LOC" muitas vezes são identificadas outras localizações para além de países, como monumentos ou cidades. Isto faz com que a *precisão* seja um pouco mais baixa, embora sempre que exista um país, o mesmo é detetado como podemos observar pelo *cobertura*. Uma melhoria que pode ser feita a este processo de identificação do país de registo é ter uma lista com os países mais vezes detetados, e apenas aceitar os resultados do *NER* que façam parte das duas listas. Desta forma, localizações que são detetadas e não são países não são consideradas. No que toca a organizações e pessoas, os resultados foram um pouco superiores aos anteriores, contudo ainda se podem observar algumas falhas na identificação de ambas as categorias. Ainda assim, é percetível que na maioria dos casos as informações necessárias são bem detetadas.

Por fim, o último grupo de regras, tenta descobrir excertos de texto através de similaridade. Para avaliar a deteção da moeda no qual um ativo financeiro é transacionado, o algoritmo começa por tentar descobrir o nome de diferentes tipos de moeda em 50 textos. Desses 50 textos apenas 30 continham nomes do moedas, mas ainda assim, apenas 22 foram identificados. Isto explica-se pelo facto de haverem muitas possibilidades de identificar a mesma moeda como "euro" e "€", e de haverem diversas moedas diferentes. O resultado pode ser melhorado adicionado ao dicionário mais representações diferentes. Este processo pode beneficiar substancialmente de aprendizagem ativa, utilizando *feedback* do utilizador para aumentar a quantidade de variações possíveis.

No caso das advertências, cada uma foi testada individualmente, e foi testada em parágrafos de documentos reais, onde, primeiro é feita uma validação manual para averiguar quantas advertências são expectáveis em cada documento. As advertências 1,2,4 e 5 encontram-se presentes em todos os documentos analisados, enquanto a advertência 3 só é expectável que seja detetada em três dos mesmos. A última, a advertência 6, não se espera encontrar em nenhum documento analisado. Neste teste não se espera que sejam encontradas mais advertências do que as que lá estão, ou seja, o número de "Encontrados" não pode ser superior ao "Expectável" da Tabela 5.1. O ideal seria ambas as colunas terem os mesmos valores. Em geral, os resultados têm níveis aceitáveis, contudo a advertência 5 ficou muito longe do esperado. Isto pode ser explicado pelo facto de ser uma frase relativamente longa, o que aumenta a probabilidade de haver erros ortográficos ou de ser escrita de outra forma, e por isso não alcançar o limiar definido.

No conjunto de todas as regras, os resultados foram positivos e fica percetível que as mesmas conseguem identificar as informações necessárias. Por vezes existem algumas falhas, contudo as mesmas são facilmente detetadas pela validação humana, uma vez que estas regras têm como objetivo ajudar na validação de um prospeto.

Tabela 5.2. Resultado da avaliação da extração da estrutura de um prospeto

	Secções		Subsecções		Notas
Ficheiro	Documento	Encontradas	Documento	Encontradas	
0					Não foi possível extrair o índice
1	5	4	30	0	
2	24	0	0	0	
3	5	**0**	37	0	
4	4	**4**	7	7	100%
5					Não foi possível extrair o índice
6	4	4	16	1	
7	5	4	37	2	
8	5	**4**	37	2	
9	13	**4**	0	0	
10					Não foi possível detetar sumário
11					Não foi possível detetar sumário
12	5	4	37	1	
13					Não foi possível extrair o índice
14					Não foi possível detetar sumário
15	37	0	0	0	
16	42	0	0	0	
17					Não foi possível detetar sumário
18	42	0	0	0	
19					Não foi possível detetar sumário
20	4	**4**	7	2	
21	26	**4**		0	
22	5	**4**	29	0	
23					Não foi possível detetar sumário
24					Não foi possível extrair o índice
25					Não foi possível detetar sumário
26	44	0	0	0	

5.2. Avaliação da extração da estrutura de um prospeto

Para validar a estrutura de um prospeto, o processo é dividido em várias fases. Neste teste a validação é feita ao processo como um todo, uma vez que partes como a deteção do índice e extração do mesmo já foram testadas anteriormente.

- O processo começa por tentar detetar o índice (já foi avaliado);
- validar o offset do mesmo;
- detetar o sumário através do índice;
- extrair o sumário;
- detetar as quatro secções e sete subsecções esperadas

Apenas se considera que o processo correu bem se as quatro secções e as sete subsecções forem detetadas com sucesso. Deste modo avalia-se o processo como um todo, ainda que seja possível tentar olhar mais em detalhe e identificar onde estão as falhas. As quatro secções e as sete subsecções são o mínimo exigido pelo regulamento, embora possam existir mais. O grande problema é, como é possível observar mais à frente, que muitas vezes os documentos não respeitam essa formatação, podendo variar muito de documento para documento.

Como se pode observar na Tabela os documentos variam bastante na sua estrutura não existindo uma estrutura padronizada. Em sete documentos não foi possível detetar o sumário. Após uma validação manual, constata-se que em cinco deles o sumário não está presente no índice, por isso o processo não consegue concluir as restantes tarefas. Quanto aos quatro documentos onde não foi possível detetar o índice, a solução pode passar por treinar o modelo de deteção com uma maior quantidade de dados.

Nos restantes documentos, percebe-se que esta abordagem não funcionou do modo esperado, por motivos variados. O modo de deteção das secções e subsecções é através do calculo da similaridade de strings, contudo o título das secções nem sempre está escrito como é procurado. Um exemplo é a subsecção "Quem é o emitente?", que muitas vezes é identificada nos documentos como "Denominação jurídica do emitente". Neste caso deviam ser consideradas a mesma subsecção, contudo o modo de validação não consegue detetar estas variações. Uma possível resolução para este problema passa por aumentar a quantidade de possibilidades para cada subsecção, e sempre que uma nova variação for detetada, adicioná-la à lista de possibilidades. Desta forma, aumenta-se a quantidade de variações para cada título de subsecção e incrementa-se as probabilidade de serem detetadas.

Este problema não se encontra tantas vezes nos títulos de secções, uma vez que os mesmos são mais fixos e não existem tantas variações, daí ser possível detetar mais vezes como se pode observar na Tabela 5.2.

CAPÍTULO 6

Conclusões e trabalho futuro

O principal objetivo deste trabalho é perceber se seria possível automatizar a validação de prospetos financeiros recorrendo a regras de *text mining*. O foco do prototipo desenvolvido foi perceber as limitações que as diferentes abordagens podiam ter na validação de um documento, recorrendo para isso a alguns documentos que foram cedidos pela CMVM e também ao regulamento que dita as regras que os documentos devem obedecer.

O projeto foi separado em duas partes. Primeiro foram criadas regras com base no regulamento para posteriormente serem aplicadas ao documento. Estas encontram-se separadas em três grupos, as REGEX's, NER e deteção de texto através da similaridade. As regras obtiveram bons resultados sendo que as regras desenvolvidas através de REGEX's foram as que melhor se comportaram. As regras desenvolvidas recorrendo ao NER podem sofrer algumas melhorias, que nas próprias regras, quer no modelo de NER que é utilizado. No que toca as regras de deteção de texto através da similaridade, as mesmas obtiveram bons resultados, contudo podem também ser melhoradas, recorrendo a métodos que consigam capturar semântica e contexto e não só palavras. Ainda assim algumas regras ficaram por desenvolver devido à sua subjetividade. Esse tipo de regras exigirá técnicas diferentes das utilizadas neste trabalho.

A segunda parte do projeto passou por tentar extrair a estrutura do prospeto recorrendo ao índice, algo que se tornou um desafio muito difícil pois não existe uma estrutura padrão para os documentos. O facto de cada empresa que emite um prospeto ter um certo grau de liberdade para o escrever, faz com que os mesmos sejam muito diferentes entre si. Embora seja obrigatório que existam certas secções e subsecções, as mesmas podem estar identificadas de diferentes maneiras e por isso não é fácil identificar as mesmas para validação. Foi então tentada uma abordagem de identificar o índice, para de seguida se conseguir identificar em que páginas se encontra o sumário, e só depois se tenta extrair o sumário, recorrendo à diferente formatação da letra com que o mesmo é escrito.

A elaboração deste trabalho pretendia responder às seguintes questões:

- Quais as dificuldades na sistematização e codificação das regras indicadas no regulamento que rege a validação de prospetos mobiliários? **R:** A codificação das regras pode passar por diferentes abordagens dependendo do tipo de regra, sendo elas utilização de REGEX's, que demonstram bons resultados embora seja um recurso com limitações, Técnicas de PLN avançadas como NER, que apenas podem ser utilizados em alguns casos, contudo apresentam resultados satisfatórios e deteção de similaridade de excertos do texto, que pode ser uma abordagem

promissora, principalmente se houver melhorias no sentido de aprendizagem automática.

- De que forma se consegue extrair o texto dos prospetos mantendo a sua estrutura? **R:** Existem algumas bibliotecas em *Python* que podem ajudar nesta fase, ainda assim este processo podia ser facilitado caso houvesse uma normalização dos prospetos para facilitar o mapeamento da sua estrutura.
- É possível mapear as regras desenvolvidas para aplicar às devidas secções? **R:** Sim, é possível mapear as regras para aplicar às secções, ainda que seja considerado o ponto mais dificil deste projeto, uma vez que há pouca normalização dos documentos recebidos pela CMVM.
- Qual a eficácia de um sistema de validação de prospetos construído deste modo? **R:** Com o trabalho realizado não é possível chegar a uma conclusão exata da eficácia de um sistema construido deste modo. Contudo, dependendo da técnica utilizada a eficácia pode ser diferente. As REGEX's atingiram resultados muito bons, contudo a sua utilização é algo limitada, pois para regras mais elaboradas podem não ser a melhor abordagem. O uso de NER também mostrou ser uma abordagem promissora embora precise de alguns ajustes e melhorias. Por fim a abordagem de deteção de textos através de similaridade também mostrou ser uma abordagem promissora que pode melhorar bastante com recurso ao *feedback* do utilizador.

No geral, este trabalho demonstra que é possível criar um processo que ajuda na validação de prospetos financeiros, ainda que seja sempre necessária uma validação humana. Havendo uma padronização dos documentos passa a ser possível extrair as secções com mais precisão, e de seguida aplicar as regras às mesmas, uma vez que as regras obtiveram bons resultados.

Embora este trabalho tenha conseguido resultados positivos, também se tornaram evidentes algumas das dificuldades e estão agora mais claros os caminhos que podem levar este tipo de aplicações mais longe. Se por um lado a padronização dos documentos resolveria alguns dos problemas detetados durante a elaboração do mesmo, a utilização de modelos que capturam a semântica e o contexto podiam ajudar bastante no desenvolvimento de regras que não foram consideradas neste trabalho. Algumas regras que obtiveram resultados menos bons podem também ser desenvolvidas de outra forma, permitindo o *feedback* de utilizadores como forma de aprendizagem ativa, o que trará com certeza resultados mais positivos. A deteção do índice também é um ponto importante que merece alguma atenção, uma vez que é a base para a identificação das restantes secções.

Bibliografia

[1] D. Biesner, R. Ramamurthy, R. Stenzel, *et al.*, "Anonymization of german financial documents using neural network-based language models with contextual word representations," *International Journal of Data Science and Analytics*, vol. 13, pp. 151–161, 2021.

[2] A. Kanapala, S. Pal e R. Pamula, "Text summarization from legal documents: A survey," *Artif. Intell. Rev.*, vol. 51, no. 3, pp. 371–402, Mar. 2019, ISSN: 0269-2821. DOI: 10.1007/s10462-017-9566-2. [Online]. Available: https://doi.org/10.1007/s10462-017-9566-2.

[3] F. Solihin, I. Budi, R. F. Aji e E. Makarim, "Advancement of information extraction use in legal documents," *International Review of Law, Computers & Technology*, vol. 35, no. 3, pp. 322–351, 2021. DOI: 10.1080/13600869.2021.1964225.

[4] K. Jasim, A. T. Sadiq e H. S. Abdullah, "A framework for detection and identification the components of arguments in arabic legal texts," in *2019 First International Conference of Computer and Applied Sciences (CAS)*, 2019, pp. 67–72. DOI: 10.1109/CAS47993.2019.9075650.

[5] A. Goyal, V. Gupta e M. Kumar, "Recent named entity recognition and classification techniques: A systematic review," *Computer Science Review*, vol. 29, pp. 21–43, 2018, ISSN: 1574-0137. DOI: https://doi.org/10.1016/j.cosrev.2018.06.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574013717302782.

[6] J. Santos, B. Consoli, C. dos Santos, J. Terra, S. Collonini e R. Vieira, "Assessing the impact of contextual embeddings for portuguese named entity recognition," in *Proceedings of the 8th Brazilian Conference on Intelligent Systems*, 2019, pp. 437–442.

[7] D. Santos, N. Seco, N. Cardoso e R. Vilela, "HAREM: An advanced NER evaluation contest for Portuguese," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy: European Language Resources Association (ELRA), May 2006. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2006/pdf/59_pdf.pdf.

[8] M. Buey, C. Roman, A. Garrido, C. Bobed e E. Mena, "Automatic legal document analysis: Improving the results of information extraction processes using an ontology," in May 2019, pp. 333–351, ISBN: 978-3-319-77603-3. DOI: 10.1007/978-3-319-77604-0_24.

[9] A. L. Garrido, A. Peiro, C. Bobed, E. Mena e C. Morte, "Icix: A semantic information extraction architecture," in *Proceedings of the 25th International Database Engineering & Applications Symposium*, ser. IDEAS '21, Montreal, QC, Canada: Association for Computing Machinery, 2021, pp. 75–83, ISBN: 9781450389914. DOI: 10.1145/3472163.3472174. [Online]. Available: https://doi.org/10.1145/3472163.3472174.

[10] H. Cunningham, "Gate, a general architecture for text engineering," *Computers and the Humanities*, vol. 36, no. 2, pp. 223–254, 2002, ISSN: 00104817. [Online]. Available: http://www.jstor.org/stable/30204529 (visited on 01/31/2023).

[11] N. Janpitak, C. Sathitwiriyawong e P. Pipatthanaudomdee, "Information security requirement extraction from regulatory documents using gate/annic," in *2019 7th International Electrical Engineering Congress (iEECON)*, 2019, pp. 1–4. DOI: 10.1109/iEECON45304.2019.8938899.

[12] B. Chandrasekaran, J. Josephson e V. R. Benjamins, "What are ontologies, and why do we need them?" *Intelligent Systems and their Applications, IEEE*, vol. 14, pp. 20–26, Feb. 1999. DOI: 10.1109/5254.747902.

[13] G. Zaman, H. Mahdin, K. Hussain, Atta-Ur-Rahman, J. Abawajy e S. A. Mostafa, "An ontological framework for information extraction from diverse scientific sources," *IEEE Access*, vol. 9, pp. 42 111–42 124, 2021. DOI: 10.1109/ACCESS.2021.3063181.

[14] D. Araujo, S. Rigo e J. Barbosa, "Ontology-based information extraction for juridical events with case studies in brazilian legal realm," *Artificial Intelligence and Law*, vol. 25, Dec. 2017. DOI: 10.1007/s10506-017-9203-z.

[15] M. Buey, A. Garrido, C. Bobed e S. Ilarri, "The ais project: Boosting information extraction from legal documents by using ontologies," Feb. 2016. DOI: 10.5220/0005757204380445.

[16] J. E. F. Friedl, *Mastering regular expressions - understand your data and be more productive: for Perl, PHP, Java, .NET, Ruby, and more (3. ed.)* O'Reilly, 2006, ISBN: 978-0-596-52812-6. [Online]. Available: http://www.oreilly.de/catalog/regex3/index.html.

[17] M. Ruiz, C. Román, A. Garrido e E. Mena, "Uais: An experience of increasing performance of nlp information extraction tasks from legal documents in an electronic document management system," Jan. 2020, pp. 189–196. DOI: 10.5220/0009421201890196.

[18] S. Saha, K. P. Joshi, R. Frank, M. Aebig e J. Lin, "Automated knowledge extraction from the federal acquisition regulations system (fars)," in *2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 3596–3603. DOI: 10.1109/BigData.2017.8258353.

[19] Y.-L. Chen, Y. Liu e W.-L. Ho, "A text mining approach to assist the general public in the retrieval of legal documents," *Journal of the American Society for Information Science and Technology*, vol. 64, pp. 280–290, Feb. 2013. DOI: 10.1002/asi.22767.

[20] S. Choi, S. Choi, J. Kim e E.-B. Lee, "Ai and text-mining applications for analyzing contractor's risk in invitation to bid (itb) and contracts for engineering procurement and construction (epc) projects," *Energies*, vol. 14, p. 4632, Jul. 2021. DOI: 10.3390/en14154632.

[21] Q. Le e T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing e T. Jebara, Eds., ser. Proceedings of Machine Learning Research, vol. 32, Bejing, China: PMLR, 22–24 Jun 2014, pp. 1188–1196.

[22] R. Řehůřek e P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, May 2010, pp. 45–50.

[23] T. Mikolov, K. Chen, G. Corrado e J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Y. Bengio e Y. LeCun, Eds., 2013.

[24] T. Nguyen. "Extract text from pdf resumes using pymupdf and python." (Aug. 2022), [Online]. Available: https://www.neurond.com/blog/extract-text-from-pdf-pymupdf-and-python.

[25] S. G. Stubblebine, *Regular expression pocket reference - regular expressions for Perl, Ruby, PHP, Python, C, Java, and .NET (2. ed.)* O'Reilly, 2007, ISBN: 978-0-596-51427-3. [Online]. Available: http://www.oreilly.com/catalog/9780596514273/index.html.

[26] M. Pasca, D. Lin, J. Bigham, A. Lifchits e A. Jain, "Organizing and searching the world wide web of facts - step one: The one-million fact extraction challenge," in *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06)*, Boston, Massachusetts, 2006, pp. 1400–1405.

[27] M. Honnibal, I. Montani, S. V. Landeghem e A. Boyd, "Spacy: Industrial-strength natural language processing in python," 2020. DOI: 10.5281/zenodo.1212303. [Online]. Available: https://spacy.io/.

[28] T. Wolf, L. Debut, V. Sanh, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[29] S. Tedeschi, V. Maiorca, N. Campolungo, F. Cecconi e R. Navigli, "WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Association for Computational Linguistics, Nov. 2021, pp. 2521–2533.

[30] S. Bird, E. Klein e E. Loper, *Natural Language Processing with Python*. O'Reilly, 2009, ISBN: 978-0-596-51649-9. [Online]. Available: `http://www.oreilly.de/catalog/9780596516499/index.html`.

APPENDIX A

Anexos

ÍNDICE

OFERTA PÚBLICA DE SUBSCRIÇÃO

ÍNDICE ***2***

CAPÍTULO 0 ADVERTÊNCIAS ***7***

CAPÍTULO 1 SUMÁRIO DAS OFERTAS ***15***

CAPÍTULO 2 FATORES DE RISCO ***25***

2.1 Risco relacionado com o contexto económico e financeiro global **26**

2.1.1 Risco associado à pandemia causada pela Covid-19 e possibilidade de futuros focos de doenças semelhantes 26

2.2 Riscos relacionados com a estratégia empresarial da Mota-Engil **27**

2.2.1 *A Mota-Engil está exposta a riscos específicos dos setores e mercados em que atua* 27

2.2.2 A Mota-Engil encontra-se exposta a alterações políticas, sociais, económicas e financeiras, a nível regional e global, e a riscos relacionados com a exposição a países/mercados emergentes (América Latina e África) 29

2.2.3 Riscos relacionados com as operações comerciais do Grupo Mota-Engil 31

2.2.4 Riscos relacionados com a implementação do acordo de investimento com a China Communications Construction Company, Ltd 33

2.3 Riscos relacionados com as operações financeiras do Grupo Mota-Engil **33**

2.3.1 As atividades da Mota-Engil estão sujeitas a risco de crédito 33

2.3.2 O Grupo Mota-Engil encontra-se exposto a riscos de liquidez 35

2.3.3 A exposição do Grupo Mota-Engil ao risco cambial resulta da presença de várias participadas em diversos mercados 37

2.3.4 A dívida financeira do Grupo Mota-Engil encontra-se indexada a taxas de juro cujas variações podem resultar em perdas 39

2.4 Riscos relacionados com o desenvolvimento da atividade da Mota-Engil **39**

2.4.1 *Existem riscos legais decorrentes do exercício da própria atividade do Grupo Mota-Engil* 39

2.4.2 As atividades da Mota-Engil podem ser afetadas pela legislação e regulamentação aplicável nos vários mercados em que opera 41

2.4.3 O desenvolvimento da atividade da Mota-Engil está sujeito a disposições legais relativas à saúde, à segurança no trabalho e a riscos laborais 42

2.4.4 A atividade da Mota-Engil pode ser afetada por incidentes associados a bens físicos e catástrofes naturais 42

2.4.5 O desenvolvimento da atividade da Mota-Engil pode ser afetado por alterações ambientais 42

2.4.6 A atividade da Mota-Engil poderá ser afetada por falhas dos sistemas tecnológicos e da segurança dos sistemas de informação 43

2.5 Riscos relacionados com a Oferta Pública de Subscrição **44**

2.5.1 Risco de crédito do Emitente e Oferente 44

2.5.2 Risco de liquidez inexistente ou reduzida no mercado no qual as Obrigações Mota-Engil 2026 estarão admitidas à negociação 44

2.5.3 Riscos decorrentes das deliberações dos Obrigacionistas e da atuação do representante comum 44

FIGURA A.1. Exemplo de índice com subsecções

TABLE OF CONTENTS

IMPORTANT INFORMATION RELATING TO THE USE OF THIS BASE PROSPECTUS AND THE OFFER OF PUBLIC SECTOR BONDS GENERALLY 7
OVERVIEW OF THE PROGRAMME 12
RISK FACTORS 24
RESPONSIBILITY STATEMENTS 64
DOCUMENTS INCORPORATED BY REFERENCE 70
FORM OF THE PUBLIC SECTOR BONDS AND CLEARING SYSTEMS 74
FINAL TERMS OF THE PUBLIC SECTOR BONDS 78
TERMS AND CONDITIONS OF THE PUBLIC SECTOR BONDS 98
CHARACTERISTICS OF THE COVER POOL 140
INSOLVENCY OF THE ISSUER 146
COMMON REPRESENTATIVE OF THE HOLDERS OF PUBLIC SECTOR BONDS 148
COVER POOL MONITOR 149
DESCRIPTION OF THE ISSUER 151
THE ORIGINATOR'S STANDARD BUSINESS PRACTICES, CREDIT ASSESSMENT AND SERVICING OF THE COVER POOL 175
USE OF PROCEEDS 179
THE PUBLIC SECTOR BONDS LAW 180
TAXATION 188
SUBSCRIPTION AND SALE AND SECONDARY MARKET ARRANGEMENTS 198
GENERAL INFORMATION 207
DEFINITIONS 219
ANNEX – ALTERNATIVE PERFORMANCE MEASURES 234

FIGURA A.2. Exemplo de índice apenas com secções principais

ÍNDICE

CAPÍTULO 1 – SUMÁRIO 10
CAPÍTULO 2 – FATORES DE RISCO DO EMITENTE E DOS VALORES MOBILIÁRIOS A OFERECER E A ADMITIR 35
2.1 Riscos gerais relativos ao Emitente e à sua atividade 35
2.2 Riscos financeiros 41
2.3 Riscos relativos às Obrigações a oferecer e admitir à negociação 48
CAPÍTULO 3 – RESPONSÁVEIS 51
3.1. Identificação dos responsáveis pela informação contida no Prospeto 51
3.1.1. Futebol Clube do Porto – Futebol, S.A.D. 51
3.1.2. Conselho de Administração 51
3.1.3. Conselho Fiscal 51
3.1.4. Intermediário Financeiro encarregado da Assistência à Oferta 52
3.1.5. Revisores Oficiais de Contas e auditores externos 52
3.1.6. Consultor jurídico no âmbito da Oferta 52
3.2. Declaração sobre a informação constante do Prospeto 52
CAPÍTULO 4 – REVISORES OFICIAIS DE CONTAS E AUDITOR EXTERNO 54
CAPÍTULO 5 – DADOS FINANCEIROS SELECIONADOS 55
5.1. Dados Financeiros Históricos 55
5.1.1. Informações Financeiras Consolidadas 55
5.1.2. Dados Financeiros Selecionados 60
CAPÍTULO 6 – ANTECEDENTES, EVOLUÇÃO, INVESTIMENTOS E POLÍTICA DE INVESTIGAÇÃO DO EMITENTE 61
6.1. Antecedentes e Evolução do Emitente 61
6.1.1. Denominação Jurídica e Comercial do Emitente 61
6.1.2. Registo e Número de Pessoa Coletiva do Emitente 61
6.1.3. Constituição do Emitente 61
6.1.4. Sede, Forma Jurídica e Legislação que Regula a Atividade do Emitente 61
6.2. Investimentos e desinvestimentos 64
6.2.1. Investimentos e desinvestimentos 64
6.2.2. Investimentos futuros 68
6.2.3. Financiamento de investimentos futuros 68
6.3. Investigação 69
CAPÍTULO 7 – PANORÂMICA GERAL DAS ATIVIDADES DO EMITENTE 71
7.1. Principais Atividades 71
7.2. Competições desportivas de caráter profissional 79
7.3. Acesso a espetáculos desportivos 82
7.4. Patrocínios e publicidade 82
7.5. Direitos de transmissão televisiva 83
7.6. Sponsorização e merchandising, licenciamento 83
7.7. Realização de eventos não desportivos 84
7.8. Exploração das vertentes de Multimédia e Internet 84
7.9. Intermediação de Seguros 85
7.10. Viagens e Turismo 86
7.11. Audiovisual, Multimédia e Comunicação 86
7.12. Estabelecimentos principais e património imobiliário 88
CAPÍTULO 8 – ESTRUTURA ORGANIZACIONAL 90
8.1. Estrutura Organizacional 90

FIGURA A.3. Exemplo de índice sem separador entre título e página correspondente

CAPÍTULO 1 – SUMÁRIO

Os sumários são elaborados com base em requisitos de divulgação denominados "Elementos". Tais Elementos são numerados em secções de A – E (A.1 – E.7).

O presente Sumário contém todos os Elementos que devem ser incluídos num sumário para o tipo de valores mobiliários e emitente em causa. A numeração dos Elementos poderá não ser sequencial uma vez que há Elementos cuja inclusão não é, neste caso, exigível.

Ainda que determinado Elemento deva ser inserido no Sumário tendo em conta o tipo de valores mobiliários e emitente, poderá não existir informação relevante a incluir sobre tal Elemento. Neste caso, será incluída uma breve descrição do Elemento com a menção "Não Aplicável".

Secção A – Introdução e advertências

A.1	**Advertências**	O presente sumário deve ser entendido como uma introdução ao prospeto (**"Prospeto"**). Qualquer decisão de investimento nas Obrigações deve basear-se numa análise do Prospeto no seu conjunto pelo investidor. Sempre que for apresentada em tribunal uma queixa relativa à informação contida num prospeto, o investidor queixoso poderá, nos termos da legislação interna dos Estados-Membros, ter de suportar os custos de tradução do Prospeto antes do início do processo judicial. Só pode ser assacada responsabilidade civil às pessoas que tenham apresentado o sumário, incluindo qualquer tradução do mesmo, e apenas quando o sumário em causa for enganador, inexato ou incoerente quando lido em conjunto com as outras partes do Prospeto ou não fornecer, quando lido em conjunto com as outras partes do Prospeto, as informações fundamentais para ajudar os investidores a decidirem se devem investir nestas Obrigações.
A.2	**Autorização para ofertas subsequentes**	Não Aplicável. O Emitente não autoriza o uso do Prospeto para a subsequente revenda ou colocação final dos valores mobiliários denominados "Benfica SAD 2018-2021" através de intermediários financeiros.

Secção B – Emitente

B.1	**Denominação jurídica e comercial do Emitente**	Sport Lisboa e Benfica – Futebol, SAD, sociedade aberta (**"Emitente"** ou **"Benfica SAD"**). A denominação comercial utilizada mais frequentemente é Benfica SAD.

FIGURA A.4. Exemplo sumário em tabela com secções identificadas

OVERVIEW OF THE COVERED BONDS PROGRAMME

This overview is qualified in its entirety by the rest of this Base Prospectus.

The following overview does not purport to be complete and is taken from, and is qualified in its entirety by, the remainder of this Base Prospectus and, in relation to the terms and conditions of any particular Tranche of Covered Bonds, the applicable Final Terms. The Issuer and any relevant Dealer may agree that Covered Bonds shall be issued in a form other than that contemplated in the Terms and Conditions, in which event, in the case of listed Covered Bonds, a new Base Prospectus will be published.

This overview constitutes a general description of the Programme for the purposes of the Prospectus Delegated Regulations, notably Article 25(1)(b) of Commission Delegated Regulation (EU) no. 2019/980.

Capitalised terms used in this overview and not otherwise defined below or under the Definitions have the respective meanings given to those terms elsewhere in this Base Prospectus.

Description:	Covered Bonds Programme.
Programme Size:	Up to EUR 12,500,000,000 (or its equivalent in other currencies) aggregate principal amount (or, in the case of Covered Bonds issued at a discount, their aggregate nominal value) of Covered Bonds outstanding at any time. The Issuer will have the option at any time to increase the amount of the Programme, subject to compliance with the relevant provisions of the Programme Agreement.
Issuer:	Banco Santander Totta, S.A. (see *Description of the Issuer*).
Issuer Legal Entity Identifier (LEI):	549300URJH9VSI58CS32
Arranger:	Barclays Bank Ireland PLC
Dealers:	Banco Santander, S.A., Banco Santander Totta, S.A., Barclays Bank Ireland PLC, BNP Paribas, BofA Securities Europe SA, Credit Suisse Bank (Europe), S.A., Deutsche Bank Aktiengesellschaft, Goldman Sachs Bank Europe SE, HSBC Continental Europe, Morgan Stanley Europe SE, Natixis, Société Générale, UBS Europe SE, UniCredit Bank AG and any other Dealer(s) appointed from time to time by the Issuer in accordance with the Programme Agreement and excludes any entity whose appointment has been terminated pursuant to the Programme Agreement.

FIGURA A.5. Exemplo sumário em modo texto