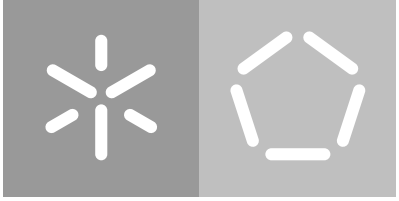
Universidade do Minho
Escola de Engenharia

Joel Soares Rodrigues

Modelação e previsão de decisões judiciais utilizando um repositório de sentenças

Dezembro, 2021



Universidade do Minho
Escola de Engenharia

Joel Soares Rodrigues

Modelação e previsão de decisões judiciais utilizando um repositório de sentenças

Dissertação de Mestrado

Mestrado Integrado em Engenharia Informática

Trabalho efetuado sob a orientação do(a)

Paulo Jorge Freitas de Oliveira Novais

Dezembro, 2021

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Agradecimentos

Chegada a fase final do percurso académico, ao analisar em retrospetiva os cinco anos do curso, são várias as pessoas a quem tenho de agradecer já que contribuíram, de uma forma ou de outra, para que esta etapa da minha vida fosse concluída com sucesso.

Aos meus orientadores, professor Paulo Novais e professor Marco Gomes, por aceitarem orientar-me e por sempre mostrarem muita disponibilidade ao longo da realização deste trabalho.

Aos meus pais por me proporcionarem todas as condições para eu poder estudar durante estes cinco anos.

Ao meu irmão por estar sempre disponível e por me apoiar em todas as fases da minha vida.

À minha namorada, Daniela, por ter sido um pilar fundamental na minha vida académica, tanto nos bons como nos maus momentos.

Aos amigos que fiz durante este percurso académico. Estiveram sempre ao meu lado desde o início e por isso também desempenharam um papel importante durante estes cinco anos.

Por último, e não menos importante, agradecer aos meus amigos fora do âmbito académico, esses estão ao meu lado desde a minha infância e por isso também foram fundamentais na conclusão do curso.

Resumo

O ritmo da evolução tecnológica e a sua relação com o ser humano tem aumentado significativamente ao longo dos tempos, sendo que um dos ramos que mais impacto está a causar no quotidiano das pessoas é a inteligência artificial. A grande ascensão desta área é um fenómeno transversal a praticamente todos os setores da sociedade. Esta dissertação enquadra-se no setor da justiça.

Ao longo dos anos, um dos principais problemas nos sistemas judiciais por todo o mundo é a morosidade na resolução dos processos judiciais. Tendo por base esta problemática, as entidades governamentais adotam cada vez mais reformas na área da justiça com recurso à tecnologia, desejando sistemas judiciais cada vez mais eficientes.

Neste sentido, esta dissertação tem como objetivo o desenvolvimento de uma solução capaz de extrair conhecimento a partir de dados jurídicos portugueses. Esta solução é caracterizada por um conjunto de mecanismos, com recurso a técnicas de inteligência artificial, que vai desde a extração de dados até à realização de duas grandes experiências: análise de sentimentos e previsão da decisão. Estes mecanismos permitiram gerar diversas informações e conhecimento. Por um lado evidenciou-se a pouca relação entre a carga emocional dos textos dos juízes e a decisão, por outro destaca-se o desenvolvimento de modelos inteligentes capazes de prever a decisão com precisões de média de 76%, recorrendo ao conteúdo textual. Além disso, ao longo de todo o processo, também foram extraídas outras informações, como por exemplo, as palavras relacionadas com a procedência e improcedência de acórdãos, a legislação que é geralmente citada em conjunto, entre outras. Em paralelo desenvolveu-se um protótipo de uma *dashboard* com a apresentação de informações e conhecimento de alto nível sobre os dados.

Palavras-chave: Justiça, Decisões Judiciais, Inteligência Artificial, Processamento de Linguagem Natural, *Machine Learning*

Abstract

The pace of technological evolution and its relationship with the human being has increased over time. One of the branches causing the most significant impact on people's daily lives is artificial intelligence. The remarkable rise in this area is a phenomenon that affects practically all sectors of society, but this project fits in the area of justice.

Over the years, one of the significant problems in all countries' judicial systems is the delay in resolving judicial processes. Based on this problem, governmental entities have been adopting more and more reforms in justice linked to technology, with a view to increasingly efficient judicial systems.

This master's dissertation aims to build up a solution capable of extract knowledge from Portuguese juridical data. This solution is characterized by a group of mechanisms that use artificial intelligence techniques that go from the data extraction until the implementation of two experiences: sentimental analysis and predict judicial decisions. These mechanisms allow generating information and knowledge. On the one hand, there was a lack of connection between the emotional side of the text from the judges and the decision of the verdict. On the other hand, an important matter was developing the innovative models capable of predicting with a precision of 76%, using textual content extracted. Besides that, throughout the entire process, another type of information was removed, for example, the words relate to the court decisions, the legislation that was most of the time quoted together, among others. At the same time, was developed a dashboard prototype with the presentation of the high-level information and the knowledge of the data.

Keywords: Justice, Judicial Decisions, Artificial Intelligence, Machine Learning, Natural Language Processing

Índice

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O presente capítulo tem como objetivo apresentar o enquadramento e a motivação para a realização deste projeto de dissertação, nomeadamente uma contextualização de um ponto de vista judicial. Para além disso, identifica-se o principal problema que se pretende resolver com este projeto, a questão de investigação à qual se procura obter uma resposta e uma visão dos objetivos a serem alcançados. O capítulo termina com a apresentação da estrutura do documento.

1.1 Contextualização e motivação

Um sistema judicial é caracterizado pelo conjunto de tribunais e autoridades judiciárias de um país ou organização soberana, que têm como principal compromisso resolver os litígios e garantir a aplicação da lei de uma forma correta e coerente. Um sistema judicial é um pilar fundamental numa sociedade e, portanto, existe uma relação direta entre um sistema judicial e o desenvolvimento social, político e económico de um país. No que toca ao sistema político-social, a justiça apresenta um papel essencial, visto que propicia a transparência e credibilidade pública que são duas características muito importantes em ambientes democráticos. Relativamente ao desenvolvimento económico, ter uma justiça funcional, eficaz e robusta aumenta a confiança e credibilidade dos cidadãos e propicia o ambiente de investimento [49]. Os credores têm maior probabilidade de emprestar dinheiro, os custos das transações são reduzidos e as novas empresas são mais propensas a captar investimentos estrangeiros [18].

Neste sentido, existem vários estudos e literatura científica que correlacionam os sistemas judiciais

aos económicos. Por exemplo, em 2017, Vincenzo e Leandro concluíram que, no que toca à duração de resolução de processos, uma redução de 1% permite aumentar em 0,04% a taxa de crescimento do número de empresas [27]. Outro estudo, elaborado por Dogru em 2012, estabelece uma correlação positiva entre a independência de um sistema judicial e o investimento estrangeiro na Europa Central e Oriental [12].

Contudo, avaliar um sistema judicial é uma tarefa complexa, por esse motivo, a Comissão Europeia estabeleceu três parâmetros de forma a comparar os sistemas judiciais dos vários Estados-Membros: a qualidade, a eficiência e a independência. A qualidade está relacionada com vários fatores: a acessibilidade da justiça aos cidadãos e empresas; os recursos financeiros e humanos adotados; a criação de instrumentos de avaliação; a aplicação de normas de qualidade. Por outro lado, a independência é avaliada tendo em conta a autonomia e a imparcialidade de cada sistema judicial. Finalmente a eficiência, é o parâmetro que está, efetivamente, relacionado com a problemática desta dissertação, e diz respeito à morosidade da resolução processual [49].

Neste sentido, para qualificar a eficiência dos sistemas judiciais a Comissão Europeia definiu vários indicadores, nomeadamente o tempo de resolução, que corresponde ao tempo estimado (em dias) necessário para a resolução de um processo em tribunal, a taxa de resolução dos processos, que diz respeito ao rácio entre o número de processos resolvidos e o número de processos iniciados e ainda o número de processos pendentes ao fim de cada ano. Atendendo a estes indicadores, um sistema judicial diz-se eficaz se gerir os processos iniciados e processos em atraso e emitir decisões em tempo útil [49].

Tendo por base todas estas métricas de comparação, anualmente, a Comissão Europeia emite um relatório com estatísticas que permitem conferir o desenvolvimento dos sistemas judiciais de cada país. Mediante estes relatórios é factual que todos os anos na Europa, milhões de processos judicias vêem proferida a sua decisão final, para além de que muitos outros são iniciados e um número elevado deles permanece no estado de pendência. De facto, estes últimos números são bastantes elevados e embora variem consideravelmente entre os vários Estados-Membros da Europa, centram grande parte da atenção das entidades competentes para assegurar a eficácia dos diversos sistemas judiciais europeus [49].

Enquadrando o país nesta matéria, Portugal encontra-se num patamar intermédio relativamente aos outros Estados-Membros da Europa. Em 2019, a Direção Geral de Política da Justiça (DGPJ) divulgou informação estatística sobre tribunais judiciais portugueses, apresentado uma taxa de resolução processual na ordem dos 124,61% e um tempo estimado para concluir um processo de, aproximadamente, 30 meses. Já no que toca à problemática da pendência de processos, Portugal segue a tendência da Europa uma vez que o número de pendências judiciais tem sido bastante elevado ao longo dos anos, apesar de em 2018 se verificar a mais baixa pendência desde 1996, com uma descida de 35% entre 2015 e 2018

(menos de 463.273 processos), principalmente devido ao elevado movimento de processos de matéria fiscal, que representavam uma grande quantidade dos processos findos nesses anos. [42].

Essencialmente é nestes números que as entidades governamentais se baseiam para definir as suas reformas no área da justiça, perspetivando sistemas judiciais cada vez mais eficientes. São várias as áreas onde as reformas judiciais se fazem sentir, no entanto, para o âmbito desta dissertação destacam-se as medidas que os governos têm tomado relativamente às áreas de Tecnologia da informação e Comunicação (TIC). Existe uma preocupação em investir nestes setores, pois acredita-se que aliar a justiça à modernização, nomeadamente em utilizar mecanismos diferentes daqueles usados até então, conduzirá os sistemas a uma maior produtividade. Portugal, tem adotado várias medidas neste sentido, que segundo o relatório de Justiça de 2019, o trabalho da modernização tem sido reconhecido mundialmente, sendo já elogiado pela Organização para a Cooperação e Desenvolvimento (OCDE) [44].

De todas as áreas que relacionam a tecnologia com a justiça aquela que mais se destaca é a Inteligência Artificial (IA), que embora em alguns países os sistemas inteligentes sejam apenas uma questão emergente e sem usabilidade, noutros já existem abordagens e leis para a aplicação dos mesmos, pois estes sistemas trazem grande impacto na eficiência dos sistema judiciais, nomeadamente na rapidez de resolução processual, na organização de informação e na aquisição de conhecimento sobre a grande quantidade de dados, que é produzida todos os dias [40]. Esta dissertação enquadra-se nesta relação, IA-justiça, no qual se pretende atingir todas estas vantagens, numa área ainda pouco explorada em Portugal.

Neste sentido, a problemática à qual se pretende apresentar uma solução com este projeto consiste na extração de conhecimento a partir de dados jurídicos portugueses semiestruturados. Não obstante, este problema principal é desdobrado em problemas de menor dimensão ao longo do projeto, nomeadamente, na estruturação e preparação dos dados para a aplicação de técnicas de inteligência artificial, na aplicação propriamente dita de mecanismos de extração de conhecimento e na representação deste conhecimento.

Do ponto de vista académico, com este projeto de dissertação pretende-se aplicar todas as competências de cariz técnico e de investigação, adquiridas e desenvolvidas ao longo do curso.

1.2 Questão de investigação

O primeiro passo num projeto deste âmbito passa pela formulação de uma questão de investigação que permita a definição de linhas de investigação e desenvolvimento do projeto.

Neste caso pretende-se dar resposta à seguinte questão de investigação: "É possível um sistema, com recurso a mecanismos inteligentes, gerar informação e conhecimento útil a partir de dados jurídicos

portugueses?".

Deste modo, com o decorrer do projeto pretende-se reunir evidências, sejam elas corroboradas pela investigação ou pelos resultados obtidos com os mecanismos desenvolvidos, que permitam obter uma resposta para esta questão.

1.3 Objetivos

Neste sentido, de forma a responder à questão de investigação e considerando as sentenças portuguesas disponíveis nas base de dados jurídicas disponibilizadas *online*, pretende-se desenvolver um conjunto de mecanismos baseados em dados para extrair, armazenar, gerir e processar as informações jurídicas e contextuais necessárias para a modelação e previsão de decisões judiciais. Assim, prevê-se o desenvolvimento dos seguintes objetivos:

- A agregação de dados jurídicos portugueses, semiestruturados, em larga escala;
- A aplicação dinâmica de algoritmos analíticos e de previsão para desvendar padrões que direcionam as decisões judiciais;
- A análise e compreensão das variáveis que estão relacionadas com as decisões judicias;
- A visualização dos dados através de um painel de controle, apresentando um resumo de alto nível dos dados e do conhecimento extraído. Para além disso, pretende-se que seja uma ferramenta de otimização de tarefas diárias recorrentes dos profissionais da justiça.

1.4 Organização do documento

Este documento está dividido em cinco capítulos. No primeiro e presente capítulo é introduzido e contextualizado o problema, apresentada a questão de investigação e os objetivos que se pretendem alcançar. O segundo capítulo diz respeito à revisão da literatura, apresenta-se um estudo sobre o estado da arte do tema deste projeto e abordam-se alguns conceitos que sustentam a solução proposta. Segue-se o capítulo do desenho e implementação da solução, onde se apresenta todo o trabalho prático desenvolvido, passando pelos processos de extração, análise, preparação e extração de características dos dados. Para além disso apresentam-se os cenários de dados construídos e as experiências desenvolvidas. No capítulo quatro, são apresentados e discutidos os resultados obtidos com as experiências anteriores. Finalmente, no último capítulo, tecem-se as conclusões consoante os objetivos inicialmente propostos e os possíveis trabalhos futuros desta dissertação.

Capítulo 2

Revisão da literatura

A revisão da literatura é a tarefa responsável pelo levantamento exaustivo e avaliação crítica da literatura relacionada com uma determinada área ou tema. É na revisão da literatura que também se apresenta e analisa o estado da arte relacionado com o tema do projeto.

No contexto desta dissertação, para a realização da revisão da literatura, recorreu-se, principalmente, a plataformas *online* que disponibilizam artigos, livros e dissertações, nomeadamente à *Ieee Xplore*, *Google Scholar*, *PeerJ Computer Science*, *RepositóriUM*, *Plos One* e *Researchgate*. Para além disso ainda se exploraram várias páginas *Web* que contêm conteúdo relevante no contexto deste problema.

Outro aspeto importante a ter em conta nesta tarefa é a qualidade do conteúdo selecionado. Uma revisão da literatura é tão boa quanto mais útil for o material. Neste sentido, adotaram-se várias estratégias para avaliar o material recolhido: a data de publicação, com a preocupação de procurar documentos atualizados, a relevância científica, tendo em conta o número de citações do documento e por fim analisando o conteúdo do documento.

A presente revisão da literatura começou por um breve estudo da justiça em Portugal, nomeadamente uma análise da organização judiciária, da hierarquia e funcionamento dos tribunais portugueses. É de referir que, neste caso, recorreu-se a um profissional da área de direito para o auxílio na seleção e avaliação do material usado. Seguiu-se uma análise das áreas resultantes da relação entre a inteligência artificial e a justiça, assente em duas vertentes: uma mais genérica com a apresentação de projetos e empresas com atividade nesta relação e, posteriormente, uma investigação dos trabalhos relacionados com uma das áreas principais desta dissertação, a justiça preditiva, onde são apresentados vários casos de estudo.

Finalmente procedeu-se a uma revisão dos conceitos tecnológicos relacionados com o tema deste projeto, partindo-se do mais genérico, descoberta de conhecimento em bases de dados até ao mais específico, processamento de linguagem natural.

2.1 Organização judiciária portuguesa

Segundo o artigo 202º da Constituição da República Portuguesa (CRP), os tribunais são "*os órgãos de soberania com competência para administrar a justiça em nome do povo (...)*". A definição é bem clara, é nos tribunais que é aplicada a lei, através do julgamento de provas apresentadas pelas partes e sempre com o propósito de alcançar a verdade.

A hierarquia dos tribunais portugueses é uma estrutura complexa e por isso, existe a organização judiciária, a disciplina que estuda a estrutura e funcionamento dos órgãos que compõem o poder jurisdicional na realização das seguintes funções: a categorização dos tribunais portugueses, as sedes e as áreas onde exercem o poder jurisdicional; a divisão territorial de atuação de cada tribunal; determinação da relação entre os tribunais, entre os tribunais e os outros órgãos de soberania (Presidente da República, Assembleia da República e Governo) e entre os juízes de cada tribunal; a hierarquização dos tribunais tendo em conta os seus recursos; e por fim a atribuição e divisão de competências a cada tribunal [5].

Em Portugal, são várias as categorias de tribunais existentes, com várias divisões e graus de decisão judicial e organizados numa estrutura robusta, como esquematizado na Figura 2.1.No contexto desta dissertação é essencial ter um conhecimento sobre esta divisão hierárquica, de forma a perceber a origem e contexto dos dados que serão alvo de estudo.

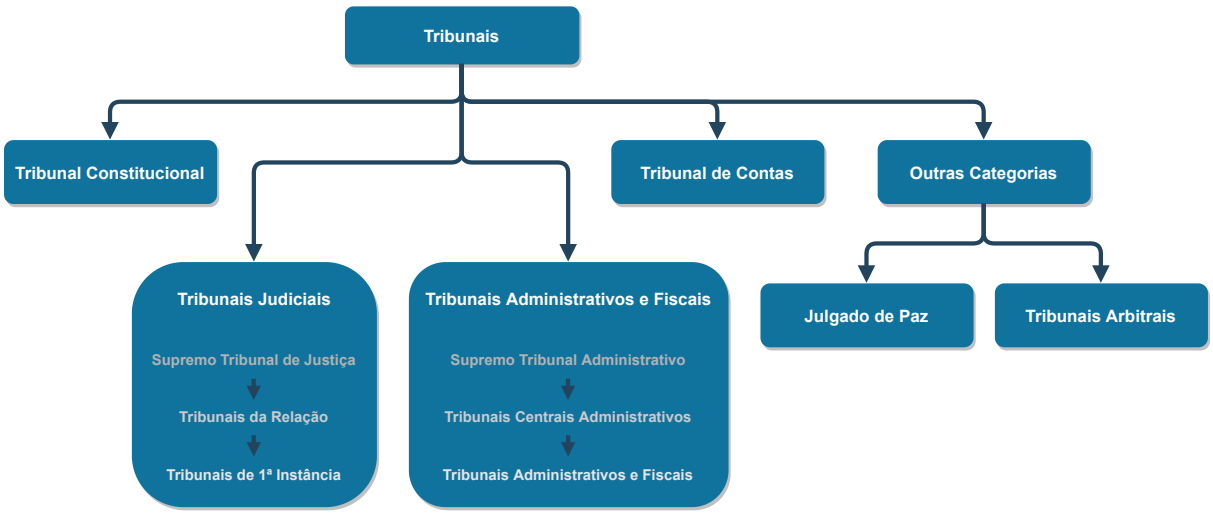


Figura 2.1: Categorização e hierarquia dos Tribunais. Adaptado de [9].

O **Tribunal Constitucional** tem como competência principal *"administrar a justiça em matérias de natureza jurídico-constitucional"* segundo o artigo 30° da Lei da Organização do Sistema Judiciário (LOSJ) [15]. Este órgão de soberania tem características semelhantes aos os outros tribunais, ainda assim diferencia-se em alguns aspetos, um deles é que grande parte dos juízes são escolhidos pela Assembleia da República. Este tribunal tem como principal objetivo fiscalizar a constitucionalidade de legislações aprovadas pelo Parlamento ou pelo Governo e por isso é o órgão de soberania que garante o cumprimento da Constituição.

Segundo o artigo 149° da LOSJ ao **Tribunal de Contas** compete a *"fiscalização da legalidade das receitas e das despesas públicas e de julgamento das contas públicas"* [15]. Além desta fiscalização, cabe a este tribunal julgar as infrações financeiras que envolvam dinheiros públicos.

Os **Tribunais Judiciais**, conforme o artigo 40° da LOSJ, têm a *"competência para as causas que não sejam atribuídas a outra ordem jurisdicional."* [15]. Estes tribunais são organizados segundo uma hierarquia para efeitos de recurso, isto é, qualquer cidadão ou instituição que não concorde com a decisão de um tribunal judicial, tem a oportunidade de recorrer da sentença para um tribunal superior, onde o processo é reapreciado. Deste modo, um cidadão para resolver um conflito, numa fase inicial, recorre a um tribunal de primeira instância, também chamado tribunal da comarca[1]. É nestes tribunais que ocorre o ponto de partida de um processo judicial. Interessa também saber que, os tribunais da comarca ainda são sujeitos a uma divisão segundo as suas competências, uns de competência mais genérica, outros mais específica (Ex. Tribunais de Trabalho, de Comércio, Família, etc...). Numa fase posterior, o cidadão pode recorrer da decisão para o tribunal de relação que está hierarquicamente definido, estes também são chamados de tribunais de segunda instância. Em Portugal, existem 5 tribunais da relação, Tribunal da Relação de Lisboa, Coimbra, Porto, Évora e Guimarães. A sua principal tarefa é, portanto, julgar os recursos interpostos. Finalmente, no topo da hierarquia dos tribunais judiciais, encontra-se o Supremo Tribunal de Justiça, responsável por julgar os recursos dos tribunais da relação. Este é o último órgão onde um cidadão pode recorrer de uma decisão tomada num tribunal judicial. Destaca-se que qualquer tribunal ao qual se recorreu e portanto, reapreciou o processo, pode tomar duas posições: 1) confirmar a decisão anteriormente tomada; 2) revogar a decisão que foi objeto de recurso [5, 9].

Por fim, aos **Tribunais Administrativos e Fiscais**, de acordo com o artigo 212° da CRP *"compete o julgamento das acções e recursos contenciosos que tenham por objecto dirimir os litígios emergentes das relações jurídicas administrativas e fiscais."* Ou seja, estes tribunais têm como objetivo tratar das relações entre os particulares e a Administração Pública. Estão organizados segundo a hierarquia apresentada na Figura 2.1, sendo o Supremo Tribunal Administrativo que decide os recursos expostos nos outros tribunais

[1] Chama-se comarca à divisão do território que está sob o âmbito de funções de um tribunal judicial.

de nível inferior [9].

2.2 Inteligência artificial na justiça

O crescente desenvolvimento da relação entre a IA e a justiça é tão evidente que, já levou a Comissão Europeia para a Eficiência da Justiça (CEPEJ) a abordar esta questão, lançando em 2018, uma carta ética europeia sobre o uso da IA na justiça [37]. A CEPEJ menciona que a utilização de instrumentos e serviços de IA nos sistemas judiciais, nomeadamente, no tratamento de decisões e dados judiciais deve seguir os seguintes princípios:

- **Princípio do respeito dos direitos fundamentais:** assegurar que a elaboração e a implementação de instrumentos e serviços de inteligência artificial respeitem os direitos fundamentais. Uma aplicação prática deste princípio é, por exemplo, quando os sistemas inteligentes são usados para resolver um caso, apoiar a tomada de decisão ou orientar o utilizador. Nestes casos, a garantia de acesso a um juiz e de um julgamento justo, não pode ser posta em causa [37].

- **Princípio da não discriminação:** prevenir especificamente o desenvolvimento ou a intensificação de qualquer discriminação entre indivíduos ou grupos de indivíduos. É conhecido o alto desempenho destes sistemas em encontrar padrões nos dados, nomeadamente, no agrupamento ou classificação de dados relativos a grupos de indivíduos. Portanto, estes mecanismos inteligentes, não devem reproduzir ou agravar as evidências discriminatórias entre pessoas [37].

- **Princípio da qualidade e da segurança:** no que diz respeito ao tratamento de decisões e dados judiciais, utilizar fontes certificadas e dados intangíveis com modelos concebidos de forma multidisciplinar, num ambiente tecnológico seguro. Isto é, os dados utilizados na construção desses sistemas têm de partir de fontes certificadas e não devem ser modificados. Para além disso, os modelos e algoritmos devem permanecer em locais seguros de forma a manter a integridade e segurança do sistema [37].

- **Princípio da transparência, imparcialidade e equidade:** tornar os métodos de tratamento de dados acessíveis e compreensíveis e ainda autorizar auditorias externas. Os sistemas devem ser o mais transparente possíveis. Têm também de permitir a auditoria das autoridades competentes, de forma a certificarem o produto [37].

- **Princípio "controlo do utilizador":** impedir uma abordagem prescritiva e garantir que os utilizadores sejam agentes informados e controlem as suas escolhas. O objetivo é que a autonomia dos

> utilizadores seja aumentada e não restringida com estes serviços. Devem poder rever as decisões tomadas e serem informados da sua origem [37].

De facto, por todo o mundo, são várias as empresas e países que desenvolveram e utilizam sistemas que aplicam a IA em várias áreas da justiça. Esta aplicabilidade pode assentar em diferentes áreas, como por exemplo, motores de pesquisa de jurisprudência avançada, resolução de processos *online* também designada Online Dispute Resolution (ODR), ferramentas de auxilio e elaboração de escrituras, análise preditiva também chamada de justiça preditiva, categorização e análise de contratos/processos/documentos e até *chatbots* para fornecer apoio jurídico a cidadãos ou auxiliar os advogados e juízes nas tarefas do seu dia-a-dia [47].

Nos Estados Unidos o investimento e aplicação destes serviços é grande, já na Europa as iniciativas partem, essencialmente, por empresas privadas, o que torna a sua consolidação mais demorada [37]. Ainda assim, já se encontram bastantes sistemas implementados, como por exemplo, na área da análise de documentos, o *software Luminance*[2], uma ferramenta de análise de documentos para advogados que entre inúmeras funcionalidades, permite a visualização e agrupamento simples dos dados e ainda a deteção de anomalias nos documentos. Todas estas funcionalidades utilizam algoritmos de Machine Learning (ML). Segundo o site da própria empresa, os seus sistemas conseguem reduzir em cerca de 85% o tempo de revisão por processo [3] [33, 47].

No Reino Unido, já existem ferramentas com o objetivo de prever o risco de um cidadão voltar a cometer um crime num determinado período de tempo, como é o caso da ferramenta *HART (Harm Assessment Risk Tool)*. Por exemplo, na Universidade de Cambridge foi desenvolvido um sistema, utilizando a *HART*, baseado num historial de cerca de 104.000 cidadãos presos em *Durham*, contendo um conjunto vasto de variáveis, como o historial de ofensas, idade, género, entre outras, para prever se num prazo de 2 anos, existe grande probabilidade de cometer um crime grave, um crime moderado ou não cometer nenhum crime [35].

Na França já são vários os produtos que estão no mercado. Em 2018, a empresa Charpentier Consulting, emitiu uma lista[4] com cerca de 94 empresas francesas com produtos de tecnologia ligados à área da justiça. Um deles, a *Doctrine*[5], uma ferramenta que disponibiliza motores de busca de jurisprudência bastante robustos com inúmeros filtros, que permite, a pesquisa de casos judiciais semelhantes, comparar leis e até oferece funcionalidades de recolha de informações, como por exemplo de noticias jurídicas relacionadas com o processo em questão.

[2] https://www.luminance.com/
[3] Valor calculado tendo em conta os 50 maiores escritórios de advocacia do Reino Unido.
[4] https://media.wix.com/ugd/c21db1_14b04c49ba7f46bf9a5d88581cbda172.pdf
[5] https://www.doctrine.fr/

Na área de ODR, são vários os países que já adotaram sistemas deste género, Holanda, Reino Unido, Letónia e Estónia são alguns deles. Uns mais automatizados do que outros, uns com mais supervisão do que outros e uns que resolvem litígios mais complexos que outros, mas todos com o objetivo de facilitar o acesso à justiça, economizar e aumentar a rapidez de resolução processual [47]. Em 2015, no Reino Unido, o Civil Justice Council (CJC) emitiu um relatório técnico a abordar esta questão e apresentou várias empresas e sistemas que já usam mecanismos deste género, tanto no seu país como no resto do mundo. Por exemplo, a empresa *eBay*, todos os anos resolve cerca de 60 milhões de casos de falta de pagamentos dos consumidores e de produtos entregues que não correspondem ao produto comprado, tudo recorrendo a uma ferramenta ODR [19]. Já na Holanda, o *e-Court*, um serviço desenvolvido por empresas privadas que resolve na integra um processo judicial. Ou seja, a decisão é calculada automaticamente na íntegra, depois é enviada para um tribunal público onde os juízes analisam manualmente os resultados e emitem a decisão final [47]. Ainda na Europa, a Comissão Europeia disponibiliza uma plataforma de ODR para resolução de litígios sobre casos de compras *online*, absolutamente regulada segundo normas Europeias[54].

Os avanços tecnológicos são maiores nos Estados Unidos, onde já existem ferramentas informáticas para calcular as probabilidade de um cidadão voltar a cometer um crime e até algoritmos a decidir, automaticamente, processos sobre infrações de trânsito. Destaca-se o *COIN (Contract Intelligence)*, desenvolvido com IA, que tem o objetivo de analisar documentos, nomeadamente, empréstimos e contratos bancários, de um dos maiores bancos dos Estados Unidos, o *JP Morgan Chase & Co*. Segundo os proprietários, estima-se que esta ferramenta, substitui cerca de 360 mil horas de trabalho ao ano de um advogado e diminui drasticamente os erros humanos normais na atribuição de empréstimos [30].

Em Portugal, começam-se a dar os primeiros passos neste sentido. No âmbito governamental, atualmente existem dois principais projetos desenvolvidos pelo Ministério da Justiça: uma ferramenta na área da pesquisa, permite a rápida recolha de jurisprudência e a obtenção de processos e acórdãos correlacionados. É baseada no serviço *Watson* da empresa *IBM*, um serviço que tem por base de funcionamento a IA; uma outra aplicação, na área do registo de informações geográficas, uma plataforma de identificação automática de proprietários através da matrizes geográficas [39, 46] .

Todos estes produtos resultantes da relação entre a IA e a justiça, têm em comum a constante procura da melhoria da eficiência e da qualidade dos serviços judiciais. Não só permitem otimizar os processos recorrentes dos profissionais de direito, como também organizar, interpretar e manipular os dados produzidos [37].

Ainda assim, analisando-se todas as áreas pertencentes à relação IA-justiça, a mais sensível e que divide mais opiniões é a área do âmbito desta dissertação, a chamada **justiça preditiva**. Entende-se por

justiça preditiva, a capacidade de um sistema determinar a probabilidade de sucesso de um caso futuro tendo por base a análise das decisões tomadas em casos anteriores [32].

Universalmente, não é unânime a sua utilização. Enquanto que para uns, estes sistemas, têm como propósito auxiliar os advogados a antecipar resultados dos casos dos seus clientes, obter padrões nas decisões tomadas pelos tribunais e até auxiliar os juízes na tomada de decisão [37], para outros, o termo "justiça preditiva"é perigoso, põe em causa o direito tal como o conhecemos até hoje e até há, quem já tenha adotado leis para travar o desenvolvimento de sistemas deste tipo. Foi o caso da França, em 2019, o primeiro país do mundo a criar uma lei que pune quem desenvolve sistemas que avaliem, analisem, comparem ou prevejam as práticas profissionais de juízes [48].

Ainda assim, ao longo dos anos, são vários os estudos e trabalhos publicados relacionados com a previsão de decisões judicias. São usadas diferentes metodologias, diferentes interpretações dos dados, diferentes resultados e são várias as conclusões obtidas. Por este motivo, foi necessário efetuar uma análise exaustiva da maior parte dos estudos, de forma a verificar o que foi feito, as lacunas e potencialidade de cada método e o que pode ser modificado e adaptado ao problema desta dissertação.

Em 2016, em Londres, um conjunto de investigadores apresentou um sistema para prever as decisões judiciais do tribunal europeu dos direitos humanos apenas recorrendo a análise textual. O objetivo era prever um valor binário, valor esse que representa se um processo viola ou não os direitos humanos. Para isso usaram uma abordagem baseada em Processamento de Linguagem Natural (PLN), ou seja, os dados de *input* dos modelos são fruto da aplicação de técnicas de PLN nos textos de processos já ajuizados. Como resultados, obtiveram modelos com uma percentagem de 79% de precisão e ainda apresentaram algumas conclusões descritivas dos dados, nomeadamente quais os fatores dos processos que mais importância têm na decisão final [24].

Mais tarde, em 2017, Katz, Bommarito II e Blackman, nos Estados Unidos apresentaram uma abordagem para prever a decisão do supremo tribunal perante as decisões dos tribunais inferiores. Neste caso, definiram que a variável de resposta podia tomar três valores: revertida, confirmada ou outra. Isto é, o supremo tribunal pode reverter ou confirmar a decisão anteriormente tomada e existe o caso quando não se enquadra em nenhuma das anteriores. Para a construção do sistema usaram métodos de árvores de decisão aleatórias, no inglês Random Forest Cassifier (RFC), que, de uma forma geral, definem-se como conjuntos de árvores de decisão. A fonte dos dados é uma base de dados com cerca de 200 anos de casos ajuizados no supremo tribunal. Cada caso contém cerca de 240 meta dados, como por exemplo,

variáveis cronológicas e variáveis especificas de justiça. Como resultados, obtiveram modelos com precisões de cerca de 70%, que, comparados com modelos de base[6], superam em 5% o seu desempenho [31].

Outro exemplo, em 2018, na Tailândia, um grupo de académicos desenvolveu um sistema de previsão de decisões do supremo tribunal da Tailândia em processos criminais, onde o objetivo passa por prever se um pessoa é culpada ou não de um crime. Também ele baseado em dados textuais de processos já ajuizados, sendo que a grande diferença para os anteriores é a introdução de modelos de Deep Learning (DL) na classificação. O algoritmo é composto por 2 vertentes: uma de análise dos factos do caso e outra de análise da legislação associada ao caso, com redes neuronais artificiais mais complexas, chamadas redes recorrentes. Obtiveram resultados em média na ordem dos 66%, sendo que o modelo é mais assertivo quando o suspeito é culpado (74%) [38]. Ainda neste ano, Dietrich, Enos e Sen, no âmbito da previsão da decisão do supremo tribunal dos Estados Unidos, ao invés de usarem dados textuais, utilizaram dados do tom vocal dos juízes durante o julgamento (cerca de 3000 horas de gravações), para prever o que será o voto final do juiz relativamente ao caso. Resumidamente, descobriram que, quão maior é a agitação e excitação emocional do juiz perante um advogado relativamente ao outro, menor é a probabilidade desse advogado ter o voto favorável do juiz [34].

Mais recentemente, em 2020, Strickson e Beatriz, no Reino Unido, desenvolveram um estudo comparativo de várias formas de prever as decisões dos tribunais. Para os testes usaram diversas combinações de variáveis, sempre associadas a mecanismos de PLN aplicados ao texto. Os melhores resultados foram obtidos utilizando um modelo de regressão logística, com dados de treino relativos à frequência das palavras nos textos das decisões anteriores. O melhor modelo obteve uma taxa de precisão de cerca de 68% [52].

De facto, existe um grande *trade-off* na comunidade de direito quanto à utilidade da justiça preditiva. Para o comum cidadão as suas vantagens são evidentes, tais como a transparência, a previsibilidade e normalização das decisões judiciais, uma vez que permitem antecipar uma possível decisão de uma ação judicial e também analisar a discrepância das decisões de processos ajuizados em contextos semelhantes. No entanto, nunca se pode descurar as desvantagens assumidas pelos entendidos do direito. Para estes, as decisões judiciais, não são o resultado de decisões anteriores, existe uma variedade de fatores que influenciam a tomada de decisão, como o contexto jurisdicional, político, social, profissional, família, entre outros. E portanto, a incapacidade de transpor para um algoritmo estes critérios, faz com que este termo não seja assumido de uma forma homogénea por toda a comunidade [32]. É por isso, que esta relação,

[6]Um modelo de base, no inglês *baseline*, é um modelo básico que serve como termo de comparação para outros resultados.

nunca deve ser interpretada como uma disputa entre *Homem vs Máquina*. Os profissionais de direito devem trabalhar em conjunto com os da tecnologia, ambos devem assumir e transformar o modo da existência da lei, compreender as potencialidades e limites associados a estes sistemas [41].

2.3 Descoberta de conhecimento em bases de dados

Estima-se que em 2020 a produção de dados foi cerca de 50 vezes superior a 2009, cerca de 40 triliões de *gigabytes* de dados. Números ainda mais impressionantes são a quantidade estimada de dados que cada pessoa gera por segundo, aproximadamente 1.7 megabytes [16]. De facto, não é por acaso que se apelidou esta era como a era do *Big Data* [10, 16]. No entanto, em paralelo a este crescimento exponencial de dados é essencial ter mecanismos de processamento computacionalmente robustos para tratar essa grande quantidade de dados.

A descoberta de conhecimento em bases de dados, no inglês Knowledge Discovery in Database (KDD), enquadra-se nesta temática, já que tem como objetivo a obtenção de conhecimento útil a partir dos dados. Um conhecimento é considerado útil se for de fácil compreensão, apresentar informação acrescida àquela que já existe sobre os dados, ou se validar/invalidar algum princípio anteriormente proposto. Trata-se de um processo organizado, que pretende identificar padrões nos dados que sejam válidos, novos, úteis e compreensíveis no contexto do problema em estudo. Sendo um processo, é constituído por várias fases, começando no pré-processamento dos dados, passando pela mineração dos mesmos e terminando na avaliação e representação do conhecimento gerado.

2.3.1 Pré-processamento de dados

Uma vez que os dados recolhidos do mundo real são em grande escala e provenientes de fontes heterogéneas, apresentam indiscutivelmente, muitas inconsistências, contêm ”lixo”e são incompletos. Esta fase do processo tem como objetivo preparar, transformar e organizar os dados para serem processados na etapa seguinte.

- **Integração dos dados**

  A fase de integração dos dados caracteriza-se pela recolha de dados de várias fontes, convertendo-os numa única base de dados consistente, permitindo uma visão unificada dos dados. É um processo pesado e interativo devido à heterogeneidade e desestruturação dos dados a recolher. Nesta fase pretende-se detetar e resolver os conflitos entre os dados, nomeadamente as redundâncias

e inconsistências. Para isso é necessário ter um conhecimento profundo das várias fontes, permitindo decidir-se qual a mais fiável, por exemplo, no caso da existência de inconsistências [11, 43].

- **Limpeza dos dados**

  A limpeza dos dados é a fase onde se analisam os dados, removendo ou modificando os que são incorretos, irrelevantes, incompletos, duplicados ou que contém ruído. Neste ponto, os problemas que maioritariamente surgem são relativos aos dados em falta e ao ruído. Relativamente ao primeiro, existem várias técnicas para lidar com esta problemática, por exemplo, remover a entrada nos dados que contém algum campo em falta, preencher o campo em falta com a média global, ou até usar mecanismos matemáticos para calcular o valor mais provável para essa variável. No que toca ao segundo, o ruído, que nada mais é que erros ou variações nos dados incompreensíveis no contexto do problema, as técnicas usadas tem como objetivo suavizar a variância e tratar dos *outliers* [11].

- **Redução dos dados**

  Conjuntos de dados enormes tornam a tarefa de mineração de dados demorada e por vezes inexequível. A redução tem como finalidade produzir uma representação do conjunto de dados reduzida, mas que mantenha a informação e não comprometa a integridade inicial dos dados. Isto é, a obtenção de resultados para um conjunto de dados reduzido deverá ser mais eficiente e produzirá os mesmos resultados, quando obtidos com o conjunto inicial [11]. De entre as inúmeras estratégias para reduzir os dados, destacam-se as de redução de dimensionalidade que, por exemplo, seleccionam o sub-conjunto de variáveis mais relevantes para o problema em estudo e as de compressão que permitem compactar os dados, por exemplo recorrendo-se a dicionários de conversão para reduzir o tamanho de uma variável.

- **Transformação dos dados**

  Por vezes, os dados provenientes do mundo real não apresentam as características mais adequadas para a sua análise. Como tal, a transformação dos dados objetiva a aplicação de várias técnicas para ajustar os dados ao mecanismo de mineração a usar, tornando-o mais eficiente. Existem imensos tipos de transformações possíveis, destacam-se, a normalização, para diminuir a gama de valores, a discretização, para atribuir *labels* a intervalos de valores e até *Smoothing* para remover o "ruído"dos dados. Esta fase é fundamental para o processo de KDD e não existe nenhuma

metodologia predefinida, a aplicação destas técnicas depende sempre de projeto para projeto. [2, 11]

Resumidamente, o pré-processamento, como primeira fase do processo de KDD é essencial para um bom desempenho das tarefas seguintes. Considera-se o estágio mais demorado do processo de KDD, estima-se que 80% do tempo é despendido a tratar os dados [17]. As técnicas do pré-processamento permitem tornar os mecanismos de *Data Mining* mais eficientes já que, ter dados com qualidade é um ponto fulcral para o bom desempenho desses mecanismos.

2.3.2 Mineração de dados

A mineração de dados, mais conhecida no inglês por Data Mining (DM), trata-se da aplicação de diferentes algoritmos para extrair padrões ou modelos dos dados. Muitas vezes, na comunidade científica, este termo é usado, erradamente, como ML, no entanto, o DM contém um conjunto vasto de algoritmos dos quais, os algoritmos de ML são apenas parte integrante (Sec 2.4) [26]. Esta fase inclui a escolha da família dos algoritmos mais adequados para o problema, seguida da seleção do próprio algoritmo, implementação e finalmente a otimização do método selecionado [2].

Existem inúmeros métodos que podem ser aplicados na fase de DM, assim sendo, é essencial ter um conhecimento sobre a diversidade de algoritmos, bem como as diferenças e semelhanças entre eles. A Figura 2.2 representa algumas das várias famílias de algoritmos existentes.

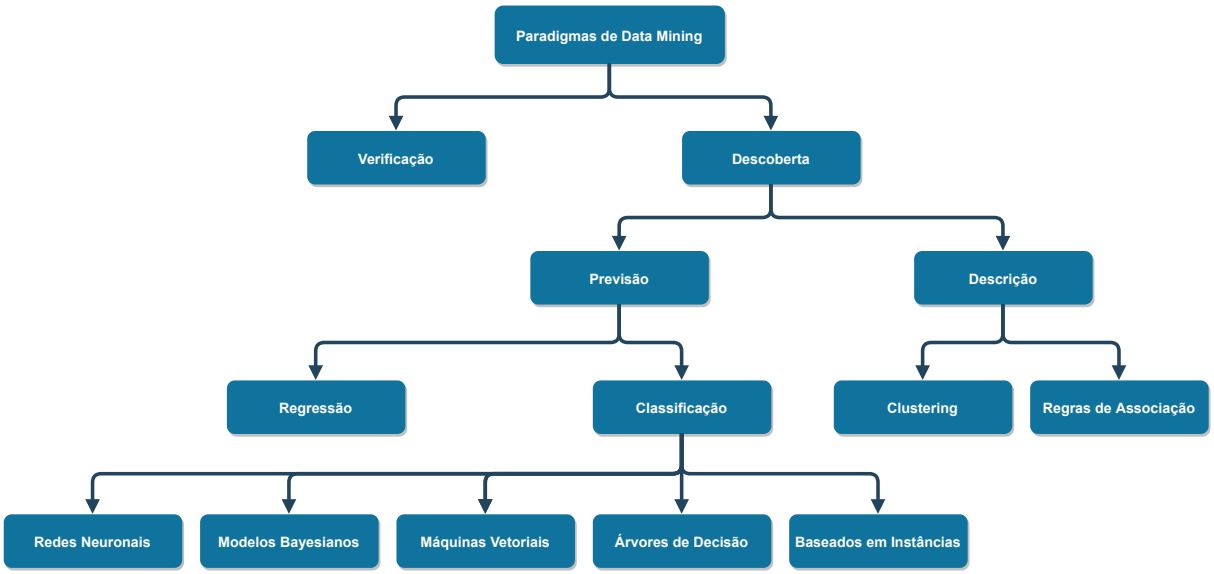


Figura 2.2: Paradigmas de *Data Mining*. Adaptado de [2, 26].

Para além destes, em que a maior parte é habitualmente associada ao ML, o DM usa ainda outras áreas da inteligência artificial para a extração de padrões, como por exemplo a visualização de dados[7].

Numa primeira instância existem dois tipos principais de paradigmas de DM: verificação e descoberta. Quanto ao primeiro, a verificação, tem como objetivo testar hipóteses do utilizador relativamente aos dados, usando diversas técnicas, como *goodness of fit*[8], testes de hipótese e análise de variância. Por outro lado, o segundo, descoberta (Sec 2.3), pretende obter conhecimento útil a partir de grandes conjuntos de dados.

Por sua vez, no processo de KDD e de acordo com o objetivo do problema em estudo existem dois objetivos da mineração de dados: descrição e previsão.

Os métodos de **descrição** são orientados à interpretação dos dados, ou seja, encontrar padrões nos dados que representem conhecimento. Entre vários métodos existentes, destacam-se *Clustering* e Regras de Associação.

Por outro lado, com a **previsão** pretende-se construir padrões relacionais entre os dados (modelos) que permitam prever o valor de alguns parâmetros a partir de outros. É de salientar que, dentro dos problemas de previsão, existem ainda duas subcategorias: regressão e classificação. De uma forma genérica, problemas de regressão caracterizam-se por previsão de valores numéricos, enquanto que na classificação prevêm-se classes para uma determinada amostra [2].

2.3.3 Avaliação e representação do conhecimento

Posteriormente à fase de mineração é necessário efetuar uma avaliação dos padrões e modelos obtidos, bem como representar essa informação. A informação é considerada conhecimento válido se for útil, de fácil compreensão e validar/invalidar algum princípio anteriormente proposto. Evidentemente que o sucesso desta fase depende das fases anteriores, portanto, existe uma interatividade entre as fases anteriores e esta de forma a encontrar conhecimento com utilidade. Este conhecimento pode ser representado sob diversas formas, como por exemplo, apresentação de tabelas, modelos de previsão, relatórios e regras. A partir deste momento, o conhecimento pode ser usado por outros sistemas [2].

[7] As técnicas de visualização de dados têm como objetivo a representação dos dados por via de gráficos. Permitindo, assim, uma fácil interpretação das correlações e padrões entre os eles

[8] O *goodness of fit* de um modelo estatístico representa o quão bem o modelo se ajusta aos dados, ou seja, a discrepância entre os resultados observados e os esperados [7].

2.4 Aprendizagem automática

Em 1997, Tom Michael Mitchell apresenta a seguinte definição de ML no seu livro *"Machine Learning"* [1]:

" A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. "

Sendo assim, entende-se por aprendizagem automática, no inglês ML, a área da IA que confere aos sistemas a capacidade de aprenderem com os dados, sendo estes considerados "experiências passadas".

Os algoritmos de ML podem ser classificados em três grandes paradigmas de aprendizagem:

- **Aprendizagem supervisionada:** Os métodos de aprendizagem supervisionada têm como objetivo encontrar a relação entre os atributos de entrada e um de saída, habitualmente denominado variável dependente. Estes relação resulta num modelo que serve para prever o valor de saída tendo a informação das variáveis de entrada [22]. Sendo assim, nestes sistemas há um conhecimento sobre os resultados esperados e, por isso, é possível interferir-se para obter melhores resultados.

- **Aprendizagem não supervisionada:** Ao contrário do anterior, os modelos de aprendizagem não supervisionada não têm variáveis de saída (variável dependente), portanto o objetivo é encontrar relacionamentos e dependências entre as próprias variáveis de entrada [22].

- **Aprendizagem por reforço:** Neste paradigma não existe o conceito de pares de variáveis de entrada/saída. O objetivo aqui é encontrar a solução para um problema, utilizando para isso um mecanismo de tentativa-erro. Os sistemas possuem uma representação de cada estado possível no ambiente e das ações que podem ser tomadas em cada estado. A partir das ações tomadas e de acordo com as recompensas recebidas, o agente aprimora a ação a escolher em situações futuras semelhantes [1].

Nos seguintes tópicos serão abordados, sucintamente, alguns algoritmos de ML que são comuns na literatura relacionada com esta dissertação.

2.4.1 Máquinas de vetores de suporte

O termo máquina de vetores de suporte, no inglês Support Vector Machine (SVM), refere-se a um algoritmo de aprendizagem supervisionada, tendo por objetivo a obtenção de modelos matemáticos adequados para a resolução de determinados problemas, como a classificação de um *input*, associando-o a uma das classes de dados em estudo [2].

Assim, a solução deste algoritmo será um classificador não-probabilístico, uma vez que a sua utilização apresenta um resultado que não tem associada uma probabilidade. Este classificador é dado na forma de um hiperplano que divide os dados em classes [2].

SVMs são um melhoramento de dois classificadores, que são limitados a dados lineares: *Maximal Margin Classifier (MMC)* e *Support Vector Classifier (SVC)*. As SVMs vêm acrescentar a possibilidade de classificação de *datasets* não-lineares e ainda a distinção de várias classes, ao invés da dualidade de categorização imposta pelos algoritmos anteriores. Apesar de serem muito conectadas a problemas de classificação, as SVMs podem também ser utilizadas em problemas de regressão.

Quanto à metodologia de funcionamento destes algoritmos, os mais simples classificam os dados através de um hiperplano (uma reta), dividindo o espaço em duas categorias. No entanto, as SVMs pretendem classificar dados que não sejam apenas lineares, através da introdução de *kernels*. Estes são funções que atuam sobre os dados originais, criando um espaço em que os dados transformados são divisíveis linearmente por um hiperplano [2]. Voltando ao espaço inicial, este hiperplano irá ser traduzido numa divisão não-linear dos dados, como se pode visualizar através da Fig. 2.3.

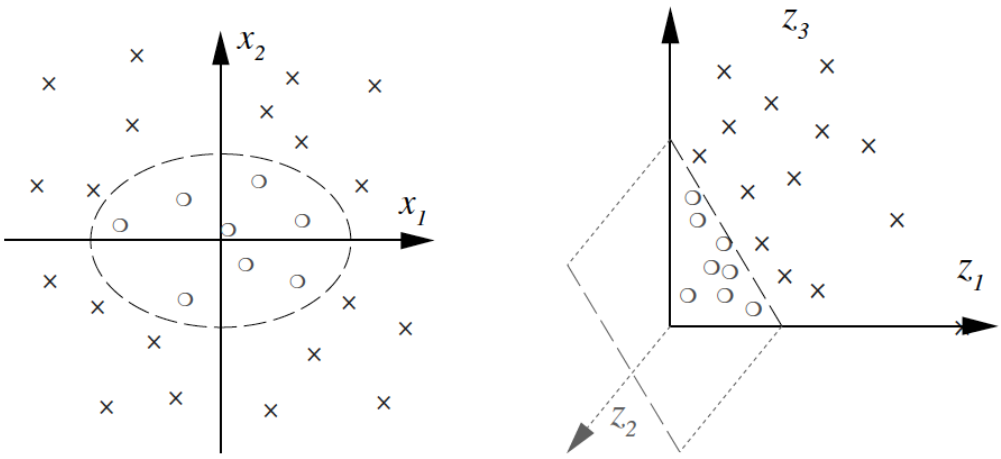


Figura 2.3: Exemplo da aplicação de um *kernel* em dados não-lineares. Fonte: [2].

Para além disto, as SVMs permitem a categorização de dados em mais que duas classes. Para isso, utiliza duas estratégias diferentes: *One-versus-One* e *One-versus-All*. Na primeira, as classes são comparadas aos pares e na segunda estratégia, cada classe é comparada assumindo que todas as outras pertencem a uma só.

2.4.2 Modelos bayesianos

Os modelos bayesianos são caracterizados por métodos de inferência com base no teorema de *Bayes*, apresentando a probabilidade para um novo evento tendo em conta conhecimento relativo a eventos anteriores relacionados.

O teorema de *Bayes* é dado pela seguinte fórmula:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{2.1}$$

em que,

P(B|A): probabilidade de B ocorrer sabendo que A aconteceu;

P(A): probabilidade de A acontecer;

P(B): probabilidade de B acontecer.

Em termos de classificação, estes algoritmos pretendem atribuir, probabilisticamente, uma classe a uma nova ocorrência, tendo em conta todas as classes das ocorrências anteriores.

Dentro dos modelos bayesianos existem várias vertentes, um dos algoritmos mais utilizados é o *Naive Bayes Classifier*, aplicado a problemas onde uma instância (X) é descrita por um conjunto de atributos e a função f(X), para a variável dependente, toma valores finitos. No entanto, este algoritmo adota um teoria mais simplista, assumindo total independência entre os atributos. A aprendizagem é tão melhor quanto maior for o conjunto de dados e mais qualidade estes dados apresentarem [1].

Estes métodos são bastante usados e até, segundo alguns estudos, comparados com outros algoritmos, como por exemplo, árvores de decisão e redes neuronais, apresentam tão bons resultados e, em alguns casos, são mais eficazes [1].

2.4.3 Árvores de decisão

As árvores de decisão são um classificador que pretende representar funções através de uma árvore abstrata. Uma árvore é um grafo acíclico direcional conectado com nós, que implementam testes a um atributo, e com folhas que representam uma classe [26].

Habitualmente o desenvolvimento de uma árvore segue dois passos: a construção e a poda da árvore. No primeiro passo, os algoritmos selecionam o melhor atributo para a raiz, depois dividem os dados em subconjuntos segundo os valores da raiz e assim, recursivamente, até o subconjunto ter elementos só de uma classe. De realçar que a escolha do melhor atributo para cada nó é consoante diversas medidas,

sendo a mais usada a entropia[9]. No passo da poda da árvore, no inglês *Pruning*, o objetivo é simplificar o modelo, removendo alguns ramos que representem lixo ou ruído, aumentando a eficiência do modelo.

Day	Outlook	Temperature	Humidity	Windy	Play Golf?
1	Sunny	85	85	False	No
2	Sunny	80	90	True	No
3	Overcast	83	78	False	Yes
4	Rainy	70	96	False	Yes
5	Rainy	68	80	False	Yes
6	Rainy	65	70	True	No
7	Overcast	64	65	True	Yes
8	Sunny	72	95	False	No
9	Sunny	69	70	False	Yes
10	Rainy	75	80	False	Yes
11	Sunny	75	70	True	Yes
12	Overcast	72	90	True	Yes
13	Overcast	81	75	False	Yes
14	Rainy	71	80	True	No

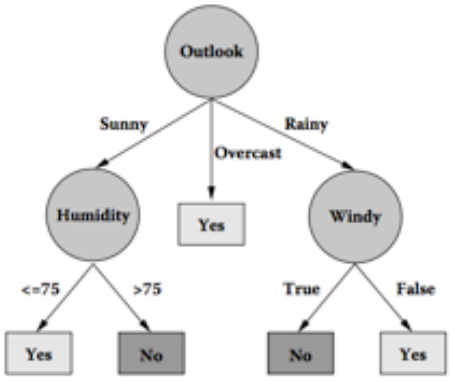


Figura 2.4: Exemplo de uma árvore de decisão construída usando o algoritmo C4.5. Fonte: [8]

Estes métodos são um dos algoritmos mais populares entre os algoritmos de inferência. São usados em diversas áreas, como por exemplo, na avaliação de risco de empréstimos e até no diagnóstico de casos médicos [1].

2.4.4 K-Vizinhos mais próximos

O algoritmo k-Nearest Neighbors (KNN) é um algoritmo baseado em distâncias. Tem como objetivo classificar um novo dado, tendo em conta as semelhanças com os dados existentes. Isto deve-se ao facto deste algoritmo considerar que dados semelhantes estão espacialmente mais próximos do que dados diferentes. Portanto, a classificação de um novo dado é calculada consoante a predominância de uma classe na vizinhança [8].

Existem alguns critérios a ter em conta no desenvolvimento deste algoritmo: 1) a distância ou métrica de similaridade que é usada para calcular a distância entre objetos; 2) o valor de K, ou seja, o número de vizinhos mais próximos que é considerado; 3) o método usado para determinar a classe dos dados de *input* com base nas classes e distâncias dos K vizinhos [8].

De facto, a escolha K é o fator mais importante neste método, e está diretamente relacionado com o desempenho do mesmo. Não existe um número pré-definido que se ajuste a cada problema. Um valor de K pequeno pode enviesar o modelo segundo pontos de ruído, se for grande, a vizinhança pode conter diversas classes e por isso, diminuir a precisão do modelo. Geralmente este valor é calculado com processos de iteração que avaliam o desempenho do modelo, por exemplo, a validação cruzada. Quando o valor de $K > 1$, para selecionar a classe, habitualmente, estabelece-se uma relação entre as distâncias

[9]Grandeza que mede o grau de impureza de um conjunto de dados. É considerado o melhor atributo o que tiver menor entropia.

dos K vizinhos e as classes. Relativamente à distância, na maior parte dos casos é utilizada a distância Euclidiana, representada na seguinte fórmula [8]:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.2}$$

onde, x_i e y_i representam as várias variáveis (atributos) de cada objeto.

2.4.5 Redes neuronais artificiais

As redes neuronais, no inglês Artificial Neural Network (ANN) são modelos computacionais baseados no sistema nervoso central do ser humano. São estruturas extremamente interconectadas de unidades computacionais, designadas neurónios ou nodos, com capacidade de aprendizagem. Estas redes, tal como o cérebro humano apresentam duas capacidades distintas [4]:

- A aptidão para adquirir conhecimento a partir do ambiente, através de um processo de aprendizagem.
- A capacidade para armazenar o conhecimento nos seus componentes, nomeadamente nas conexões dos neurónios.

Na biologia um neurónio é uma célula do sistema nervoso que responde a sinais eletro-químicos, sendo composto por um núcleo, por um conjunto de dendrites que recebem sinais de outros neurónios via sinapses e ainda um axónio que transmite o sinal entre neurónios [4].

Segundo a informática um nodo, ou neurónio artificial, é a unidade mais importante no funcionamento de uma ANN, sendo composto por um conjunto de conexões indexadas com um peso (w_i), podendo estas ter um papel excitatório ou inibitório, um integrador ou valor de ativação ($\sum$), que transforma os vários valores de entrada em um único valor e ainda uma função de ativação (Θ) que tem a função de condicionar o sinal de saída [4].

As funções de ativação são importantes para otimizar o treino das redes neuronais e apresentam diferentes fórmulas e são usadas em contextos diferentes, dependendo do problema em questão. As funções mais usadas são a *ReLU*[10] e *sigmoid*[11] [13].

Para além disso, tal como no cérebro humano, os nodos artificiais são organizados de diferentes formas. Nas ANN as disposições dos nodos podem ser classificadas de três diferentes maneiras:

[10] Fórmula matemática da função *ReLU*: f(x) = $\begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$

[11] Fórmula matemática da função *sigmoid*: f(x) = $\frac{1}{1+e^{-x}}$

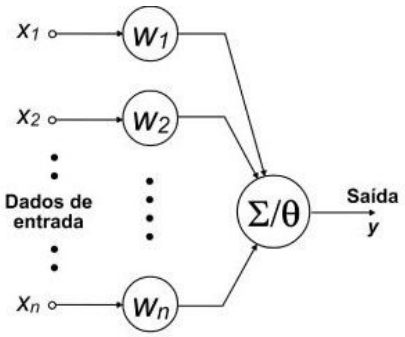


Figura 2.5: Modelo de um nodo artificial. Adaptado de [4].

- **Redes Feedforward de uma só camada:** são organizadas por apenas duas camadas, uma de entrada e uma de saída. Não existem ciclos pois as conexões são todas unidirecionais. Realça-se ainda que apenas são efetuados cálculos na camada de saída [4].
- **Redes Feedforward Multicamada:** ao contrário das anteriores apresentam camadas intermédias que permitem à rede usar funções mais complexas quando o número de nodos iniciais é elevado. Isto acarreta como consequência o aumento do tempo de treino da rede [4].
- **Redes Recorrentes:** de maneira oposta às arquiteturas anteriores, surgem as redes recorrentes que apresentam ciclos, isto é, a saída de alguns nodos pode estimular neurónios da mesma camada, inclusive o mesmo nodo [4].

A propriedade mais importante das ANN é a sua capacidade de aprender a partir do seu ambiente. Esta aprendizagem assenta em 3 passos: a ANN recebe estímulos do ambiente, isto é, recebe dados de treino; os pesos das sinapses são alterados em resultado deste estímulo; e finalmente uma ANN é capaz de responder ao ambiente graças às alterações na sua estrutura interna. Esta aprendizagem é realizada consoante um algoritmo de treino selecionado. Existem diversos algoritmos de treino, o mais usado atualmente é o *back-propagation*, faz parte dos paradigmas supervisionados e baseia-se na alteração dos pesos das sinapses. Este algoritmo utiliza dois passos [13]:

- **Em frente**, onde o vetor de entrada é propagado ao longo da rede até à última camada e no fim é calculado um erro em função do resultado obtido e o desejado. Nesta fase os pesos não são alterados.
- **Retropropagação**, onde o erro é propagado desde a última camada até à inicial segundo algumas funções de alteração dos pesos das sinapses, por exemplo, a mais usada é o gradiente descendente[12].

[12] O gradiente descendente é um método para otimizar a alteração dos valores dos pesos das sinapses que tem o objetivo de encontrar os valores que minimizam o valor da função de erro.

Para o bom desenvolvimento de uma *Rede Neuronal Artificial* é necessário que se faça um estudo com o objetivo de selecionar os melhores métodos a implementar. Ou seja, é fundamental selecionar uma topologia de rede, quantidade de neurónios nas diversas camadas, função de ativação, função de transferência, algoritmos de treino e tarefa de aprendizagem.

2.5 Aprendizagem profunda

Entende-se por aprendizagem profunda, no inglês DL, uma área do ML caracterizada pela maior complexidade dos modelos e capacidade de aprender as representações dos dados. Os modelos de DL são baseados em ANN estruturadas em diversas camadas de processamento, com neurónios e arquiteturas de diferentes tipos. Estes modelos podem ser aplicados em qualquer paradigma de aprendizagem: supervisionado, não supervisionado e por reforço [28].

Os modelos de DL mais populares são as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) e Long Short-Term Memory Networks (LSTM). Atualmente, estes modelos já são aplicados em inúmeras áreas com bons resultados, como por exemplo na classificação de imagem, reconhecimento de texto, transcrição de textos, tradução automática e até em jogos.

2.6 Processamento de linguagem natural

O PLN é um dos termos mais frequentes da literatura relacionada com esta dissertação. Na generalidade, os dados judiciais disponíveis são de tipologia textual e, portanto, a área do PLN é essencial para o desenvolvimento de sistemas que têm por base este tipo de dados.

O PLN caracteriza-se por um ramo das áreas da ciência da computação e inteligência artificial que estuda as formas das máquinas processarem, interpretarem e manipularem a linguagem natural. Entende-se por linguagem natural, ou humana, o tipo de linguagem que nós, humanos, utilizamos no nosso dia-a-dia. As máquinas e algoritmos conhecidos até então, são incapazes de entender texto, portanto, cabe às técnicas de PLN converter estas informações textuais em dados passíveis de serem compreendidos pelas máquinas [45].

Embora o interesse pelo estudo desta área não seja efetivamente recente, os rápidos avanços têm-se verificado nos últimos anos, não só pela era em que vivemos, do *Big Data*, como também pelo crescente interesse na comunicação homem-máquina. Devido a estes grandes avanços, atualmente, são várias as áreas de uso do PLN. Áreas que vão desde a conversão de fala em texto e vice-versa, a tradução de

textos, a sumarização[13], extração de entidades dos textos (nomes próprios, locais, adjetivos, etc.), análise de sentimentos[14], a marcação de classes gramaticais [15](*Part of Speech (POS)*), a extração de tópicos de textos e finalmente, a área desta dissertação, que é a categorização de textos e documentos tendo em conta o seu conteúdo.

O desenvolvimento de um sistema de PLN, geralmente, assente em 3 etapas: o pré-processamento do texto bruto, a extração de variáveis e, por fim, a aplicação de algoritmos, que podem ser de ML, estatísticos, ou simplesmente de análise dos dados.

2.6.1 Técnicas de pré-processamento de texto

O pré-processamento do texto é a primeira etapa num processo de PLN e apresenta um papel indispensável para a obtenção de bons resultados nos estágios seguintes. O objetivo é remover dados irrelevantes e organizar as informações de uma forma mais estruturada que sirva de *input* para as fases seguintes [6, 23]. Existem inúmeras técnicas de pré-processar os dados textuais dependendo dos algoritmos que se pretendem usar posteriormente. A seguir apresentam-se algumas das mais importantes técnicas.

2.6.1.1 Tokenização (*Tokenization*)

A tokenização é um processo de normalização do texto, fundamental em qualquer mecanismo de PLN. Tem como objetivo separar um texto em blocos de símbolos, palavras, frases ou outros elementos textuais mais pequenos, aos quais se dá o nome de *tokens* [45]. A dificuldade da tokenização difere de linguagem para linguagem, tendo em conta as suas características. Na generalidade dos casos, como acontece com as linguagens europeias, as palavras são delimitadas por espaços, e portanto usa-se esse caracter para efetuar a divisão dos *tokens*. No entanto, em linguagens não segmentadas, como a chinesa, requerem mais informações para se efetuar este processo [6].

Exemplo: A frase "Como está o tempo hoje?"depois de tokenizada resultaria na seguinte lista de *tokens*:

["Como", "está", "o", "tempo", "hoje", "?"]

[13]Uma técnica que permite efetuar resumos de textos ou documentos, as principais abordagens passam pela descoberta das palavras-chaves dos textos e organiza-las [6].

[14]O principal objetivo das técnicas desta área, são a categorização de textos ou documentos tendo em conta a sua carga emocional, negativa, positiva ou neutra [45]

[15]Uma técnica caracterizada pela atribuição de rótulos às palavras contendo as suas classes gramaticais, o objetido é diminuir a ambiguidade normal das línguas. Por exemplo, uma palavra numa frase pode ser um verbo e noutra pode ser um substantivo [45].

2.6.1.2 Remoção de palavras irrelevantes (*Removing Stop Words*)

Este método trata-se de remover do texto aquelas palavras que têm muito pouco significado, que são, consequentemente, as palavras mais frequentes numa determinada língua [45]. Estas palavras não apresentam valor semântico no texto onde estão inseridas, e por esse motivo, não têm qualquer importância na classificação e extração de informações dos dados textuais. Geralmente, para se proceder à remoção das palavras irrelevantes, recorre-se a listas genéricas de *"stop words"* disponíveis de cada língua, no entanto, é necessário efetuar uma análise prévia das listas, já que dependendo do contexto do problema, tem de se remover ou adicionar palavras a essas listas [29]. Este processo é importante, pois permite diminuir significativamente o conjunto de dados (estima-se que entre 30%-50%) e, por isso, aumentar o desempenho dos algoritmos posteriormente aplicados [6].

Exemplo: A frase "Hoje está um dia de sol."sujeita ao processo de tokenização e remoção das palavras irrelevantes resultaria na seguinte lista de *tokens*:

["Hoje", "dia", "sol", "."]

2.6.1.3 Remoção de pontuação (*Removing Ponctuation*)

A remoção da pontuação e caracteres especiais é outra prática frequentemente utilizada para pré-processar dados textuais. Existem casos de estudo onde a manutenção da pontuação é importante para as análises, como por exemplo, os *hashtags* na análise de dados do *Twitter* [29]. No entanto, na maior parte dos casos, a pontuação é removida porque não adiciona informação nenhuma ao problema. Tal como no caso anterior, este processo ajuda a diminui o tamanho do conjunto dos dados e, por isso, a aumentar a eficiência computacional dos algoritmos [45].

Exemplo: A frase "Como vai estar o tempo sábado, dia 15?"sujeita ao processo de tokenização e remoção da pontuação resultaria na seguinte lista de *tokens*:

["Como", "vai", "estar", "o", "tempo", "sábado", "dia", "15"]

2.6.1.4 *Stemming*

Stemming é uma técnica de pré-processamento que tem como objetivo diminuir o número de palavras indexadas num documento. Caracteriza-se pela redução de uma palavra à sua forma mais básica (*stem*), isto é, palavras que se encontrem em formas derivadas são convertidas para a sua forma base. Por exemplo, as variações do tipo plural, gerúndio, prefixos e sufixos de uma palavra são mapeadas na mesma raiz, e por isso diminui, significativamente, o vocabulário do problema [23, 29].

Exemplo: As palavras "Corrida", "Correr"e "Correria"são transformadas no *stem* "Corr" referente à ação de correr.

2.6.1.5 Lematização (*Lemmatization*)

O processo de lematização é muito semelhante ao anterior e tem o mesmo objetivo, reduzir as palavras à sua forma original (lema). No entanto, esta técnica é mais complexa, enquanto no *stemming*, resumidamente, são removidos os sufixos e prefixos das palavras e não existe nenhum conhecimento sobre o contexto da palavra, na lematização têm-se em conta um vocabulário e a análise morfológica das palavras. O resultado da lematização é sempre uma palavra válida enquanto que na outra, pode resultar como *stem* uma palavra que não existe no dicionário [45].

Exemplo: Comparando as duas técnicas, enquanto na lematização as palavras "Carro", "Carros"e "Automóvel"são mapeadas no mesmo lema "Carro", no *stemming* apenas são mapeadas as palavras "Carro"e "Carros", com o *stem* "Carro".

2.6.1.6 N-Grams

Em grande parte dos projetos, as palavras são usadas de forma individual, no entanto, algumas delas têm um significado ambíguo quando são retiradas do contexto [29]. Por exemplo, a palavra "processo", nas frases "Processo judicial"e "Processo criativo", apresenta significados diferentes. Por este motivo, surgiu a técnica *n-grams*, caracterizada pela criação de uma sequência contígua de *tokens* de comprimento *n*. Esta técnica permite melhorar a análise dos dados, mas também aumenta o número de *tokens* quando adicionados aos *tokens* individuais, e portanto, o mais comum nestes casos, é filtrar os *n-grams* de maior relevância [23].

Exemplo: A frase "Amanhã vai estar sol.", quando submetida a uma técnica de *2-grams*, resultaria na seguinte lista de *tokens*:

["Amanhã vai", "vai estar", "estar sol"]

2.6.2 Extração de variáveis

O estágio da extração de variáveis é o principal em qualquer processo de PLN. Não podendo as máquinas entender o texto na sua forma bruta, é nesta etapa que os documentos são convertidos em formatos capazes de serem interpretadas por máquinas, que maioritariamente são números. Este processo, de conversão de texto em números, também é muitas vezes denominado de *feature engineering*, pois são usadas diversas competências da engenharia de dados para a estruturação dos dados. Logicamente, o

desempenho dos mecanismos posteriormente usados está diretamente relacionado com a técnica de engenharia de dados usada [45]. A seguir são apresentadas algumas das técnicas mais usadas, com um sucinto resumo da sua metodologia.

2.6.2.1 Representação binária

A representação binária, também conhecida como *One Hot Encoding*, é das técnicas mais simples de PLN, o objetivo é representar a presença de um termo num determinado documento de forma binária. Isto é, se o termo ocorre no documento uma ou mais vezes, na matriz [*documentos x termos*][16], a sua presença é classificada como 1, caso contrário é atribuído o número 0 [23].

doc1 Eu gosto de NLP.

doc2 Eu vou estudar NLP.

doc3 NLP é o futuro.

	Eu	gosto	de	NLP	vou	é	o	estudar	futuro
doc1	1	1	1	1	0	0	0	0	0
doc2	1	0	0	1	1	0	0	1	0
doc3	0	0	0	1	0	1	1	0	1

Figura 2.6: Exemplo da aplicação da técnica *One Hot Encoding*. Cada frase é considerada um documento. Adaptado de [45].

2.6.2.2 Representação por frequência do termo

A representação por frequência, também conhecida como *Count Vectorizer*, acrescenta mais informação que a técnica anterior. Enquanto que na representação binária, não se tem em consideração a frequência com que a palavra ocorre num determinado documento, aqui, o funcionamento, é precisamente o contrário. O facto de contarmos as vezes que um termo ocorre num documento, ganha-se a informação da relevância daquele termo no universo do problema [23, 45].

doc1 Eu gosto de NLP porque o NLP é o futuro.

doc2 Eu vou aprender NLP.

doc3 NLP é o futuro.

	Eu	gosto	de	NLP	vou	é	o	aprender	futuro	e	porque
doc1	1	1	1	2	0	1	2	0	1	0	1
doc2	1	0	0	1	1	0	0	1	0	0	0
doc3	0	0	0	1	0	1	1	0	1	0	0

Figura 2.7: Exemplo da aplicação da técnica *Count Vetorizer*. Cada frase é considerada um documento. Adaptado de [45].

[16] A matriz [*documentos x termos*] é caracterizada pela relação entre os documentos(textos) e os termos(*tokens*). Ao conjunto de termos chama-se vocabulário, que é o conjunto único de *tokens* existente em todo o universo dos documentos. Com isto resulta uma matriz, onde as linhas representam os documentos e as colunas os termos do vocabulário.

2.6.2.3 Representação por TF-IDF

A representação por TF-IDF é das mais utilizadas no PLN e tem como objetivo determinar a importância de um termo relativamente aos outros. Esta técnica invalida a suposição de que termos que aparecem com muita frequência são os mais importantes para o problema. Por exemplo, uma palavra que apareça muitas vezes e em muitos documentos deixa de ter valor informativo, diminuindo assim a sua importância [23] no contexto do problema em questão. Por isso, esta técnica resulta do **produto de duas métricas**: a Term Frequency (TF) e a Inverse Document Frequency (IDF) [20].

- **Term Frequency**: mede a frequência com que um termo aparece num documento e é dada pela seguinte fórmula [20]:

$$TF_t = \frac{n}{N_t} \tag{2.3}$$

  onde,

  n é o número de vezes que o termo aparece no documento;

  N_t é o número de termos do documento.

- **Inverse Document Frequency**: mede a importância do termo, relativamente a todos os documentos e é dada pela seguinte fórmula [20]:

$$IDF_t = \log \frac{D}{N_t} \tag{2.4}$$

  onde,

  D é número total de documentos;

  N_t é o número de documentos onde o termo t aparece.

Com a representação por TF-IDF é possível tecer algumas conclusões importantes, tais como: quando um termo tem alta frequência em poucos documentos, o TF-IDF aumenta; quando um termo aparece em muitos documentos, deixa de ser relevante, e por isso o TF-IDF diminui; quando um termo aparece em todos os documentos o TF-IDF é nulo [23].

doc1	Eu gosto de NLP porque o NLP é o futuro.
doc2	Eu vou aprender NLP.
doc3	NLP é o futuro.

	Eu	gosto	de	NLP	vou	é	o	aprender	futuro	e	porque
doc1	0.017	0.048	0.048	0	0	0.048	2	0	0.018	0	0.048
doc2	0.044	0	0	0	0.119	0	0	0.119	0	0	0
doc3	0	0	0	0	0	0.044	0.044	0	0.044	0	0

Figura 2.8: Exemplo da aplicação da técnica TF-IDF. Cada frase é considerada um documento. Repare-se, que neste exemplo, como o termo "NLP"aparece em todos os documentos a sua cotação é 0.

2.6.2.4 Word Embeddings

Na generalidade dos problemas, as técnicas anteriormente abordadas são suficientes para obter os resultados esperados. No entanto, quando é necessário obter as relações semânticas entre as palavras já não são tão competentes. Essas técnicas baseiam-se, essencialmente, na frequência dos termos nos documentos, não contendo nenhuma informação sobre o contexto do termo [45]. Neste sentido, surgiram os mecanismos de *word embeddings* onde uma palavra é mapeada num vetor de números que representam o significado da mesma no conjunto dos dados [25].

Num exemplo prático, imagine-se que se tem um vocabulário com cinco palavras: "Rei", "Rainha", "Homem", "Mulher"e "Realeza". Numa abordagem de *One Hot Encoding*, a palavra "Rei"seria codificada no vetor representado na tabela 2.1.

Rainha	Rei	Princesa	Mulher
0	1	0	0

Tabela 2.1: Exemplo da representação da palavra "Rei", segundo o método de *One Hot Encoding*. Adaptado de [25].

De facto, através desta representação não é possível extrair informação nenhuma, para além da constatação do facto, da ocorrência ou não ocorrência dos termos, nem tão pouco é possível obter comparações significativas entre vetores, a não ser os que são, efetivamente, iguais (mesma palavra) [25].

Em contrapartida, apareceram os *word embeddings*, que passam a representar cada termo com um vetor, que permite inferir as relações semânticas entre as palavras. Ainda no contexto do exemplo anterior, se assumirmos algumas dimensões contextuais, apenas exemplificativas, a representação dos termos resultaria nos vetores representados na Figura 2.9. É claro que, na resolução algoritmica, estas dimensões contextuais não existem, fica a carga da rede neuronal defini-las [25].

	Rainha	Rei	Princesa	Mulher
Realeza	0.99	0.99	0.98	0.02
Masculino	0.05	0.99	0.02	0.01
Feminino	0.93	0.05	0.94	0.99
Idade	0.6	0.7	0.1	0.5
...	...	...	...	...

Figura 2.9: Exemplo da representação de um vocabulário segundo vetores de palavras (*word embeddings*). Adaptado de [25].

Com esta metodologia já é possível efetuar comparações semânticas entre palavras. Termos semanticamente relacionados terão vetores (*embeddings*) semelhantes. Por exemplo, "Rei"e "Rainha"são mais

próximos no contexto do termo "Realeza"e portanto, os vetores são mais próximos. Para além disso, como os termos são representados por vetores, matematicamente, é possível efetuar operações aritméticas entre eles, adicionando e subtraindo significados de diferentes termos, permitindo a extração de novos contextos, por exemplo, $Rei - Homem + Mulher = Rainha$ [25].

Quanto á implementação dos *word embeddings*, eles são baseados em sistemas de previsão e por isso, usam redes neuronais, maioritariamente algoritmos de DL, para treinar os modelos (aprender os pesos) e assim representarem os vetores das palavras. Neste sentido, o modelo mais usada é o *word2vec*, um modelo baseado em DL desenvolvido pela *Google* [45]. Resumidamente, este modelo, durante o processo de aprendizagem da rede, tem como objetivo aprender os pesos dos neurónios de uma camada intermédia (oculta), que na verdade, corresponderão aos *embeddings* das palavras. O processo é simples, são passados dados para a rede, sempre rotulados, e durante a aprendizagem a rede ajusta os pesos da camada intermédia. Existem dois tipos de *word2vec*, embora sejam semelhantes, diferem na forma como efetuam a previsão [36]:

- **Modelo Skip-Gram:** O modelo *skip-gram* dada uma palavra prevê as palavras circundantes. O intervalo de palavras vizinhas é um hiperparâmetro, isto é, cabe ao utilizador escolher quantas palavras vizinhas (tanto à esquerda como à direita) é que deseja ter em conta. Repare-se que este valor está diretamente relacionado com a complexidade de computação e a qualidade dos *embeddings* [36].

- **Modelo CBOW (Continuous Bag-of-Words):** O modelo *CBOW*, tem o funcionamento oposto, ou seja, prevê a palavra central dando como *input* as palavras circundantes. Na verdade, a camada intermédia acaba por ser igual ao modelo anterior, no entanto as dimensões da camada de entrada e saída mudam.

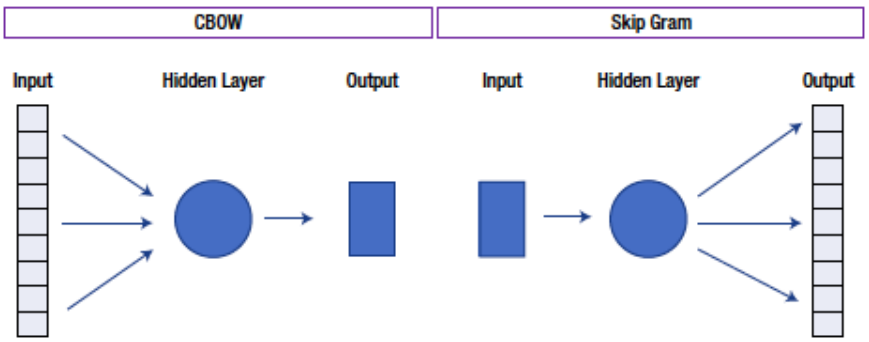


Figura 2.10: Esquema das arquiteturas dos modelos *CBOW* e *Skip-Gram*. Fonte: [45].

2.7 Conclusão

Terminada a revisão da literatura elabora-se agora uma análise crítica daquilo que foi investigado.

No diz que diz respeito à organização judiciária portuguesa, é possível concluir que o estudo realizado permitiu ter uma visão unificada do sistema judicial português, que, a partir deste momento possibilita contextualizar os dados que servirão de base para este projeto.

Relativamente ao estudo da inteligência artificial aplicada na área da justiça, para além de uma apresentação dos princípios elaborados pela CEPEJ, que serão tidos em conta no desenvolvimento do projeto, também se apresentaram vários estudos resultantes da interligação destas duas áreas. De facto, existe uma panóplia de estudos científicos relacionados com o tema deste projeto, e como tal, é importante realizar uma correlação entre os mesmos e a solução a ser desenvolvida. Uma das palavras-chave comum à maior parte da literatura é PLN. Visto que os dados jurídicos são maioritariamente de tipologia textual esta área de computação apresenta um papel fundamental na extração de conhecimento. Outro aspeto importante que foi analisado é que as taxas de precisão dos estudos de previsão da decisão são, predominantemente, valores entre os 70% e 80%, que desde logo, permitirá, no término do projeto, estabelecer uma relação de qualidade entre a solução desenvolvida e os trabalhos relacionados.

Para além disso, todos os estudos são associados a dados específicos, geralmente de tribunais superiores dos locais onde os estudos são efetuados. Neste sentido, aplicados a dados jurídicos portugueses ainda não se efetuaram trabalhos relacionados, portanto este é um fator diferenciador da solução que se pretende desenvolver. Por outro lado, as soluções existentes são muito técnicas com produtos finais com pouca utilidade e transparência para o utilizador. Neste projeto, para além da parte de extração de conhecimento, a solução que se pretende desenvolver destaca-se pela apresentação, ao utilizador final, de informações de alto nível que serão úteis no seu quotidiano.

Capítulo 3

Desenho e implementação de um *pipeline* de análise de dados judiciais

O principal problema, o qual se pretende apresentar uma solução com este projeto, passa pela necessidade de estruturação e extração de conhecimento a partir de dados jurídicos portugueses. Neste sentido, tenciona-se desenvolver uma solução caracterizada por um conjunto de mecanismos, que permita atingir todos os objetivos inicialmente propostos.

O primeiro passo no desenvolvimento de uma solução técnica passa pela definição da arquitetura do sistema. O objetivo desta arquitetura é facilitar a compreensão e organização das várias componentes da solução. A Figura 3.1 apresenta a arquitetura da solução desenvolvida neste projeto.

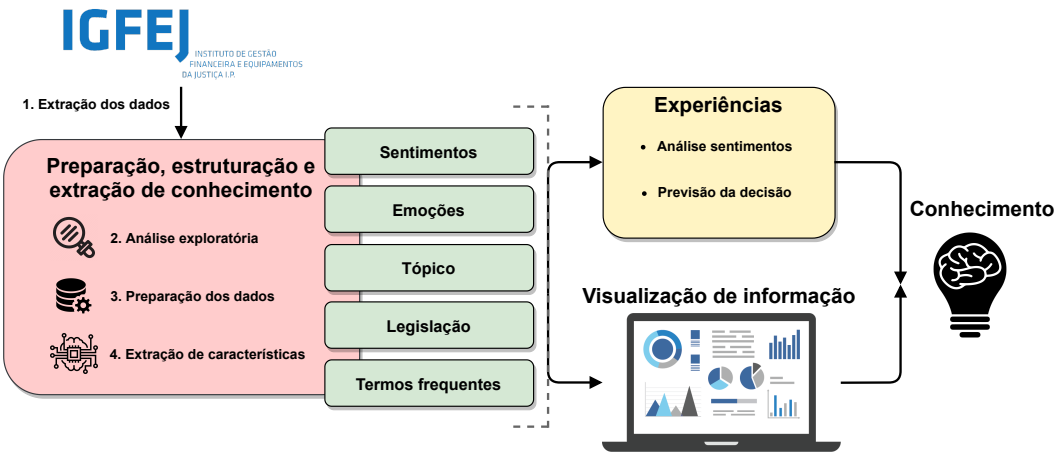


Figura 3.1: Arquitetura conceptual da solução desenvolvida.

Nesta arquitetura destacam-se quatro componentes: a origem dos dados, que são os dados disponibilizados pelo Instituto de Gestão Financeira e Equipamentos da Justiça (IGFEJ); a componente relativa aos métodos e materiais; a componente que diz respeito às experiências realizadas com os dados; e a componente de apresentação e visualização de informações dos dados. Para além disso, também é possível visualizar o conjunto de mecanismos cronológicos desenvolvidos, bem como os resultados obtidos através dos mesmos, nomeadamente o conjunto de informações e conhecimentos intermédios: sentimentos, emoções, tópico, legislação e termos frequentes.

Neste sentido, no presente capítulo é apresentado todo o trabalho prático desenvolvido. A divisão do mesmo diz respeito ao conjunto de componentes apresentados na arquitetura de Figura 3.1.

3.1 Preparação, estruturação e extração de conhecimento a partir dos dados judiciais

Esta secção visa apresentar as primeiras e fundamentais etapas do processo de extração de conhecimento. A divisão das secções segue um processo cronológico que passa por várias fases, nomeadamente a extração e agregação dos dados, seguindo-se a etapa de análise exploratória, onde são apresentadas as principais características dos dados. Posteriormente, expõem-se os mecanismos usados no pré-processamento e finalmente, umas das fases mais importante do projeto, a extração de características, onde se apresentam as metodologias usadas na extração de informações e conhecimento dos dados. A secção termina com uma síntese das tecnologias usadas neste projeto de dissertação.

3.1.1 Extração dos dados

A extração de dados é a primeira etapa de um processo de KDD que dependa dos dados de fontes externas. Recolher os dados de diferentes fontes e integrá-los num único local permite ter uma visão unificada e homogénea dos mesmos. Para isso, é preponderante estudar as fontes dos dados antes de se efetuar o processo de extração. Deste modo, nesta secção apresenta-se a fonte, o processo de extração e finalmente o sub-conjunto de dados que vai ser objeto de estudo.

Os dados que estão na origem deste projeto são provenientes do IGFEJ. O IGFEJ é um instituto público, fundado em 2012, que tem como principal propósito a prestação de serviços fundamentais para o bom funcionamento do sistema judiciário português. As suas responsabilidades assentam em várias áreas, nomeadamente na gestão financeira e orçamental do Ministério da Justiça, monitorização das custas processuais, gestão e manutenção do património do Ministério da Justiça e gerir, dinamizar e desenvolver

os sistemas de informação relacionados com a justiça [50]. É, sobretudo, em consequência desta última área de intervenção, que esta entidade disponibiliza uma base de dados jurídico-documental que contém dados de acórdãos dos tribunais judiciais e administrativos superiores, dados de acórdãos do tribunal constitucional, do tribunal de contas e julgado de paz, de vários anos[1]. Este instituto disponibiliza a base de dados gratuitamente na *internet* no seu *website*. A organização da interface que contém a base de dados é apresentada no esquema da Figura 3.2.

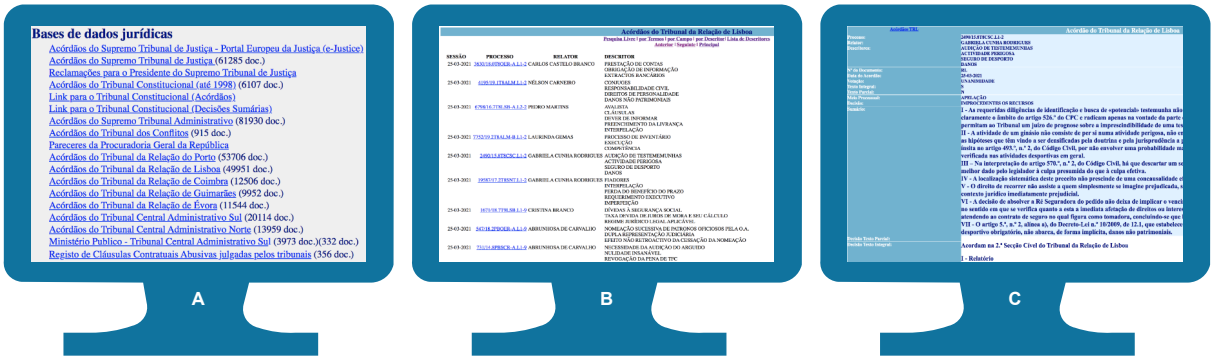


Figura 3.2: Organização do *website* que contém a base de dados. Numa primeira página são apresentados a lista de tribunais (A) e o número de acórdãos de cada um. Selecionando um tribunal, é apresentada uma lista com um número X de acórdãos, ordenados cronologicamente, com alguns metadados de cada processo (data, descritores, relator, número do processo) (B). Para além da lista, apresenta-se um cabeça-lho que permite filtrar processos e avançar para as páginas seguintes. Finalmente, ao selecionar um processo é exibida uma página com as informações do processo, na forma de uma tabela (C).

Esta base de dados *online* apenas contém os dados para efeitos de rápida consulta e visualização via uma interface *web*. Deste modo, a primeira etapa deste projeto de dissertação passou pela extração dos dados de modo a tornarem-se passíveis de serem trabalhados e processados. Numa primeira instância contactou-se o Conselho Superior da Magistratura, entidade responsável pelo esclarecimento de dúvidas relativas aos dados, solicitando um método mais simples e rápido para obter os dados num outro formato, por exemplo uma API, no entanto, a resposta foi de que não existia outra forma externa de acesso aos dados. Neste sentido e já que o único acesso a esta base de dados é por via de um *browser*, as possibilidades de extração dos dados são limitadas. Assim sendo decidiu recorrer-se a mecanismos de *web scrapping* para esse efeito.

Apesar do termo *web scrapping* não ser recente, segundo Ryan Mitchell, este conceito de extração automática de dados da *internet* tem vindo a apresentar designações diferentes, como por exemplo, *screen scraping* e *data mining*. Este conceito é caracterizado pelo desenvolvimento de um programa automático que acede a um servidor *web*, solicita dados (maioritariamente páginas na forma de HyperText Markup Language (HTML)) e posteriormente extrai e armazena as informações selecionadas [21]. Tendo por base este processo de *web scrapping*, desenvolveu-se um algoritmo, ajustável a cada tribunal, para extrair os

[1] http://www.dgsi.pt/home.nsf?OpenDatabase

acórdãos e armazená-los no formato Comma-separated values (CSV). Apesar de esteticamente as páginas de cada tribunal serem semelhantes, a sua estrutura HTML apresenta pequenas diferenças, que fez com que o algoritmo tivesse de sofrer algumas alterações consoante o tribunal em questão, no entanto, a formulação do mesmo foi igual em todos os casos e baseou-se no seguinte algoritmo:

1. Em primeiro lugar solicita-se a página inicial da lista de acórdãos do tribunal **X**.
2. De seguida entra-se num ciclo que itera sobre todas as páginas dessa lista.
3. Em cada página recorre-se às ferramentas da biblioteca *BeautifulSoup* para extrair a lista de *urls* dos acórdãos.
4. Seguidamente itera-se sobre essa lista e para cada *url* solicita-se a página com a informação detalhada do acórdão. Recorrendo à biblioteca *BeautifulSoup* extrai-se a informação do acórdão e guarda-se numa estrutura de dados global.
5. Terminado o ciclo que percorre todas as páginas de listas de acórdãos, a estrutura de dados global é convertida num ficheiro de formato CSV.

Este processo foi repetido para todos os tribunais e por conseguinte foi bastante moroso e computacionalmente pesado, uma vez que entre cada solicitação ao servidos que contém os dados foi feito um *sleep* de 8 segundos, isto é, um compasso de espera, para que o mesmo não fosse sobrecarregado. O tempo de execução do programa para cada tribunal foi proporcional à quantidade de acórdãos do mesmo, e portanto, no fim, a média de tempo de extração de todos os acórdãos de um tribunal foi aproximadamente 26 horas. Como resultado deste processamento obteve-se o conjunto de *datasets* representados na tabela 3.1, perfazendo um total de 11 *datasets* com um tamanho acumulado de cerca de 9 GB.

Numa primeira análise superficial dos dados, verificou-se que, apesar de alguns parâmetros serem diferentes de tribunal para tribunal, na generalidade, apresentam a mesma estrutura, ou seja são dados semiestruturados. Por este motivo e porque os recursos computacionais são limitados, decidiu-se basear todo o projeto de dissertação apenas nos dados dos tribunais da relação, nomeadamente, no tribunal da relação do Porto, Évora, Lisboa, Coimbra e Guimarães. Note-se que, tudo o que se desenvolveu para estes dados é facilmente replicável para todos os outros.

Tabela 3.1: Tabela representativa dos *datasets* obtidos com o processo de extração. Contém o tamanho em *GigaBytes* de cada *dataset* e o número de acórdãos presente em cada um.

Tribunal	Tamanho (Gb)	Número de acórdãos
Supremo Tribunal de Justiça	1.6	59779
Supremo Tribunal Administrativo	1.5	80185
Tribunal Central Administrativo do Sul	0.4	3973
Tribunal Central Administrativo do Norte	0.5	11892
Jurisprudência dos Julgados de Paz	0.3	6732
Tribunal dos Conflitos	0.1	872
Tribunal da Relação de Guimarães	0.4	8526
Tribunal da Relação de Coimbra	0.3	11823
Tribunal da Relação do Porto	1.4	52171
Tribunal da Relação de Évora	0.6	10250
Tribunal da Relação de Lisboa	1.4	48100

3.1.2 Análise exploratória dos dados

A fase de exploração dos dados é crucial para o bom desenvolvimento de projetos de descoberta de conhecimento. Antes de se efetuar qualquer processamento nos dados é essencial visualizá-los e interpretá-los de forma a poder tomar-se as melhores decisões nas etapas seguintes. Nesta secção apresenta-se o conteúdo e as principais características do conjunto de dados a ser estudado.

O *dataset* inicial é constituído pelas variáveis representadas na tabela 3.2.

Tabela 3.2: Constituição do *dataset* inicial.

Variável	Descrição
Processo	Código único identificativo do processo.
Data do Acórdão	Data em que o acórdão foi escrito.
Relator	Juiz relator que é o titular do processo.
Descritores	Conjunto de palavras-chave, selecionadas pelo analista ou relator presente no acórdão.
Meio Processual	Tipo de meio processual que deu origem ao acórdão (Recurso, Agravo, etc..).
Decisão	Decisão tomada pelo juiz.
Decisão Texto Integral	Texto integral escrito, normalmente, pelo juiz relator do acórdão.
Sumário	Resumo do acórdão.
Votação	Tipo de votação.
Tribunal	Tribunal responsável pelo acórdão.

Na realidade, aquando da extração dos dados, foi identificado um maior número de variáveis, no entanto, estavam presentes num número muito reduzido de acórdãos, o que não adjudicaria informação relevante num cômputo geral. Deste modo aquelas que em mais de 90% dos acórdãos não tinham valor, foram removidas de imediato, dando origem à constituição presente tabela anterior.

Este *dataset* é constituído por 130870 acórdãos e as datas da escritura dos acórdãos variam entre 06-04-1965 e 25-10-2019. As figuras 3.3 e 3.4 representam a distribuição de acórdãos por ano e mês.

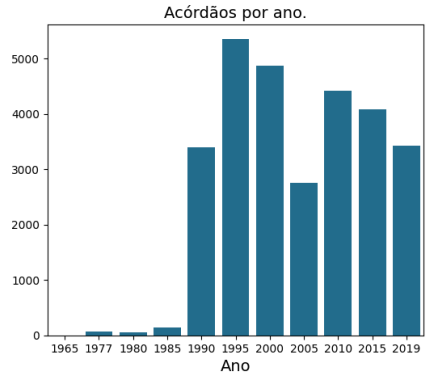


Figura 3.3: Número de acórdãos ao longo dos anos.

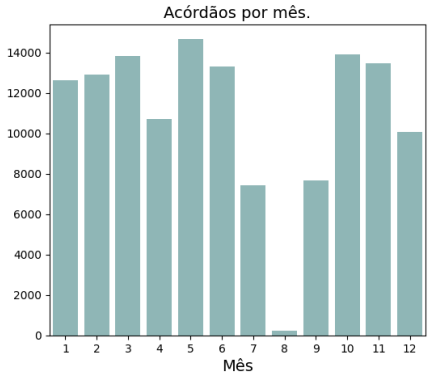


Figura 3.4: Número de acórdãos por mês.

De facto, o número de acórdãos inseridos na base de dados até ao ano de 1995, seguiu sempre uma tendência de crescimento, isto deveu-se, possivelmente, aos avanços da tecnologia e informatização da informação ao longo desses anos. Já nos últimos anos, a tendência tem sido de decréscimo.

Quanto à distribuição de acórdãos por mês observa-se quase uma constante ao longo dos meses do ano, exceptuando uma clara diminuição nos meses de verão: julho, agosto e setembro. Isto, seguramente, deve-se ao facto de os tribunais estarem fechados devido a férias laborais.

Por outro lado, analisando a distribuição de acórdãos pelos tribunais onde foram ajuizados os recursos, Figura 3.5, constata-se que os tribunais de regiões com mais população, Porto e Lisboa, são os que, efetivamente, reapreciam mais recursos, como já era espectável.

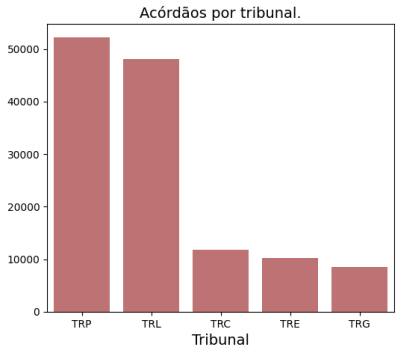


Figura 3.5: TRP - Tribunal da Relação do Porto; TRL - Tribunal da Relação de Lisboa; TRC - Tribunal da Relação de Coimbra; TRE - Tribunal da Relação de Évora; TRG - Tribunal da Relação de Guimarães;

Quanto aos descritores, palavras-chave associadas a cada acórdão, são extremamente importantes no processo de pesquisa, uma vez que através dos mesmos é possível filtrar os acórdãos tendo em conta as matérias tratadas. Segundo o IGFEJ os descritores são termos pré-definidos que constam numa tabela sujeita a permanente atualização com vista a evitar uma expansão desmedida. No entanto, numa primeira fase de análise exploratória, investigou-se a variabilidade deste campo e obteve-se um total de 30724 diferentes descritores usados neste *dataset*, que de facto, é um número bastante elevado, não permitindo uma rápida categorização dos acórdãos. A Figura 3.6 apresenta os 15 descritores mais frequentes no *dataset*.

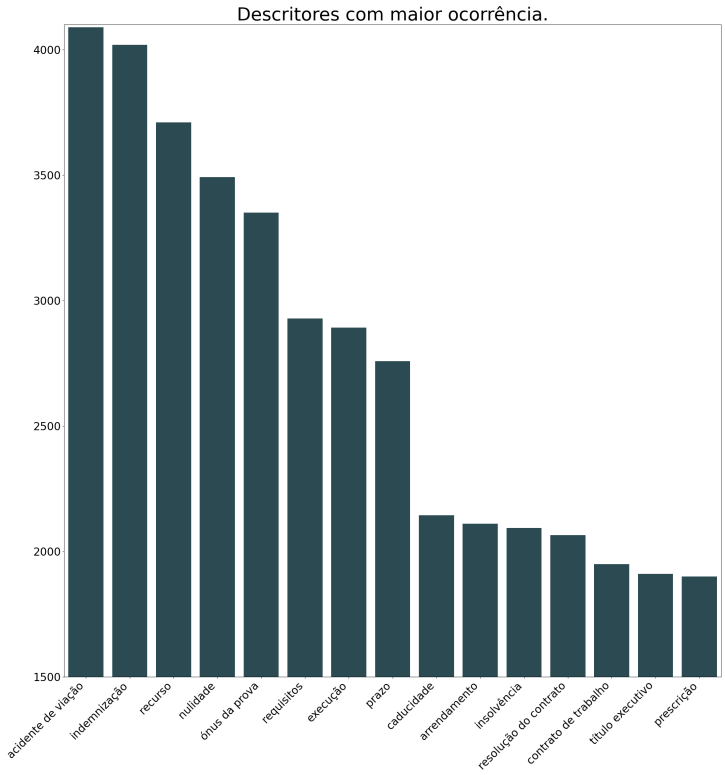


Figura 3.6: Frequência dos 15 descritores com maior ocorrência no *dataset*.

Através da figura, pode destacar-se que existem pelo menos 4090 acórdãos que, de alguma forma, estão relacionados com acidentes de viação, e que 4020 serão alusivos a indemnizações.

Por outro lado, observou-se a decisão do acórdão, que é a variável objetivo deste projeto. Como à partida se pretende que esta seja uma variável categórica, analisou-se também a variabilidade da mesma e constatou-se que existem 2799 tipos de decisão diferentes. A Figura 3.7 apresenta os 10 tipos de decisão mais frequentes no *dataset*. Através do gráfico pode auferir-se desde logo que apesar de alguns tipos apresentarem grafia diferente, semanticamente têm o mesmo significado.

Finalmente fez-se uma análise muito superficial do texto integral da decisão e verificou-se que na

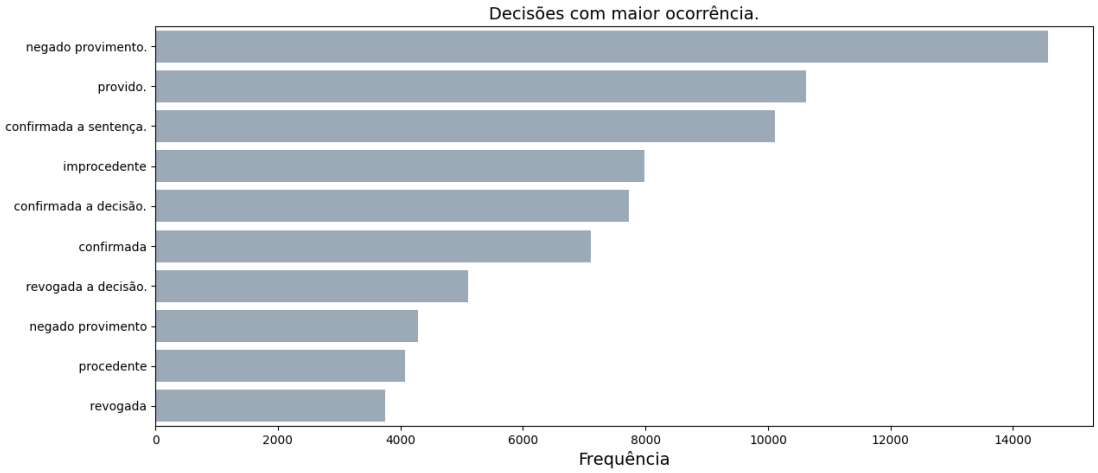


Figura 3.7: Número de ocorrência dos 10 tipos de decisões mais frequentes no *dataset*.

maior parte dos acórdãos o texto está subdividido em três partes: o **relatório**, que contém uma síntese do processo e dos termos recorridos; a **fundamentação** onde o juiz fundamenta a decisão e apresenta a legislação em que se baseou; e finalmente a **decisão** que contém objetivamente a decisão propriamente dita. Apesar desta divisão ser quase transversal a todos os acórdãos, em alguns casos é mais explícita do que em outros. Um exemplo do texto integral da decisão de um acórdão pode ser consultado em apêndice na Figura A.1.

3.1.3 Preparação dos dados

A etapa de preparação dos dados tem como objetivo preparar, organizar e transformar os dados para serem processados nas fases seguintes. Os procedimentos adotados nesta fase, apresentados nas seguintes secções, foram selecionados tendo em conta um conhecimento do domínio do problema e dos mecanismos a serem aplicados posteriormente.

3.1.3.1 Valores em falta

Os números de valores em falta no *dataset* estão apresentados na tabela 3.3.

De facto, o campo com mais valores em falta é o texto integral da decisão. Sabendo à partida para este projeto, que a principal fonte de conhecimento é o texto da decisão, todos aqueles acórdãos que não apresentam texto não têm relevância no problema, deste modo, foram removidos. O mesmo se aplicou ao campo decisão, já que, todos os acórdãos que não têm uma decisão não serão passíveis de serem categorizados, e por isso também foram removidos. Todas as outras variáveis com valores em falta, permaneceram com os valores *NaN*.

Tabela 3.3: Valores em falta do *dataset*.

Variável	Valores em falta
Processo	0
Data do Acórdão	0
Relator	0
Descritores	19
Meio Processual	3043
Decisão	4907
Decisão Texto Integral	46205
Sumário	631
Votação	1163
Tribunal	0

Como resultado deste processamento, originou-se um *dataset* com 79758 acórdãos.

3.1.3.2 Adição do ECLI

O European Case Law Identifier (ECLI) é um identificador de documentos jurídicos, normalizado, reconhecível e compreensível tanto por humanos como por computadores. É um sistema de identificação comum a todos os Estados-Membros da Europa que tem como principais objetivos facilitar a utilização de bases de dados, otimizar a pesquisa de jurisprudência nacional e europeia e ainda simplificar a citação de jurisprudência. É constituído, obrigatoriamente, pelas seguintes 5 partes [51, 53]:

- A sigla **"ECLI"**: que caracteriza o identificador como sendo europeu de jurisprudência;
- O código do país;
- O código do órgão jurisdicional que proferiu a decisão;
- O ano em que a decisão foi proferida;
- Um numeral ordinal com um formato a decidir por cada país, nunca excedendo 25 caracteres. Em Portugal, tem por base o número de processo.

Cada Estado-Membro é responsável por designar uma entidade governamental ou judicial para coordenar o ECLI a nível nacional. Uma das tarefas desta entidade é a criação da lista dos códigos para os orgãos jurisdicionais, que no caso português, corresponde à identificação do Supremo Tribunal de Justiça - «STJ» e aos Tribunais das Relações de Lisboa, Porto, Coimbra, Évora e Guimarães - «TRL», «TRP», «TRC», «TRE», «TRG», respetivamente. O coordenador deste identificador em Portugal é o Concelho Superior da Magistratura [53].

Um exemplo de um ECLI português é o seguinte:

ECLI:PT:TRL:2017:1062/14.9TVLSB.L1-2,

que corresponde à decisão proferida em Portugal, pelo Tribunal da Relação de Lisboa, em 2017, no processo 1062/14.9TVLSB.L1-2.

No caso do projeto desta dissertação a fonte de dados não contém o ECLI como metadado dos acórdãos. Portanto, para acompanhar a adesão de Portugal a esta identificação e todas as suas vantagens, elaborou-se um *script* para adicionar o ECLI aos dados do *dataset*.

O intuito deste *script* foi cruzar os dados deste projeto com uma fonte externa que contenha o ECLI. A fonte externa selecionada foi o portal *e-Justice*, um portal europeu da justiça, criado para tirar o máximo partido da tecnologia na justiça que disponibiliza o melhor e mais eficiente acesso à justiça e à informação. Uma das características deste portal é que disponibiliza um motor de busca ECLI, ou seja, que permite a procura de decisões judiciais/sentenças através do identificador ECLI.

Tendo por base este motor, para cada acórdão, fez-se um pedido *Web* com o filtro de identificador simples do acórdão (Número do processo) e a data do acórdão. Como resultado obtém-se uma página com a informação do processo e através de mecanismos de *web scraping* (Secção 3.1.1) extraiu-se o ECLI correspondente.

3.1.3.3 Categorização da variável dependente

Como se verificou na secção 3.1.2, ao analisar os valores únicos da decisão observou-se um valor extremamente mais elevado do que era expectável, e por outro lado, como um dos principais objetivos deste projeto de dissertação passa pela análise da correlação entre as várias variáveis de um acórdão com a decisão tomada, uma das primeiras fases de pré-processamento foi obter uma categorização desta variável que seja passível de ser processada.

De facto, a heterogeneidade da variável impossibilita a aprendizagem de qualquer modelo de ML. No entanto, como já referido, verificou-se que na maior parte dos casos apesar da grafia das decisões ser diferente, apresentam valores semânticos semelhantes. Ainda assim, como o *know how* no que toca a termos jurídicos não é muito elevado recorreu-se a um advogado para esclarecer as expressões que são juridicamente semelhantes e que nos possibilitam a categorização do mesmo modo, diminuindo assim a diversidade de decisões.

Geralmente, os juízes aquando da análise do recurso, podem tomar uma das seguintes decisões:

- Aceitar o recurso e portanto reformular a sentença nos pontos recorridos;

- Não aceitar o recurso e assim, manter a sentença do tribunal inferior;
- Aceitar parcialmente o recurso;
- Recusar parcialmente o recurso;

O esquema da Figura 3.8 apresenta a categorização da variável decisão bem como alguns dos sinónimos que foram agrupados em cada categoria.

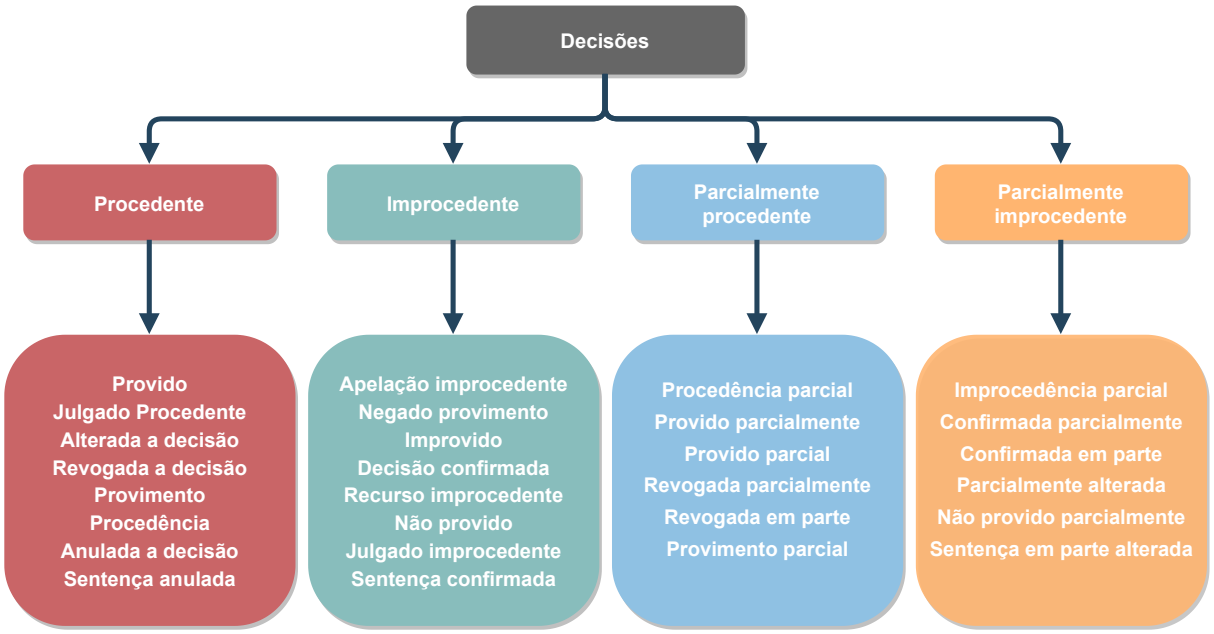


Figura 3.8: Categorização das decisões dos acórdãos. A cada categoria correspondem alguns dos sinónimos que foram agrupados.

Todos os outros tipos de decisão que não se enquadrem nestes 4, foram categorizados como **"Outra"** uma vez que não apresentam relevância no contexto do nosso problema.

Portanto, como consequência deste pré-processamento o número de acórdãos de cada categoria é o que está apresentado na tabela 3.4. Reduziu-se drasticamente a variância dos valores o que permite, a partir deste momento, aplicar mecanismos de descoberta de conhecimento.

Tabela 3.4: Número de acórdãos por categoria.

Decisão	Frequência
Procedente	38496
Improcedente	24552
Parcialmente procedente	8246
Parcialmente improcedente	8140
Outra	324

3.1.3.4 Pré-processamento do texto

Como em todos os casos de análise e processamento de textos, antes de aplicar qualquer modelo de aprendizagem é necessário efetuar um pré-processamento para adaptá-lo aos mecanismos a usar.

Antes de qualquer processamento, retirou-se do texto integral da decisão a parte correspondente à decisão propriamente dita para não conter informação da decisão nas variáveis, tornando os modelos posteriores mais realistas.

Neste caso de estudo foram implementados os seguintes mecanismos, de forma a limpar e adequar a representação do texto para a posterior análise.

1. **Tokenização:** Aplicou-se um mecanismo de tokenização simples, ou seja, usou-se o delimitador espaço para dividir o texto em *tokens*.

2. **Remoção de palavras irrelevantes:** Também se utilizou o procedimento de remoção das palavras que são muitas vezes repetidas num texto e são irrelevantes para o problema. Para além de usar as listas de *stopwords* da língua portuguesa já existentes, também se adicionar palavras que neste contexto não têm importância para o problema. A lista de palavras removidas está apresentada no apêndice B.2.

3. **Remoção de pontuação:** Outra técnica adotada foi a remoção da pontuação.

4. **Lematização:** Finalmente, e de modo a reduzir substancialmente o vocabulário, foi aplicado o processo de lematização a cada *token*.

3.1.4 Extração de características

A presente secção diz respeito à extração de características dos dados, que no inglês é designada por *"features extraction"*. É uma das etapas mais relevantes deste projeto. Partindo do facto que a base de extração de conhecimento é o texto da decisão dos acórdãos judiciais, é importante transformar estes dados textuais brutos numa forma passível de ser processada sem que haja perda de informação. Neste sentido, as quatro subdivisões desta secção, representam as quatro principais informações e conhecimentos extraídos do texto, que serão usados nas experiências posteriormente desenvolvidas.

3.1.4.1 Termos mais frequentes

A principal fonte de conhecimento dos dados é o texto da decisão e portanto, uma forma de obter esse conhecimento é obter uma representação do texto através dos dos termos mais utilizados pelos juízes nos seus textos das decisões.

Neste sentido, foi feita uma vectorização do texto segundo uma técnica de PLN, denominada TF-IDF, explicada na secção 2.6.2.3. Optou-se por esta técnica uma vez que, para além de permitir obter a frequência dos termos, permite ainda o cálculo da relevância de cada um, sendo que o resultado é o produto destas duas variáveis. Neste sentido, no contexto jurídico onde os termos são muito específicos, ter conhecimento da importância de cada um é muito relevante. Para além disso o resultado da aplicação desta técnica é uma matriz que serve de *input* para os mecanismos usados posteriormente.

Os parâmetros que foram ajustados são os seguintes:

- ***Min_df***: parâmetro que representa o valor mínimo da frequência de um termo para este pertencer ao vocabulário. Neste caso foi 10%, ou seja, os termos que apenas aparecem em menos de 10% dos acórdãos são ignorados.

- ***Max_df***: parâmetro que representa o valor máximo da frequência de um termo para este pertencer ao vocabulário. Neste caso foi 90%, ou seja, os termos que aparecem em mais de 90% dos acórdãos são ignorados, uma vez que são termos muito repetidos e por esse motivo deixam de contribuir para qualquer conhecimento.

- ***Ngram_range***: propriedade que representa o intervalo de limite inferior e superior dos *n-grams* a serem extraídos. Neste contexto usou-se o intervalo (1,3), isto é, apenas foram tidos em conta *unigrams*, *bigrams* e *trigrams*.

- ***Max_features***: valor representativo do número limite de termos que são tidos em conta para a construção do vocabulário. Neste projeto o número escolhido foi 2500 termos, tendo em conta o número total de termos em todos os acórdãos e o limite computacional disponível.

Ainda que o limite de termos selecionado tenha sido 2500, no final da aplicação desta técnica, obteve-se a matriz:

$$DT_{axt} \tag{3.1}$$

onde,

a é o número total de acórdãos;

t são os 2305 termos mais frequentes nos textos.

Deste modo, cada célula da matriz *DT* representa a frequência do termo t no acórdão a.

3.1.4.2 Tópico

Uma das formas de extração de conhecimento de dados textuais é a categorização dos textos consoante o conteúdo semântico que tratam. Neste projeto de dissertação o único campo que à partida permitiria categorizar os acórdãos segundo a área temática seria os descritores. No entanto, como se chegou à conclusão na análise exploratória, existem **30724** tipos de descritores diferentes, o que, com os recursos de computação e temporais que se dispõe, impossibilita o uso de qualquer mecanismo de agrupamento e categorização. Neste sentido, procedeu-se à classificação de acórdãos tendo em conta o conteúdo semântico do texto da decisão.

Assim sendo, decidiu-se utilizar mais uma técnica de PLN, denominada *topic modeling*, uma vez que esta técnica tem como objetivo obter tópicos abstratos associados a cada documento, e como consequência a associação semântica entre os documentos, que é o desejado neste problema. Um dos algoritmos mais usados neste âmbito e portanto, que também se usou neste projeto foi o Latent Dirichlet Allocation (LDA). O LDA é um modelo estatístico iterativo que tem como principio que um tópico é um conjunto de palavras e um documento é uma mistura de tópicos. Portanto, tendo por base um número **K** de tópicos, no final das iterações do algoritmo, a cada acórdão correspondem **K** associações, no entanto, no contexto deste problema apenas se associa o tópico cujo valor de associação é mais elevado.

Neste sentido, partindo da matriz anterior *DT* 3.1, este algoritmo agrupa os acórdãos tendo em conta as frequências dos termos nos mesmos, ou seja, acórdãos cujas frequências dos mesmos termos são semelhantes terão uma classificação no mesmo tópico também ela semelhante.

O principal hyperparâmetro do LDA é o **K**, que é o número de tópicos que se espera como resultado do algoritmo. Sendo um hyperparâmetro, trata-se de um valor que precisa de ser ajustado tendo em conta o conhecimento profundo do domínio dos dados, para se perceber o que melhor se ajusta aos mesmos. Ainda que existam algumas métricas de avaliação destes algoritmos, a mais usada continua a ser a técnica manual, ou seja, treinar o algoritmo com vários valores de **K** e escolher o que melhor se adapta ao domínio dos dados. No contexto do projeto desta dissertação testaram-se valores do intervalo de 10 a 100, e tendo por base os termos que pertencem a cada tópico o que obteve melhores resultados foi o valor **K=20**.

No final do algoritmo ainda se fez uma nova avaliação e os tópicos que apresentavam um conjunto de palavras semanticamente semelhantes foram agrupados, foi o caso do *0-9-10* e o *2-18*. O resultado deste processamento está apresentado na tabela 3.5.

Em síntese, no fim da aplicação desta técnica, a cada acórdão corresponde um número que caracteriza o seu tópico, ou seja, a categoria do conteúdo semântico do seu texto.

Tabela 3.5: Resultado do processo de modelação de tópicos. A cada tópico correspondem as 10 palavras mais frequentes. A descrição foi atribuída manualmente, tendo em conta a semântica do conjunto de palavras pertencentes ao tópico, bem como o conjunto de descritores mais frequente por tópico.

Tópico	Descrição	Palavras
0	Prescrição de penas / Caducidade	ação, pedir, autor, réu, causar, civil, prescrição, prazo, julgar, sentença, despachar
1	Insolvência / Crédito	devedor, passivar, credor, crédito, requerente, administrador, situação, dívida, sociedade, restante, requerido
2	Nulidade / Contraordenação	arguir, penal, peno, crime, prisão, acusação, criminal, instrução, despachar, suspensão, assistente
3	Provas / Ónus da prova / Factos	provar, depoimento, testemunhar, matéria, matéria facto, autor, quesito, recorrente, réu, gravação, sobrar
4	Família / Paternidade	provar, depoimento, testemunhar, matéria, matéria facto, autor, quesito, recorrente, réu, gravação, sobrar
5	Resolução contratuais / Contrato-Promessa	contratar, cláusula, vender, autor, comprar, celebrar, autora, contrato, comprar vender, contratual, preço
6	Usucapião / Propriedade	prédio, terreno, réu, propriedade, predial, escriturar, autor, vender, fração, imóvel, posse
7	Dívidas / Penhoras / Execuções	exequente, executivo, penhorar, executar, negócio, título, oposição, crédito, obrigação, pagamento, bem
8	Inventário / Arresto	requerente, requerido, quota, providência, sociedade, bem, partilhar, casal, sócio, requerer, ação, procedimento
11	Acidente de viação / Indeminização	acidentar, veículo, dano, segurar, indemnização, autor, danar, responsabilidade, patrimonial, dano patrimonial, provar
12	Competência / Tribunal competente	competência, administrativo, estado, público, ação, competente, civil, jurisdição, social, segurança social, jurídico
13	Apoio judiciário / Taxas justiça	prazo, despachar, taxar justiça, notificação, requerimento, justiçar mandatário, judiciário, processual, taxar, pagamento,
14	Expropriação	parcelar, construção, indeminização, terreno, valor, área, prédio, utilidade, avaliação, público, justo
15	Acidente trabalho / Incapacidade	incapacidade, acidentar, médico, trabalhar, anual, permanente, exame, lesão, revisão, trabalho, alta
16	Mútuo / Ónus da prova	banco, bancário, crédito, pagamento, juro, contar, réu, contratar, autor, quantiar, obrigação
17	Contrato trabalho / Despedimento	trabalhar, trabalhador, contratar, autor, retribuição, trabalho, subsídio, serviço, disciplinar, provar, empresar
19	Arrendamento / Obras	obras, contratar, render, autor, prédio, alocar, resolução, reparação, provar, fração, autor

3.1.4.3 Legislação

Os textos das decisões estão repletos de citações de legislação, principalmente na parte da fundamentação da decisão. Neste sentido, uma das formas de obter conhecimento a partir dos dados, é identificar as citações no texto.

Uma citação básica apresenta a seguinte norma: *"Artigo X, alínea Y do Z"*, onde *X* é o número do artigo, *Y* o número/letra da alínea e *Z* o código legislativo. No entanto, numa primeira análise dos textos verificou-se que, apesar de seguirem a base da norma, existe muita variedade na forma de citações da legislação, por exemplo: *"artigo 368.°-A, n.° 1 e n.° 2, do Código Penal"* , *"Artº 87.°, n.° 1 e n.° 2, alíneas a) e b), da Lei n.° 5/2006"* e *"arts 25.°, alínea a) e 21.°, n.° 1, ambos do Decreto-Lei n.° 15/93"*.

Deste modo, para extrair as citações dos textos foi elaborada uma expressão regular bastante complexa, apresentada no apêndice B.1. Apesar da complexidade da expressão devido à heterogeniedade das citações, na prática um dos objetivos é captar todas as expressões textuais que começam com um dos seguintes termos: "artº", "artigo", "arts", "art". No que diz respeito ao caso de paragem é que gera mais dificuldade uma vez que o nome do código legislativo pode apresentar muitas formas diferentes, como por exemplo, *"Código Civil"*, *"CC"*, *"Decreto - Lei"*, *"Lei nº"*, *"C.P.C"*, *"c.proc.civil"* etc... Portanto todas estas formas foram tidas em conta e no final, com esta expressão, alcançou-se uma eficácia de 90-95% de filtragem das referências dos textos da decisão.

Chegados a esta ponto, associado a cada acórdão tem-se as expressões textuais referentes às citações na sua forma bruta, ou seja, tal e qual como o juiz as citou. No entanto, ainda não é possível categorizar esta variável já que, neste ponto, citações do mesmo artigo podem estar escritas de formas diferentes, por exemplo, *"artigo 5 do código civil"*, terá de ter a mesma correspondência que *"artigo 5 do cc"*, *"art. 5 do C.Civil"* ou até *"Art 5 do Cod. Civ."*.

Neste sentido, desenvolveu-se um sistema de padronização de referências com o objetivo de uniformizar as citações numa forma passível de ser categorizada. Para esse efeito, criou-se uma base de dados em ficheiro *Json*, onde se associou a cada legislação portuguesa[2] as diferentes formas como estas surgem nos textos. O excerto exemplificativo 3.1.4.3 é um trecho da base de dados do Código Civil.

Listagem 3.1: Excerto da base de dados de códigos legislativos e as diferentes formas que surgem no textos.

```
{
   "Sigla": "CC",
   "Nome": "Código Civil",
   "Sinónimos": [ "cc", "cciv", "código civil", "c civil", "c c", "c civ", "códcivil", "ccv",
      "cód civ", ... ]
}
```

[2]Para elaborar a lista da legislação consultou-se diferentes bases de dados de direito *online*, nomeadamente: Diário da República Eletrónico - https://dre.pt/; Base de dados Juridica Almedina - http://bdjur.almedina.net/; Autoridade Tributária e aduaneira - https://info.portaldasfinancas.gov.pt/; Procuradia Geral Distrital de Lisboa - http://www.pgdlisboa.pt/.

A lista final dos códigos legislativos pertencentes à base de dados pode ser consultada em apêndice na tabela B.1.

Construída a base de dados, o passo seguinte foi, para cada acórdão, estabelecer uma correlação entre a citação propriamente dita e a base de dados. No final, e ignorando as alíneas dos artigos legislativos citados, uma vez que são muito distintas e impossíveis de serem tratadas, para cada processo tem-se uma lista das suas citações, por exemplo, da seguinte forma:

Listagem 3.2: Exemplo a lista final de citações associadas a um acórdão.

```
[
   'Artigo_734_Código de Processo Civil',
   'Artigo_527_Código de Processo Civil',
   'Artigo_1362_Código Civil',
   'Artigo_868_Código da Estrada',
   'Artigo_70_Lei n 98/2009'
]
```

Como resultado deste processamento, obteve-se um total de **28985** artigos legislativos diferentes citados em todo o *dataset*. Como tal, para se poder usar esta informação em contextos de aprendizagem e previsão, mais uma vez, foi necessário reduzir a dimensionalidade da mesma. O conceito para esse efeito foi baseado no agrupamento de artigos legislativos consoante a sua ocorrência simultânea nos acórdãos. Isto é, artigos que aparecem maioritariamente citados juntos, pertencerão à mesma categoria. Neste sentido efetuou-se o seguinte *pipeline*:

1. Construiu-se uma matriz de coocorrência de *artigo vs artigo*. Ou seja, cada célula representa o número de vezes que os dois artigos aparecem citados em simultâneo no mesmo texto.

2. Aplicou-se a métrica de similaridade por cosseno[3] à matriz anterior. Desta forma obteve-se uma nova matriz, agora com as similaridades entre os artigos segundo a sua ocorrência.

3. De seguida foi aplicado um algoritmo de *clustering*, nomeadamente o *spectral clustering*[4], à matriz de similaridade para reduzir a dimensionalidade. O número de *clusters* é um hyperparâmetro e por isso foi alvo de avaliação tendo em conta o desempenho final dos modelos. O número que, em média, proporcionou melhores resultados em todos os tópicos foi o 15.

[3] A similaridade por cosseno é uma métrica que calcula a similaridade entre vetores tendo em conta o cosseno do ângulo compreendido entre eles. Caso os vetores sejam iguais a similaridade é 1.

[4] O *spectral clustering* é umas das técnicas mais utilizadas no agrupamento de dados baseados em matrizes e que proporciona melhores resultados.

4. Finalmente, como em cada acórdão são citados inúmeros artigos e cada artigo tem uma categoria correspondente, para associar uma categoria geral àquele acórdão, calculou-se a categoria mais representada no mesmo.

Em suma, o resumo do *pipeline* do processamento de extração do conhecimento relativo aos artigos citados está presente na Figura 3.9. Note-se que este *pipeline* foi desenvolvido para cada tópico. Isto é, para o conjunto de acórdãos de cada tópico passam a existir 15 *clusters* de artigos legislativos.

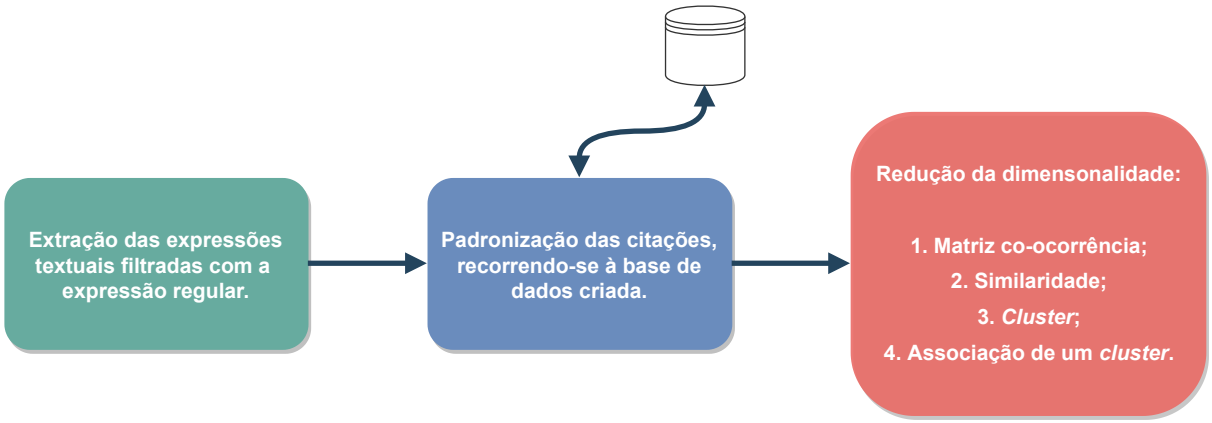


Figura 3.9: Resumo do processo usado para a extração do conhecimento relativo aos artigos legislativos citados nos textos.

3.1.4.4 Sentimentos e emoções

Outra forma de obtenção de conhecimento dos dados é identificar a carga emocional e sentimental presente nos textos da decisão para depois verificar a correlação com a decisão. Neste sentido, recorreu-se ao *NRC Emotional Lexicon* em português, um léxico com uma lista de 14182 palavras e as suas associações a dois sentimentos: positivo e negativo; e a oito emoções básicas: raiva, medo, tristeza, alegria, surpresa, confiança, entusiasmo e repulsa. Cada associação é representada por 1 ou 0, caso a palavra esteja ou não associada à emoção ou sentimento [14].

Posteriormente, efetuou-se uma correspondência entre o *dataset* deste projeto e este *NRC Emotional Lexicon*. O algoritmo de correlação foi simples: para cada palavra do texto extraiu-se o vetor de associações, no final do texto somaram-se todas as associações de cada emoção/sentimento e dividiu-se pelo número total de associações, por forma a obter a percentagem de cada emoção presente no texto e a sua carga sentimental (positiva ou negativa).

Portanto no fim da aplicação desta técnica, obteve-se a matriz:

$$DS_{axe} \tag{3.2}$$

onde,

a é o número total de acórdãos;

e são as 8 emoções e os 2 sentimentos.

Deste modo, cada célula da matriz *DS* representa a percentagem da emoção/sentimento e no acórdão a.

3.1.5 Tecnologias

Este projeto de dissertação foi maioritariamente desenvolvido recorrendo à linguagem de programação *Python*. Esta linguagem teve um papel fundamental em todas as etapas do projeto, tanto na extração dos dados, análise e preparação dos mesmos, como também na construção e avaliação dos modelos. Além de *Python* ser uma linguagem prática, simples e versátil, uma das suas principais características é o conjunto enorme de bibliotecas que podem ser usadas e que nos fornecem, de uma maneira rápida, várias funcionalidades já implementadas. Portanto, como em todos os projetos baseados nesta linguagem, usaram-se várias bibliotecas, destacam-se algumas mais relevantes, tais como: *BeautifulSoup*, *Spacy*, *Pandas*, *Seaborn*, *Nltk*, *Numpy*, *Gensim*, etc.

Por outro lado, no que toca a recursos computacionais, devido ao volume elevado de dados e, principalmente, por grande parte ser de tipologia textual, a computação requerida é elevada. Deste modo, recorreu-se à computação na *cloud*, mais propriamente à *Google Cloud Platform* para desenvolver o trabalho. Apesar de se variar entre ambientes de desenvolvimento, grande parte do trabalho foi desenvolvido numa máquina virtual com um *CPU Intel Haswell* com o sistema operativo *Debian* e 45Gb de *RAM*.

3.2 Experiências

Nesta secção é apresentada a componente da solução desenvolvida referente às experiências realizadas com os dados. Aqui pretende-se expor o conjunto de mecanismos aplicados aos dados para extrair conhecimento. A divisão da secção está relacionada com as duas grandes experiências realizadas: análise de sentimentos e previsão da decisão.

A primeira caracteriza-se pela análise da carga emocional e sentimental presente nos textos e a sua relação com a decisão do juiz. O estudo destas relações é importante, uma vez que vai de encontro aos objetivos definidos, nomeadamente à aplicação de algoritmos analíticos e descritivos para desvendar padrões nos dados que direcionem as decisões judiciais.

A segunda experiência passa pelo desenvolvimento de um sistema capaz de prever as decisões judiciais, partindo das informações e conhecimentos extraídos dos dados anteriormente. Para além disso,

como resultado desta modelação, pretende-se analisar e compreender a relação entre as informações extraídas dos textos com a decisão do juiz, visando, mais uma vez, atingir os objetivos propostos.

3.2.1 Análise de sentimentos

Nesta secção é apresentada a primeira experiência realizada, que foi subdividida em dois ensaios. O primeiro caracteriza-se pelo estudo da relação entre a carga emocional e sentimental e o segundo passa pela análise da relação entre a carga emocional e a decisão do acórdão.

3.2.1.1 Relação emoções-sentimento

Segundo a psicologia o sentimento é o resultado de uma experiência emocional, ou seja, é caracterizado pela interpretação do cérebro à experiência, física ou não, que a emoção proporcionou. Por exemplo, numa situação do quotidiano em que uma pessoa sinta pânico de algo, normalmente, é gerado um sentimento negativo.

Neste sentido, o primeiro ensaio passou pela análise da relação entre a carga emocional presente no texto e o sentimento global do texto, que neste caso apenas pode ser um sentimento positivo, negativo ou neutro.

Para esse efeito elaborou-se um gráfico de dispersão bidimensional de forma a poder-se visualizar a relação entre essas duas componentes. O objetivo passa por verificar se existe uma dispersão espacial no gráfico, que possibilite associar determinadas emoções a sentimentos. Cada ponto do gráfico representa as emoções e a sua cor é representativa do sentimento do texto, como demonstra a Figura 3.10.

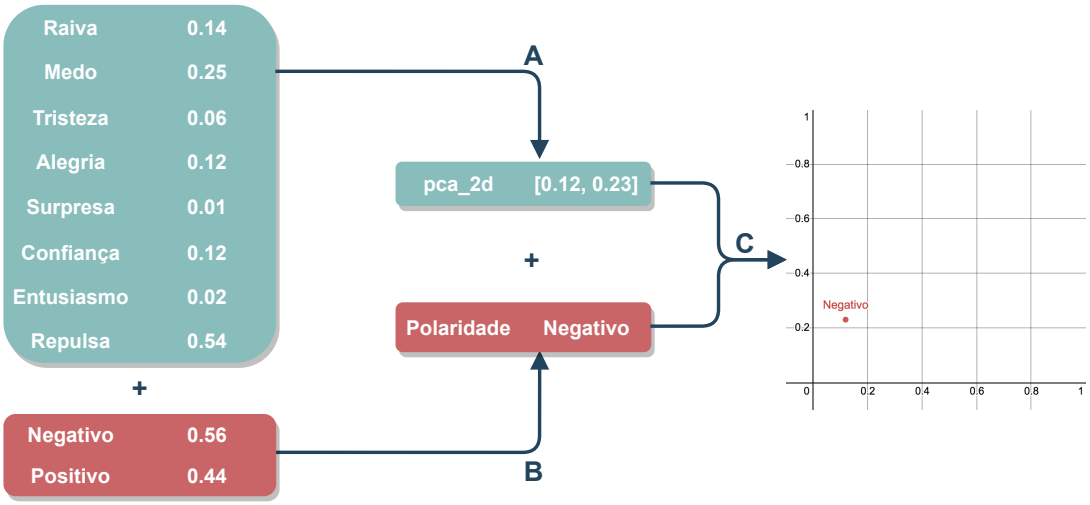


Figura 3.10: Processo de construção do gráfico de dispersão para análise da relação emoções-sentimento. Numa primeira fase, reduziu-se a dimensionalidade das emoções de 8 para 2, usando a técnica de Principal Components Analysis (PCA) (A). Em paralelo calculou-se a polaridade do texto, ou seja, aquele sentimento com mais percentagem é o que é atribuído (B). No fim, tendo as coordenadas e a polaridade, construiu-se o gráfico de dispersão (C).

3.2.1.2 Relação emoções-decisão

O segundo ensaio passou pelo estudo da relação entre a carga emocional do texto e a decisão do juiz, que foi subdividido em duas experiências.

A primeira forma de estudo desta relação foi, tal como no caso anterior, através de um gráfico de dispersão entre as duas componentes. O processo da construção do gráfico foi o mesmo da Figura 3.10, no entanto, neste caso, cada ponto do gráfico representa as emoções e a sua cor é representativa da decisão do juiz.

A segunda forma de estudo desta relação consistiu numa análise mais exaustiva da correlação entre o conjunto de emoções e a decisão. Para tal definiu-se o problema como um problema de classificação, isto é, de previsão da decisão tendo como principais variáveis explicativas, isto é, *input* dos modelos, o vetor de emoções presentes no texto.

Como a variável decisão é categórica usaram-se alguns modelos de classificação simples para auferir aquele que se ajustava melhor aos dados e, consequentemente apresentaria melhores resultados na previsão, tais como:

- **SVM (Support Vector Machine):** O primeiro modelo construído foi um SVM (Secção 2.4.1) recorrendo-se ao módulo *SVC* da biblioteca *Sklearn*.
- **DT (Decision Tree):** O segundo modelo foi uma árvore de decisão (Secção 2.4.3), usando-se o módulo *DecisionTreeClassifier* da biblioteca *Sklearn*.
- **MLP (MLP Classifier):** O terceiro modelo implementado foi MLP, uma rede neuronal artificial *feedforward* multicamada (Secção 2.4.5), recorrendo-se ao módulo *MLPClassifier* da biblioteca *Sklearn*.
- **KNN (K Neighbors Classifier):** O quarto e último modelo treinado e testado, foi o KNN (Secção 2.4.4), utilizando-se o módulo *KNeighborsClassifier* também da biblioteca *Sklearn*.

Todos os modelos foram treinados e testados segundo o modelo de validação cruzada *K-Fold*. Esta técnica caracteriza-se pela divisão do conjunto de dados em K subconjuntos e em cada iteração do procedimento, os *(k-1)* subconjuntos são usados para treino e o restante para teste do modelo. A principal vantagem deste método é a sua generalização, já que permite que todos os dados façam parte, pelo menos uma vez, do conjunto de treino. Neste projeto de dissertação foi usado um valor de K igual a 10.

Para além disso, de forma obter os melhores modelos, efetuou-se um ajuste dos hiperparâmetros de cada modelo, recorrendo-se a um método de pesquisa bruta, nomeadamente, ao módulo *GridSearch* da

biblioteca *Sklearn*. É um método que avalia o modelo com todas as combinações possíveis de hiperparâmetros pré-estabelecidos, tabela 3.6.

Tabela 3.6: Conjunto de parâmetros usados para ajustar os modelos.

Modelo	Parâmetro	Valores
SVM	C	[0.1, 1, 10, 100, 1000]
	Gamma	[1, 0.1, 0.01, 0.001]
	kernel	['rbf', 'sigmoid']
DT	Criterion	['gini','entropy']
	Max_depth	[5, 10, 15, 20, 30, 40, 50, 70, 100]
	Min_samples_leaf	[1, 2, 5]
MLP	Solver	['lbfgs','adam']
	Max_iter	[100, 200, 400, 800, 1000]
	Hidden_layer_sizes	[(50,50,50), (50,100,50), (100,)]
KNN	N_neighbors	[3, 4, 5, 8,10]
	Weights	['uniform','distance']

A métrica de avaliação usada em cada iteração do processo de validação cruzada foi a métrica simples *Accuracy*, que contabiliza a percentagem de acerto do modelo, isto é, as vezes que o modelo previu de forma correta a decisão do acórdão.

3.2.2 Previsão da decisão

A segunda grande experiência realizada passou pelo desenvolvimento de um sistema geral de previsão da decisão judicial, tendo em conta todo o conhecimento extraído do texto.

Neste sentido, definiu-se o problema de previsão como um problema de classificação binária, isto é, previsão da decisão como procedente ou improcedente. Todos os acórdãos com outro tipo de decisão foram ignorados, já que, contêm muita variedade de tipos de decisão e não permitem a criação de bons modelos preditivos e consequente extração de conclusões.

Para efeitos de análise posterior definiu-se a decisão dos acórdãos improcedentes com -1 e procedentes com 1, o que permitiu analisar os pesos resultantes do treino do modelo, isto é, variáveis com pesos positivos serão maioritariamente indicativas de procedência e as variáveis com pesos negativos serão mais indicativas de improcedência.

Quanto ao desenvolvimento dos modelos, optou-se por implementar um sistema distribuído, isto é, treinar um modelo para cada tópico, o que permite auferir, para cada área temática de acórdãos, as variáveis que mais influenciam a decisão.

Posteriormente, para encontrar a melhor combinação de variáveis que permitem a previsão da decisão, para cada tópico foram definidos, treinados e testados vários cenários, tabela 3.7.

Tabela 3.7: Combinações de variáveis usadas para a previsão da decisão.

Cenário 1	Termos mais frequentes + Cluster legislação
Cenário 2	Termos mais frequentes + Cluster legislação + Ano
Cenário 3	Termos mais frequentes + Cluster legislação + Juiz

Todos estes cenários foram usados para treinar e testar modelos de SVM. Ao contrário da experiência anterior, onde se testaram diferentes modelos de ML, neste caso, apenas se implementaram modelos SVM, por várias razões:

- São modelos que apresentam bons resultados em problemas de classificação usando variáveis textuais;
- Em grande parte da literatura relacionada com o projeto desta dissertação, estes modelos são os que apresentam melhores resultados;
- São modelos que, usando um *kernel* linear, permitem obter várias conclusões explicativas, permitindo interpretar os resultados [3]. Neste problema não se pretende um sistema "caixa negra", ou seja, que não permita entender o porquê dos resultados. Portanto, este ponto foi fundamental nesta escolha.

Tal como na experiência anterior, todos estes modelos foram treinados e testados segundo o modelo de validação cruzada *K-Fold*. Quanto à hyperparametrização, como o *kernel* usado é linear, apenas se ajustou o parâmetro de regularização do erro, C, que se fez variar entre os valores do intervalo $[1, 10, 100]$.

A precisão dos modelos foi calculada segundo a métrica simples de *Accuracy* que contabiliza a percentagem de acerto do modelo, ou seja, o rácio entre o número de vezes que o algoritmo acertou na previsão da decisão e o número de amostras. A *Accuracy* final corresponde à media do resultado obtido nos 10 *folds*.

Em suma, o *pipeline* do desenvolvimento dos modelos está exemplificado no esquema da Figura 3.11.

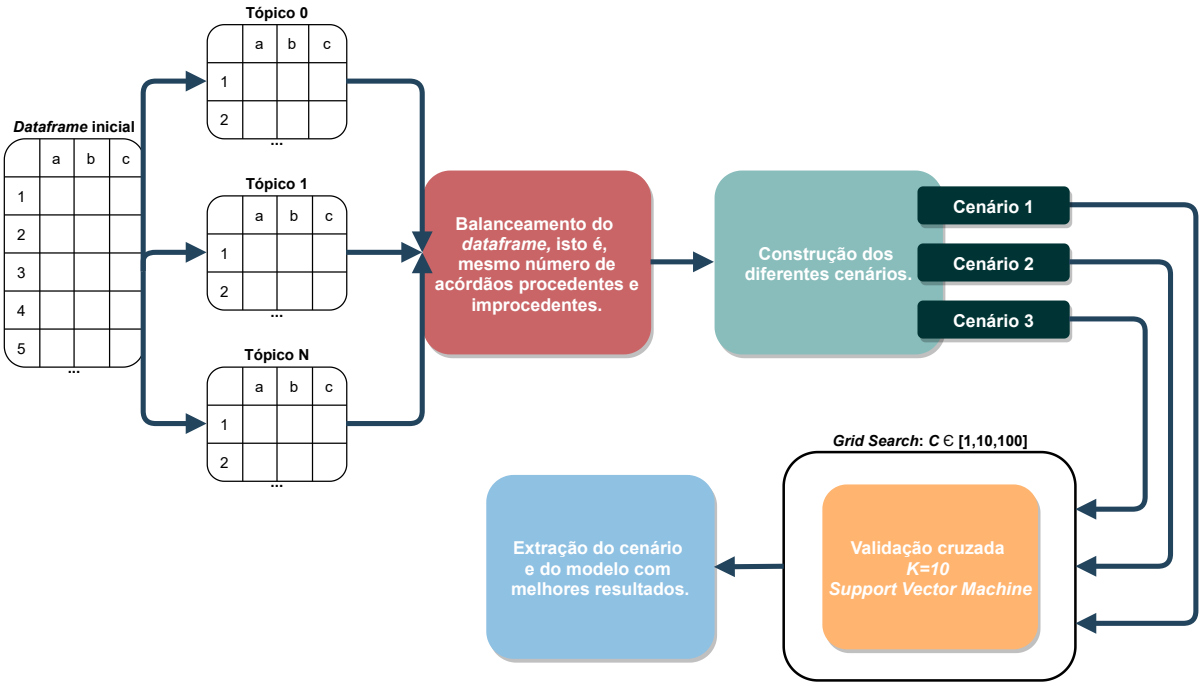


Figura 3.11: *Pipeline* de desenvolvimento dos modelos preditivos. Depois de se computar e associar a cada processo o seu tópico, separou-se o *dataframe* inicial tendo em conta o tópico correspondente. Posteriormente, para cada tópico foi calculado o número de acórdãos da menor classe da variável decisão, e feito um balanceamento, isto é, definir o mesmo número de acórdãos procedentes e improcedentes. Seguidamente, passou-se à criação das três combinações de variáveis e finalmente treinaram-se e testaram-se os modelos recorrendo ao modelo de validação cruzada, ajustando o hyperparâmetro *C*, para que no fim se obtive-se o cenário e o modelo com melhores resultados.

3.3 Visualização de informação

Em paralelo ao processo de modelação e previsão das decisões judiciais, desenvolveu-se um protótipo de uma *dashboard*, indo de encontro ao objetivo inicialmente definido onde se pretendia uma visualização dos dados através de um painel de controle.

No cômputo geral, as *dashboards* são painéis que exibem métricas e indicadores para a monitorização dos resultados de uma entidade ou empresa. No entanto, no contexto desta dissertação a *dashboard* desenvolvida tem como finalidade a otimização da pesquisa de jurisprudência, a pesquisa de legislação, a visualização de estatística sobre os tribunais nacionais e ainda a correlação de vários parâmetros apresentada graficamente.

O principal propósito destes sistemas de auxílio dos profissionais de direito, sejam eles advogados, juízes ou procuradores é essencialmente a otimização das tarefas diárias que são demoradas e repetitivas. Neste caso, para além disso, pretende-se que este também seja, num futuro, um sistema de apoio à tomada de decisão já que contempla várias informações de decisões passadas, que podem servir de auxílio para decisões futuras semelhantes.

3.3.1 Desenvolvimento

No que toca ao desenvolvimento deste sistema, optou-se por dividi-lo em duas aplicações: o *frontend* e o *backend*. O *frontend* é a aplicação responsável pela interação com o utilizador, enquanto que o *backend* trata de todo o processamento dos dados. Quando às tecnologias e *frameworks* optou-se por utilizar a *stack* tecnológica apresentada na Figura 3.12.

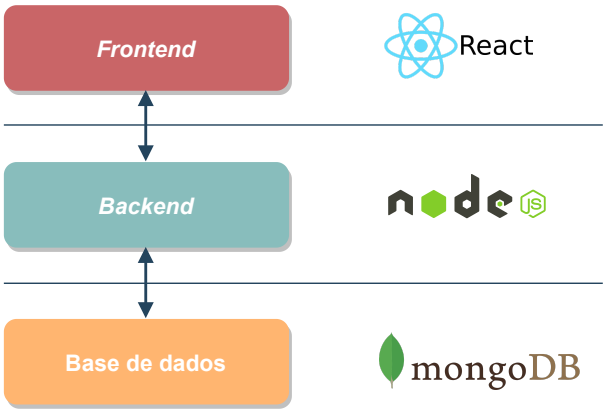


Figura 3.12: *Stack* tecnológica utilizada no desenvolvimento do sistema.

Relativamente à camada de dados, decidiu usar-se *MongoDB* para a persistência dos dados, uma vez que este motor está bem integrado com a *framework* utilizada na camada intermédia e por isso permite uma fácil conexão. Para além disso utiliza base de dados orientadas a documentos, que são as que mais se ajustam à tipologia dos dados deste projeto. Os dados são os mesmos que se usaram na previsão, ou seja, os acórdãos dos tribunais da relação já pré-processados e com a adição do conhecimento extraído, que foi apresentado na secção 3.1.4.

No que diz respeito ao *backend*, utilizou-se *NodeJS*, uma *framework* de *javascript*. Esta escolha deveu-se ao facto da grande familiarização com a mesma e ainda porque permite a fácil criação de uma *API REST*. Por fim, na camada de apresentação, no *frontend*, utilizou-se *ReactJs*, uma *framework* muito usada nos dias de hoje no desenvolvimento *web*. Para além disso apresenta uma robusta documentação e muitas bibliotecas que permitem a replicação de componentes de uma forma rápida.

3.3.2 Principais funcionalidades

Esta secção tem como objetivo apresentar as principais características do protótipo desenvolvido. Decidiu dividir-se a aplicação em quatro principais funcionalidades que são abordadas mais detalhadamente nas secções seguintes.

3.3.2.1 Estatísticas globais

Partindo do conceito de disponibilizar ao utilizador uma ampla gama de estatística sobre o estado da justiça em Portugal, ao entrar na aplicação é apresentada uma página inicial contendo estatísticas globais sobre os dados presentes na base de dados e ainda sobre a eficácia dos tribunais portugueses, estas últimas são informações recolhidas da fonte externa *PORDATA*[5].

A Figura 3.13 apresenta parte da página inicial onde é exibido o número de acórdãos da base de dados, o número de juízes, o número de tribunais e um gráfico que contém a evolução da taxa de eficácia dos tribunais da relação ao longo dos anos. A restante informação apresentada nesta página pode ser consultada em apêndice na Figura D.1.

Figura 3.13: Excerto da página inicial da aplicação.

3.3.2.2 Pesquisa simples

Uma das funcionalidades da aplicação é um motor de busca simples. Este motor caracteriza-se pela aplicação de filtros simples aos dados, nomeadamente, a filtragem dos acórdãos por número de processo, ano, juiz, tribunal, ECLI e área temática. Como resultado da pesquisa é apresentada uma lista com a meta informação de cada acórdão correspondentes ao filtro. Além disto também se dá a possibilidade ao utilizador de, para cada acórdão, expandir a sua informação mais detalhada, como exemplifica a Figura 3.14.

[5] https://www.pordata.pt/Home

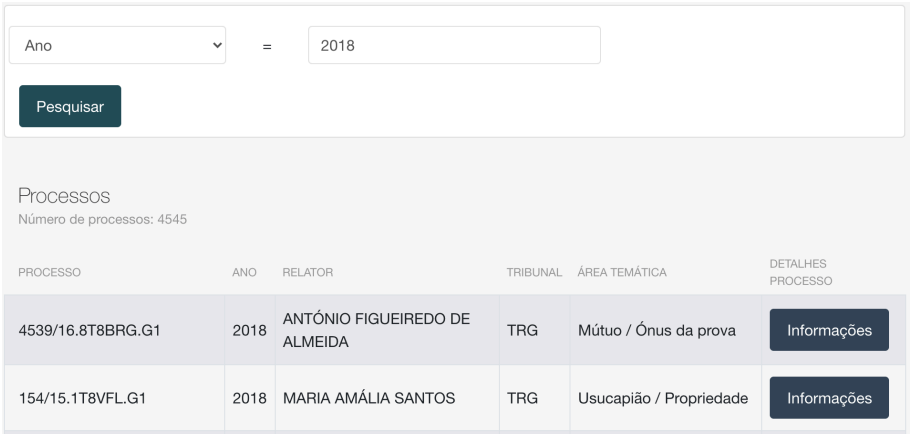


Figura 3.14: Excerto da página da funcionalidade de pesquisa simples com a aplicação do filtro de ano igual a 2008.

3.3.2.3 Pesquisa avançada

Ao contrário do motor de busca simples, esta funcionalidade permite conjugar vários filtros obtendo-se uma informação mais precisa daquilo que é desejado. Neste sentido, é possível combinar filtros de quatro variáveis: área temática, descritores, tribunal e juiz.

Como resultado, para além da lista dos acórdãos selecionados através do filtro, são apresentados gráficos que, dependendo da conjugação escolhida, podem ser diferentes. Por exemplo, no caso do utilizador apenas selecionar a área temática e o tribunal, são fornecidos três gráficos: o primeiro sobre os oito juizes que mais decidiram nos acórdãos filtrados e a percentagem de cada tipo de decisão que tomaram, Figura 3.15, o segundo sobre os descritores mais frequentes e ainda um mais genérico com a percentagem global de cada tipo de decisão, estes últimos podem ser consultados em apêndice na Figura D.3.

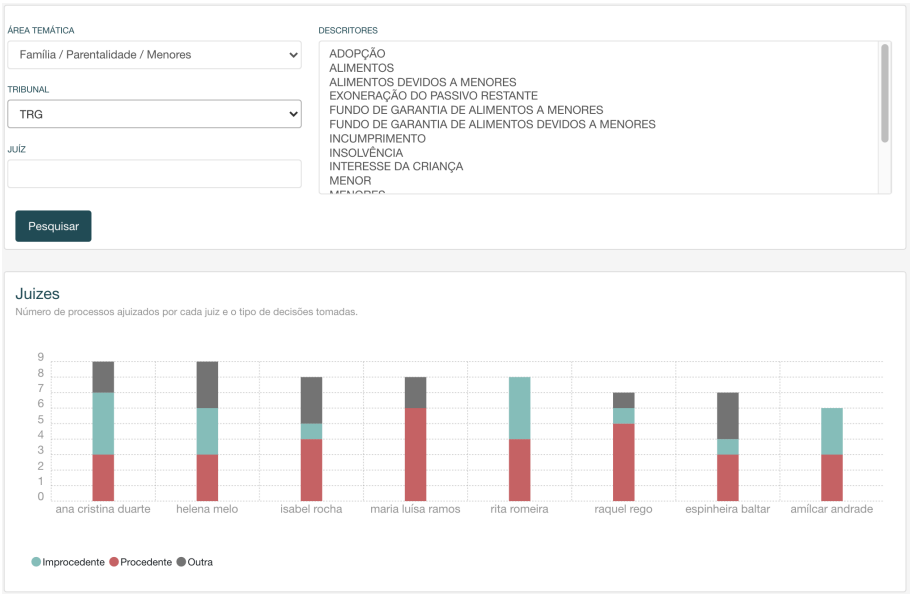


Figura 3.15: Excerto da página da funcionalidade de pesquisa avançada com a aplicação do filtro da área temática igual a "Família/Parentalidade/Menores"e o tribunal igual a Tribunal da Relação de Guimarães.

3.3.2.4 Legislação

Por fim, implementou-se uma funcionalidade para apresentar informações da relação de artigos citados e decisão do acórdão. Deste modo, é possível procurar por acórdãos de uma área temática nos quais um determinado artigo legislativo foi referenciado. Tal como nos casos anteriores é apresentada a lista de acórdãos selecionados pelo filtro mas também dois gráficos: um referente à percentagem de cada tipo de decisão nos acórdãos filtrados e outro relativo aos artigos legislativos mais referenciados em acórdãos da área temática selecionada. A Figura 3.16 apresenta um exemplo desta funcionalidade.

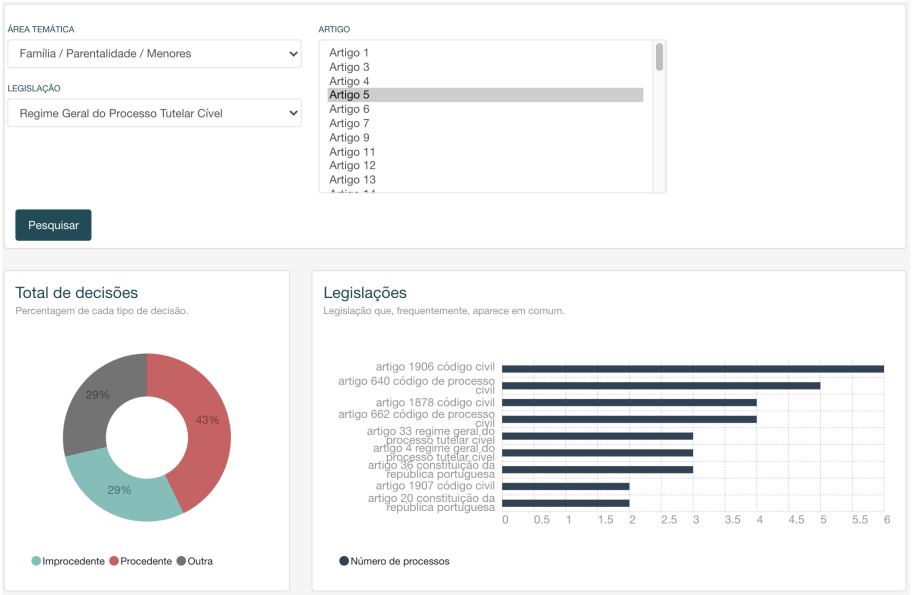


Figura 3.16: Excerto da página da funcionalidade de pesquisa de legislação com a aplicação do filtro de área temática igual a "Família/Parentalidade/Menores"com o artigo 5 do Regime Geral do Processo Tutelar Cível. Através desta informação pode-se tecer várias conclusões, como por exemplo, a de que em 43% dos acórdãos desta área temática em que este artigo foi referenciado a decisão tomada foi Procedente, e que o artigo legislativo mais usado pelos juízes, na sua fundamentação, em acórdãos desta área é o 1906 do Código Civil.

Capítulo 4

Resultados e discussão

O presente capítulo visa apresentar, interpretar e discutir os resultados obtidos neste projeto. Neste sentido o capítulo divide-se em três partes: a primeira relativa à experiência de análise de sentimentos, a segunda diz respeito aos resultados obtidos na experiência de previsão da decisão e termina com um tópico referente à componente de visualização de informação.

4.1 Análise de sentimentos

Nesta secção são apresentados os resultados obtidos com a realização da primeira experiência. A divisão da secção diz respeito aos dois ensaios realizados.

4.1.1 Relação emoções-sentimento

Como resultado do primeiro ensaio que foi realizada e que diz respeito à análise da relação entre a carga emociona e sentimental no texto, obteve-se o gráfico de dispersão da figura 4.1.

De facto, à primeira vista é possível verificar uma ligeira diferença das distribuições das duas polaridades (Negativa, Positiva), isto é, uma concentração de acórdãos com carga negativa mais à esquerda, e positiva mais à direita do gráfico.

No nosso quotidiano é normal que várias emoções estejam geralmente relacionadas com sentimentos. No entanto, no mundo judicial é natural que assim não o seja. Os juízes são por norma muito pouco

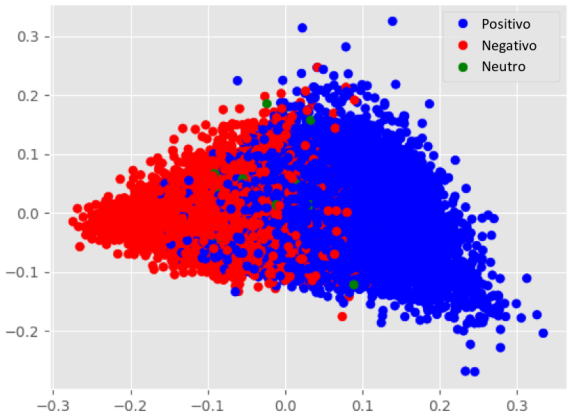


Figura 4.1: Distribuição no espaço bidimensional dos processos segundo a sua carga emocional e polaridade sentimental (Negativa, Positiva, Neutra)

expressivos sentimentalmente. Deste modo, se por um lado se verifica uma ligeira diferença no gráfico, também é notório que não é possível identificar um padrão muito explicativo que nos permitam tecer conclusões mais concretas relativamente a esta relação. Para tal acontecer, os dados das diferentes categorias, graficamente, teriam de apresentar uma dispersão maior entre as polaridades o que permitia concluir, por exemplo, que as emoções *x,y,z* estão, maioritariamente relacionadas com o sentimento positivo ou negativo. Com este resultado tal não é possível, já que a diferença é muito diminuta.

4.1.2 Relação emoções-decisão

Já no que toca ao segundo ensaio, o gráfico de dispersão que se obteve foi o da figura 4.2.

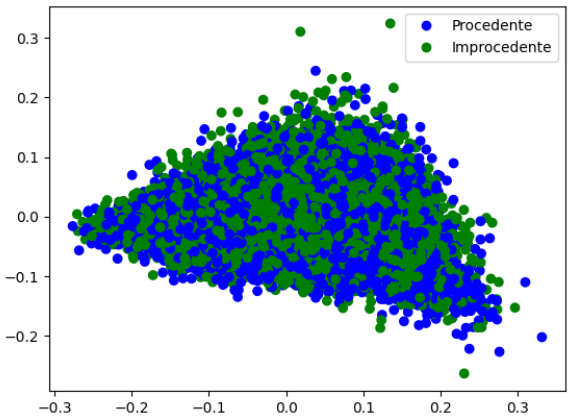


Figura 4.2: Distribuição no espaço bidimensional, dos processos segundo a sua carga emocional e a decisão do tribunal.

Os resultados foram semelhantes aos anteriores, o que vem reforçar a ideia de que não é possível

auferir padrões nos dados. Neste caso, ainda é mais vincada a aleatoriedade dos dados, o que nos demonstra a pouca correlação existente entre as emoções e a decisão.

Ainda assim, efetuou-se a análise mais exaustiva recorrendo a modelos preditivos e os resultados obtidos foram os presentes na tabela 4.1.

Tabela 4.1: Precisão dos diferentes modelos implementados. A coluna "3 categorias"corresponde aos resultados com as 3 variáveis dependentes (Procedente, Improcedente, Outra), a coluna "2 categorias"diz respeito aos resultados dos modelos cujos dados apenas contêm as 2 variáveis principais (Procedente, Improcedente). A negrito está assinalado o melhor resultado.

Modelo	**Melhores parâmetros**	***Acc* (%) 3 categorias**	***Acc* (%) 2 categorias**
SVM	C = 10 Gamma = 0.1 Kernel = 'rbf'	48.02	**62.10**
DT	Criterion = 'entropy' Max_depth = 5 Min_samples_leaf = 5	49.41	60.05
MLP	Solver = 'lbfgs' Max_iter = 1000 Hidden_layer_sizes = (100,)	49.01	59.10
KNN	N_neighbors = 10 Weights = 'uniform'	44.20	55.71

De facto, analisando a tabela dos resultados verifica-se que o desempenho dos modelos não é muito favorável ao compararmos com os modelos base de aleatoriedade. Ainda assim, é de destacar que a remoção dos processos cuja decisão pertence à categoria "Outra", melhorou cerca de 10% a precisão dos modelos. Isto deve-se ao facto desta variável englobar uma grande heterogeneidade de decisões e como tal introduz enviesamento nos modelos.

Na prática, o melhor modelo implementado, com 62.1% de precisão, permite auferir que, em cerca de 62.1% dos casos de teste, o modelo acerta na decisão do tribunal, tendo por base apenas a carga emocional do texto de decisões passadas. Como existem duas variáveis *outputs*, o valor de *accuracy* de referência seria 50%, e portanto com os resultados obtidos, pode-se afirmar que o melhor modelo supera o modelo base em 12%, que, de facto, é pouco satisfatório.

Deste modo, o modelo não é explicativo o que não permite afirmar que, a carga emocional presente no conteúdo textual de um processo está diretamente relacionada com a decisão tomada. Na verdade,

os maus resultados obtidos já eram esperados. À partida para este estudo, já se tinha uma ideia pré-concebida da prevalência da racionalidade sobre a emoção dos juízes no ato de julgar. No entanto, o objetivo passava por investigar se, computacionalmente, seria possível extrair a carga emocional dos textos e encontrar correlações nos dados, corroboradas pelos resultados da investigação, que permitissem tecer conclusões sobre a relação emoção-decisão.

Uma explicação jurídica para estes resultados é que um dos princípios mais relevantes de um juiz é o equilíbrio emocional, isto é, o juiz deve ter um domínio sobre as suas emoções, não se deixando levar por elas. Um juiz tem de ser sensato, sereno e principalmente imparcial, nunca tomando partido de um dos lados, permitindo uma abstração emocional que lhe permite decidir.

4.2 Previsão da decisão

No que toca à segunda grande experiência efetuada, os resultados do *pipeline* da figura 3.11 estão apresentados na tabela 4.2.

Tabela 4.2: Precisão dos modelos de cada tópico tendo em conta os diferentes cenários de conjuntos de variáveis.

Tópico	Descrição	*Acc (%)* Cenário 1	*Acc (%)* Cenário 2	*Acc (%)* Cenário 3
0	Prescrição de penas / Caducidade	78	67	68
1	Insolvência / Crédito	77	64	67
2	Nulidade / Contraordenação	84	57	62
3	Provas / Ónus da prova / Factos	72	65	64
4	Família / Paternidade	75	60	68
5	Resolução contratuais / Contrato-Promessa	74	55	57
6	Usucapião / Propriedade	75	70	73
7	Dívidas / Penhoras / Execuções	78	67	70
8	Inventário / Arresto	76	61	62
11	Acidente de viação / Indeminização	79	62	61
12	Competência / Tribunal competente	77	65	66
13	Apoio judiciário / Taxas justiça	77	61	63
14	Expropriação	75	66	68
15	Acidente trabalho / Incapacidade	73	71	73
16	Mútuo / Ónus da prova	74	63	63
17	Contrato trabalho / Despedimento	80	66	67
19	Arrendamento / Obras	75	63	68
	Média	**76**	**64**	**66**

Partindo destes resultados, o primeiro facto a ser auferido é que o **cenário 1** de combinações de variáveis é aquele que proporcionou modelos com melhores resultados. Em consequência deste facto pode-se afirmar que tanto a variável "Ano", no **cenário 2**, e a variável "Juiz", no **cenário 3**, não acrescentam conhecimento ao modelo, ou seja, a decisão do acórdão não é influenciada pelo ano em que foi escrito, nem pelo juiz que relatou o acórdão.

Para além disto é notório que tópicos cuja quantidade de dados é maior apresentam melhores resultados, o que indica que a *accuracy* dos modelos depende do número acórdãos usado no processo de treino. Por este motivo, estes modelos serão passíveis de serem melhorados ao longo do tempo com o treino de novas entradas de dados. Ainda assim, explorando a *accuracy* obtida, atingir uma média de 76% e ter modelos em tópicos a ultrapassar os 80%, já é matéria substancial para se poder assumir que as informações textuais extraídas estão algo relacionados com a decisão do juiz. Para além disso ao estabelecer uma relação com outros trabalhos relacionados, nomeadamente de previsão de decisões judiciais, apresentados na revisão da literatura, é percetível que estes resultados estão num patamar idêntico, superando até grande parte deles.

Como o cenário que proporcionou melhores resultados foi o 1, decidiu-se analisar mais exaustivamente esses modelos, em duas vertentes: relação da legislação citada com a decisão e a relação dos termos mais frequentes com a decisão.

No que diz respeito à primeira relação, analisou-se a correlação entre os *clusters* de legislação e a categoria da decisão. Note-se, mais uma vez, que cada *cluster* de legislação representa o conjunto de artigos que geralmente são citados em conjunto. Deste modo, construíram-se gráficos de barras para cada tópico, onde a cada *cluster* corresponde o número de acórdãos procedentes (vermelho) e improcedentes (verde). As figuras 4.3 e 4.4, representam esses gráficos para o tópico 1 - "Insolvência / Crédito"e 19 - "Arrendamento / Obras", respetivamente.

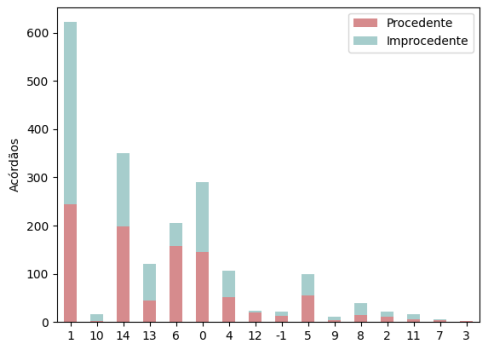


Figura 4.3: Distribuição dos acórdãos pelos diferentes *clusters* de legislação e pelas duas categorias de decisão no tópico 1.

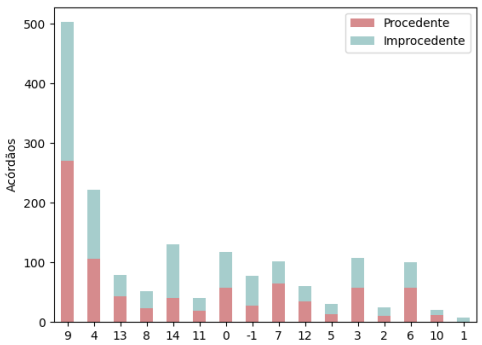


Figura 4.4: Distribuição dos acórdãos pelos diferentes *clusters* de legislação e pelas duas categorias de decisão no tópico 19.

Como é possível visualizar nestes gráficos, existe uma distribuição uniforme das duas categorias de decisão em cada *cluster* de artigos citados. Isto é, a maior parte dos *clusters* apresenta o mesmo número de acórdãos com decisão procedente e improcedente. Deste modo é possível concluir que a citação dos artigos legislativos não está diretamente associada à procedência ou improcedência dos recursos. Os gráficos dos outros tópicos podem ser consultados no apêndice C.

Quanto à segunda relação, termos mais frequentes com a decisão, tirando partido do uso de um *kernel* linear no modelo implementado, é possível analisar os termos que mais influenciam, tanto na previsão de procedência como de improcedência de recursos. Neste sentido, a tabela 4.3 apresenta, para cada tópico, os quatro termos com maiores e menores pesos. Aqueles termos com maiores pesos, são os que têm maior relação com o facto do acórdão ser ajuizado como procedente e, por outro lado, os que têm menores pesos são os que têm mais influência na decisão improcedente. À primeira vista, estes termos não apresentam informação e apenas são encarados como palavras soltas, uma vez que não se tem o conhecimento sobre o motivo que deu origem ao recurso. No entanto, num trabalho futuro, se fosse possível saber o que levou uma das partes a recorrer da decisão de um tribunal inferior, teríamos um contexto perfeito que nos permitiria perceber a associação das palavras à decisão.

Tabela 4.3: Análise dos pesos dos modelos implementados. Apresentação dos termos com maiores e menores pesos do melhor modelo *SVM* obtido, ou seja, aqueles que têm maior influência na decisão procedente e improcedente.

Tópico	Descrição	Termos com pesos maiores	Termos com pesos menores
0	Prescrição de penas / Caducidade	formular 9.89 parcelar 4.16 companhia 4.11 vencimento 4.02	defender -8.13 revogar -4.02 patente -3.82 comarca -3.43
1	Insolvência / Crédito	agravante 9.51 coletivo 6.22 rua 5.42 identificação 4.23	verdade -3.85 âmbito processar -3.12 bancário -3.02 titulado -2.98
2	Nulidade / Contraordenação	provimento recurso 6.87 acordam 6.32 editora 5.23 criminal 4.23	factualidade provar -5.23 conclusões -4.32 recebimento -4.25 emitir parecer -3.52
3	Provas / Ónus da prova / Factos	procedência 3.02 ausência 2.94 afirmar 2.67 facto assentir 2.21	provar -4.23 impedir -3.65 impugnação -3.42 especificar -2.32
4	Família / Paternidade	remuneração 5.23 instituto 3.23 assumir 2.98 nascer 2.32	futuro -4.32 residir -3.84 provar facto -2.78 salvaguardar -2.54
5	Resolução contratuais / Contrato-Promessa	elemento 6.97 permitir 5.68 formulação 4.56 matéria 3.87	litígio -4.29 julgar -4.20 concordar -3.56 recorrer -2.75
6	Usucapião / Propriedade	recusar 4.25 indemnizar 4.02 dificuldade 3.85 qualidade 2.58	proporção -5.23 conjuntar -4.23 derivar -4.21 forçar -3.84
7	Dívidas / Penhoras / Execuções	agravar 4.23 anual 4.20 emprestar 3.48 concretização 3.25	peticionar -5.24 repetir -5.20 instauração -4.23 agente -3.89
8	Inventário / Arresto	empresa -5.23 vencer -4.76 requerer -4.56 segundar -3.89	instauração -5.23 relatório -5.23 repetir -4.69 totalidade -3.47
11	Acidente de viação / Indeminização	indemnização dano 6.34 participação 5.75 automóvel 4.67 seguro 3.58	improceder -6.23 autoridade -5.43 conhecimento -4.52 conformidade -4.23
12	Competência / Tribunal competente	provimento 7.23 proibir 6.23 violação 5.42 intenção 5.23	proferir despachar -5.25 apelante -5.23 euro -4.98 autoridade -3.53
13	Apoio judiciário / Taxas justiça	tribunal 6.98 suscetível 5.23 agravar 4.67 preescrever 3.65	completar -6.87 abrigar -6.23 regime jurídico -5.75 legitimidade -4.52
14	Expropriação	imóvel 7.24 rua 7.21 zona 6.42 indeminização 4.21	ordenamento -6.23 constitucional -6.12 matriz predial -4.23 identificar -3.76

15	Acidente trabalho / Incapacidade	atribuição 6.52 acórdão 5.23 procedente 4.23 recorrente 3.53	manter -6.23 ministério -4.42 saúde -4.21 recusar -3.98
16	Mútuo / Ónus da prova	papel 7.23 relação 6.23 código civil 5.23 prova 4.45	integralmente -5.23 nenhum -4.23 praticar -3.45 companhia -2.32
17	Contrato trabalho / Despedimento	diligência 6.23 conclusão recurso 5.23 ilegitimidade 4.23 competente 3.23	recorrente -5.23 conclusão alegação -4.23 residência -3.34 factualidade -2.24
19	Arrendamento / Obras	fiscalização 5.23 audiência julgamento 4.23 exigir 4.21 crime 3.64	improceder -6.23 tribunal -5.23 procedimento -5.21 conformar -4.23

4.3 Visualização de informação

No que diz respeito à componente da visualização de informação, a *dashboard* apresentada surge no contexto do objetido definido inicialmente que pretendia a implementação de um painel de controle para visualizar os dados. É de ressaltar que a *dashboard* ainda está numa fase embrionária, podendo até, ser considerada um protótipo, por isso é passível de melhoramentos e aumento das suas funcionalidades. Ainda assim o que se implementou já se considera bastante útil no quotidiano dos profissionais de direito, pois se por um lado esta plataforma permite uma visualização de alto nível dos dados, por outro permite gerar informações e conhecimento a partir dos dados recorrendo a mecanismos gráficos de fácil interpretação. Ainda assim é importante, no futuro, ter uma apreciação crítica de um profissional da área da justiça, para saber as potenciais melhorias a serem aplicadas.

Capítulo 5

Conclusões e trabalho futuro

A integração da tecnologia nos sistemas judiciais está cada vez mais presente da realidade atual. Também é factual que a inteligência artificial está intrinsecamente ligada aos sistemas tecnológicos que se vão desenvolvendo, e por isso, é inevitável a sua presença nos sistemas judiciais. É percetível que a tecnologia tem um poder transformacional em várias áreas da justiça, no entanto, a mais sensível, aquela que divide opiniões e que está em diferentes fases de implementação nos vários países é a justiça preditiva. Como um dos principais focos desta dissertação é precisamente nessa área, é de notar que se deu um passo importante no que toca a esta matéria, especificamente em Portugal.

No que diz respeito aos ojetivos que foram inicialmente propostos, cabe agora avaliar se foram ou não atingidos.

Num primeiro objetivo pretendia-se agregar dados semiestruturados em larga escala. Este objetivo foi atingido na medida em que a grande quantidade de dados, mais propriamente, dos diferentes tribunais foi extraída e agregada em formatos passíveis de serem processados. É de realçar que, depois de definidas as fontes dos dados, as primeiras etapas do desenvolvimento do projeto, nomeadamente a extração e preparação dos dados foram as mais morosas, tal como já é normal acontecer em projetos deste âmbito.

O segundo objetivo consistia na aplicação dinâmica de algoritmos analíticos e de previsão para desvendar padrões que direcionam as decisões judiciais. Se por um lado se fez uma análise exploratória dos dados, examinando as principais características dos mesmos, por outro efetuaram-se duas experiências que permitiram desvendar padrões nos dados. Enquanto que na primeira experiência, análise de sentimentos, se obtiveram resultados pouco satisfatórios, o que já era expectável, nomeadamente a pouca

relação entre a carga emocional dos textos dos juízes e a decisão do acórdão, na segunda experiência, previsão da decisão, obtiveram-se resultados bastante satisfatórios, uma média de *accuracy* de 76%, que permitiu auferir relações entre as variáveis textuais e a decisão judicial. Neste sentido, pode-se concluir que este objetivo também foi atingido com sucesso.

No que toca ao terceiro objetivo que passava pela análise e compreensão das variáveis relacionadas com a decisão do juiz, de facto, neste projeto não se limitou apenas à apresentação dos resultados, foi também analisado o porquê dos mesmos e deste modo, foram escrutinadas as variáveis que mais ou menos influenciavam o resultado. Portanto, este objetivo também foi alcançado.

Finalmente, o último objetivo proposto consistia no desenvolvimento de um painel de controlo com informações dos dados. A *dashboard* desenvolvida surgiu no âmbito deste objetivo. Para além de ser uma plataforma com inúmeras estatísticas dos dados, apresenta funcionalidades que vêem otimzar as tarefas recorrentes dos profissionais do direito, principalmente na pesquisa de jurisprudência e legislação.

Tendo em conta todos os objetivos atingidos e estando o projeto concluído é agora possível responder à questão de investigação inicialmente proposta, nomeadamente, "É possível um sistema, com recurso a mecanismos inteligentes, gerar informação e conhecimento útil a partir de dados jurídicos portugueses?". A resposta a esta questão começa na revisão da literatura, quando se aborda a relação da inteligência artificial com a justiça e se expõem vários casos de estudo, que desde logo, demonstram esta possibilidade. Para além disso, a resposta está patente em todo o desenvolvimento prático do projeto, uma vez que os mecanismos desenvolvidos, usando a inteligência artificial, permitiram a extração de inúmeras informações e conhecimentos a partir dos dados, tais como:

- As informações recolhidas na análise exploratória, como por exemplo, os descritores mais frequentes, os vários tipos de decisões dos juízes, a distribuição dos acórdãos por ano e mês, entre outras.
- Categorização dos acórdãos a partir do conteúdo textual.
- A extração da legislação citada nos textos dos acórdãos.
- O conhecimento dos artigos legislativos que geralmente são citados simultaneamente nos textos de cada área temática.
- A carga emotiva e sentimental presente nos textos dos juízes.
- A baixa relação entre a carga sentimental e emotiva com a decisão.
- Modelos inteligentes capazes de prever a decisão tendo em conta os termos frequentes e a legislação.

- Os termos que mais estão relacionados com a decisão procedente e improcedente.
- A baixa relação entre os artigos legislativos e a decisão.
- A visualização gráfica, através da *dashboard*, de várias informações sobre os juízes, tribunais e artigos legislativos.

Outro aspeto importante a ter em consideração, após o término do projeto, é a análise do trabalho desenvolvido, de modo a verificar se respeita os princípios elaborados pela CEPEJ sobre o uso da inteligência artificial na justiça:

- **Princípio do respeito dos direitos fundamentais:** Todos os mecanismos tanto de modelação como previsão não põem em causa os direitos fundamentais. Não substituem qualquer decisão jurídica nem afetam o direito a um julgamento justo.
- **Princípio da não discriminação:** Os padrões que foram detetados nos dados, tanto na análise exploratória como no resultado da aplicação de mecanismos de ML não demonstram padrões discriminatórios. Um dos principais motivos para isto, é o facto dos acórdãos omitiram a informação da identificação dos intervenientes nos mesmos.
- **Princípio da qualidade e da segurança:** No que diz respeito à qualidade dos mecanismos desenvolvidos, de facto, pode dizer-se que não houve um ambiente multidisciplinar no desenvolvimento do projeto, no entanto, existiram troca de ideias com os supervisores do projeto que permitiram avaliar a qualidade do mesmo. Quanto à segurança, os dados utilizados são de fontes seguras, nomeadamente de instituições oficiais portuguesas. Para além disso, o ambiente de desenvolvimento foi a *Google Cloud*, que usa infraestruturas altamente monitorizadas ao nível da segurança.
- **Princípio da transparência, imparcialidade e equidade:** Este princípio é maioritariamente aplicado a sistemas já difundidos e implementados no quotidiano, uma vez que diz respeito, principalmente, à possibilidade de efetuar auditorias aos sistemas. Como tal, neste momento não existe uma forma concreta de avaliar se este princípio é ou não respeitado.
- **Princípio "controlo do utilizador":** Tal como o anterior, este princípio também é aplicado a sistemas já difundidos, ainda assim, durante o desenvolvimento optou-se por implementar mecanismos que permitissem explicar os resultados, o que, na possibilidade de expansão deste sistema para a realidade jurídica, permitiria respeitar este princípio.

Tecnicamente, os principais desafios encontrados neste projeto foram a extração, preparação e estruturação dos dados e ainda o treino e validação dos modelos implementados, mais propriamente, devido à grande quantidade de tempo e recursos computacionais exigidos nestas tarefas. Não obstante, finalizado o trabalho, é importante salientar o grande poder das tecnologias de inteligência artificial, do processamento de linguagem natural e da análise de dados, aplicadas a dados judiciais. Por este motivo é de extrema importância a cooperação entre os profissionais da justiça e da tecnologia para acompanhar a evolução da relação justiça-tecnologia.

Como trabalho futuro pretende transformar-se todos os mecanismos desenvolvidos num sistema efetivo de apoio ao processo de tomada de decisão, mais propriamente, na incorporação na *dashboard* de um sistema de previsão. Para além disso, como foi referido durante o projeto, o mesmo foi desenvolvido com os dados dos tribunais da relação, por isso, outro trabalho futuro importante seria aplicar todo o *pipeline* aos acórdãos de outros tribunais, repare-se que, tecnicamente, o trabalho está preparado para esse efeito. Por outro lado, também se pretende uma futura cooperação com colaboradores da área da justiça para saber, efetivamente, as vantagens e as limitações deste projeto, nomeadamente, nas tarefas do seu quotidiano.

Bibliografia

[1] Tom M. Mitchell. *Machine Learning*. McGraw-Hill Science/Engineering/Math, mar. de 1997. isbn: 0070428077.

[2] Lior Rokach Oded Maimon, ed. *Data Mining and Knowledge Discovery Handbook*. Springer, 2005. isbn: 978-0-387-09822-7. doi: 10.1007/978-0-387-09823-4. url: https://tanthiamhuat.files.wordpress.com/2015/04/data_mining_and_knowledge_discovery_handbook.pdf.

[3] Yin-Wen Chang e Chih-Jen Lin. "Feature Ranking Using Linear SVM". Em: *Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008*. Ed. por Isabelle Guyon, Constantin Aliferis, Greg Cooper, André Elisseeff, Jean-Philippe Pellet, Peter Spirtes e Alexander Statnikov. Vol. 3. Proceedings of Machine Learning Research. Hong Kong: PMLR, jun. de 2008, pp. 53–64. url: http://proceedings.mlr.press/v3/chang08a.html.

[4] *Neural Networks and Learning Machines*. Third. Upper Saddle River, New Jersey: Pearson Education, 2009. isbn: 978-1-4200-8964-6.

[5] António Alberto Vieira Cura. *ORGANIZAÇÃO JUDICIÁRIA (Elementos para umas Lições)*. Coimbra, 2010.

[6] Nitin Indurkhya e Fred J. Damerau. *Handbook of Natural Language Processing*. 2nd. Chapman & HallCRC, 2010. isbn: 1420085921.

[7] Alberto Maydeu-Olivares e Carlos Forero. "Goodness-of-Fit Testing". Em: vol. 7. Dez. de 2010, pp. 190–196. isbn: 9780080448947. doi: 10.1016/B978-0-08-044894-7.01333-6.

[8] Vipin Kumar Xindong Wu, ed. *The Top Ten Algorithms in Data Mining*. CRC Press, 2010. isbn: 978-1-4200-8964-6. url: https://doc.lagout.org/Others/Data\%20Mining/The\%20Top\%20Ten\%20Algorithms\%20in\%20Data\%20Mining\%20\%5BWu\%20\%26\%20Kumar\%202009-04-09\%5D.pdf.

[9] JOEL TIMÓTEO RAMOS PEREIRA. "ORGANIZA ORGANIZAÇÃO JUDICIÁRIA PORTUGUESA". Em: 6.º Congresso Internacional da Anamatra (Centro Cultural de Belém, Lisboa). Mar. de 2011. url: https://www.joelpereira.pt/direito/2011-03-17organizacaojudiciaria.pdf.

[10] John Gantz e David Reinse. *THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*. Rel. téc. EMC Corporation, 2012. url: https://www.speicherguide.de/download/dokus/IDC-Digital-Universe-Studie-iView-11.12.pdf.

[11] Jiawei Han, Micheline Kamber e Jian Pei. *Data Mining: Concepts and Techniques*. Third Edition. Elsevier, 2012.

[12] Bulent Dogru. "The Effect of Judicial Independence to FDI into Eastern Europe and South Asia". Em: 106 (jun. de 2013), pp. 1450–2887.

[13] Daniel Graupe. *Principles of Artificial Neural Networks*. 3nd. USA: World Scientific Publishing Co., Inc., 2013. isbn: 978-981-4522-73-1.

[14] Saif M. Mohammad e Peter D. Turney. "Crowdsourcing a Word-Emotion Association Lexicon". Em: 29.3 (2013), pp. 436–465.

[15] Assembleia da República. *Lei da Organização do Sistema Judiciário, Diário da República n.º 163/2013, Série I*. Ago. de 2013. url: https://data.dre.pt/eli/lei/62/2013/08/26/p/dre/pt/html.

[16] Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed Kamaleldin Mahmoud Ali, Muhammad Alam, Muhammad Shiraz e Abdullah Gani. "Big Data: Survey, Technologies, Opportunities, and Challenges". Em: *The Scientific World Journal* 2014 (2014), p. 18. doi: http://dx.doi.org/10.1155/2014/712826. url: http://downloads.hindawi.com/journals/tswj/2014/712826.pdf.

[17] Steve Lohr. "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights". Em: *The New York Times* (ago. de 2014). Online. Accessed: 2020-10-12. issn: 1084-6654. url: https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html.

[18] Marilene Lorizio e Antonia Rosa Gurrieri. "Efficiency of Justice and Economic Systems". Em: *Procedia Economics and Finance* 17 (2014). Innovation and Society - Statistical methods for the evaluation of services, pp. 104 –112. issn: 2212-5671. doi: https://doi.org/10.1016/S2212-5671(14)00884-3. url: http://www.sciencedirect.com/science/article/pii/S2212567114008843.

[19] Online Dispute Resolution Advisory Group. *ONLINE DISPUTE RESOLUTION FOR LOW VALUE CIVIL CLAIMS*. Rel. téc. UK, fev. de 2015. url: https://www.judiciary.uk/reviews/online-dispute-resolution/.

[20] A. Mishra e S. Vishwakarma. "Analysis of TF-IDF Model and its Variant for Document Retrieval". Em: *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*. 2015, pp. 772–776. doi: 10.1109/CICN.2015.157.

[21] Ryan Mitchell. *Web Scraping with Python: Collecting Data from the Modern Web*. 1st. O'Reilly Media, Inc., 2015. isbn: 1491910291.

[22] Francisco Herrera Salvador Garcia Julián Luengo. *Data Preprocessing in Data Mining*. Springer, Cham, 2015. isbn: 978-3-319-10246-7. doi: https://doi.org/10.1007/978-3-319-10247-4.

[23] Carlos Manuel Jorge da Silva Pereira Cedric Michael dos Santos. "Classificação de Documentos com Processamento de Linguagem Natural". Tese de mestrado. Instituto Superior de Engenharia de Coimbra, 2015.

[24] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro e Vasileios Lampos. "redicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective". Em: (2016). doi: 10.7717/peerj-cs.93. url: https://doi.org/10.7717/peerj-cs.93.

[25] Adrian Colyer. "The amazing power of word vectors". Em: *the morning paper* (abr. de 2016). Online. Accessed: 2020-10-23. url: https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/.

[26] Alek Gavrilovski, Hernando Jimenez, Dimitri Mavris, Arjun Rao, Karen Marais e Sanghyun Shin. "Gavrilovski et al SciTech2016 DataMining". Em: (fev. de 2016). doi: 10.2514/6.2016-0923.

[27] Vincenzo Bove e Elia Leandro. *The judicial system and economic development across EU Member States*. Rel. téc. Luxembourg: Publications Office of the European Union, 2017. url: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC104594/jrc104594_

_2017_the_judicial_system_and_economic_development_across_eu_member_states.pdf.

[28] Francois Chollet. *Deep Learning with Python*. Manning Publications, 2017. isbn: 9781617294433.

[29] Matthew James Denny e Arthur Spirling. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It". Em: *SSRN Electronic Journal* (2017). doi: 10.2139/ssrn.2849145. url: https://doi.org/10.2139/ssrn.2849145.

[30] Kristin Houser Dom Galeon. "An AI Completed 360,000 Hours of Finance Work in Just Seconds". Em: (mar. de 2017). Online. Accessed: 2020-11-16. url: https://futurism.com/an-ai-completed-360000-hours-of-finance-work-in-just-seconds.

[31] Daniel Martin Katz, Michael J. Bommarito II e Josh Blackman. "A general approach for predicting the behavior of the Supreme Court of the United States". Em: *PLOS ONE* 12.4 (abr. de 2017), pp. 1–18. doi: 10.1371/journal.pone.0174698. url: https://doi.org/10.1371/journal.pone.0174698.

[32] M. Eloi Buat-Ménard. "La justice dite« prédictive » en matière judiciaire : prérequis, risques et attentes – la réflexion en cours en France". Em: *L'INTELLIGENCE ARTIFICIELLE AU SERVICE DU POUVOIR JUDICIAIRE*. Riga, Letônia: CEPEJ - Administration judicaire de Lettonie, 2018. url: https://www.coe.int/en/web/cepej/practical-examples-of-ai-implemented-in-other-countries.

[33] Jyoti Dabass e Bhupender Singh Dabass. "Scope of Artificial Intelligence in Law". Em: (jun. de 2018). doi: 10.20944/PREPRINTS201806.0474.V1. url: https://www.preprints.org/manuscript/201806.0474/v1.

[34] Bryce J. Dietrich, Ryan D. Enos e Maya Sen. "Emotional Arousal Predicts Voting on the U.S.Supreme Court". Em: *Political Analysis* 27.2 (2018), pp. 237–243. doi: 10.1017/pan.2018.47.

[35] Lawrence Sherman Dr Geoffrey Barnes. "Police at the "front line" of difficult risk-based judgements are trialling an AI system trained to give guidance using the outcomes of five years of criminal histories." Em: *Research Horizons* (fev. de 2018). url: https://www.cam.ac.uk/system/files/issue_35_research_horizons_new.pdf.

[36] Palash Goyal, Sumit Pandey e Karan Jain. *Deep Learning for Natural Language Processing*. Apress, 2018, pp. 75–118. doi: 10.1007/978-1-4842-3685-7. url: https://doi.org/10.1007/978-1-4842-3685-7.

[37] European Commission for the Efficiency of Justice. *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment*. Dez. de 2018. url: https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c.

[38] Kankawin Kowsrihawat, Peerapon Vateekul e Prachya Boonkwan. "Predicting Judicial Decisions of Criminal Cases from Thai Supreme Court Using Bi-directional GRU with Attention Mechanism". Em: *2018 5th Asian Conference on Defense Technology (ACDT)*. IEEE, out. de 2018. doi: 10.1109/acdt.2018.8592948. url: https://doi.org/10.1109/acdt.2018.8592948.

[39] Mariana Oliveira. "Como a tecnologia está a revolucionar o mundo do Direito". Em: *Público* (nov. de 2018). Online. Accessed: 2020-11-20. url: https://www.publico.pt/2018/11/01/sociedade/noticia/justica-artificial-tecnologia-revolucionar-mundo-direito-1849618.

[40] Simon Stern. "Introduction: Artificial intelligence, technology, and the law". Em: *University of Toronto Law Journal* 68.supplement 1 (jan. de 2018), pp. 1–11. doi: 10.3138/utlj.2017-0102. url: https://doi.org/10.3138/utlj.2017-0102.

[41] Mireille Hildebrandt. "Data-Driven Prediction of Judgment. Law's New Mode of Existence?" Em: *SSRN Electronic Journal* (2019). doi: 10.2139/ssrn.3548504. url: https://doi.org/10.2139/ssrn.3548504.

[42] Direção de Serviços de Estatísticas da Justiça e Informática. *OS NÚMEROS DA JUSTIÇA*. Rel. téc. Direção-Geral da Política de Justiça, 2019. url: https://estatisticas.justica.gov.pt/sites/siej/pt-pt/Destaques/20191213_Os_numeros_da_justica_2018.pdf.

[43] Huma Jamshed, M. Sadiq Ali Khan, Muhammad Khurram, Syed Inayatullah e Sameen Athar. "Data Preprocessing: A preliminary step for web data mining." Em: *3C Tecnología. Glosas de innovación aplicadas a la pyme* (2019), pp. 206–221. doi: https://dx.doi.org/10.17993/3ctecno.2019.specialissue2. url: https://www.3ciencias.com/articulos/articulo/data-preprocessing-a-preliminary-step-for-web-data-mining/.

[44] Ministério da Justiça. *Relatório Justiça (2015/2019)*. Rel. téc. Ministério da Justiça, 2019. url: https://www.portugal.gov.pt/download-ficheiros/ficheiro.aspx?v=2679055e-f6d0-411a-8b45-7c21bdc8173e.

[45] Akshay Kulkarni e Adarsha Shivananda. *Natural Language Processing Recipes*. Apress, 2019. doi: 10.1007/978-1-4842-4267-4.

[46] David Mandim. "A Inteligência Artificial ao serviço da Justiça. Pode haver um juiz-robô?" Em: *Diário de Notícias* (out. de 2019). Online. Accessed: 2020-11-20. url: https://www.dn.pt/pais/a-inteligencia-artificial-ao-servico-da-justica-pode-haver-um-juiz-robo--11408704.html.

[47] Elisa Alfaia Sampaio, João J.Seixas e Paulo Jorge Gomes. *Artificial Intelligence and the Judicial Ruling*. Jul. de 2019. url: http://www.ejtn.eu/PageFiles/17916/TEAM%20PORTUGAL%20I%20TH%202019%20D.pdf.

[48] Luís Barreto Xavier. "A Inteligência Artificial é inimiga dos juízes?" Em: *Jornal Económico* (jul. de 2019). Online. Accessed: 2020-11-26. url: https://jornaleconomico.sapo.pt/noticias/a-inteligencia-artificial-e-inimiga-dos-juizes-463242.

[49] European Commission. *The 2020 EU justice scoreboard*. Rel. téc. European Union, 2020. doi: 10.2838/71558. url: https://ec.europa.eu/info/sites/info/files/justice_scoreboard_2020_en.pdf.

[50] Instituto de Gestão Financeira e Equipamentos da Justiça. "SOBRE O IGFEJ "Quem somos"". Em: (2020). Online. Accessed: 2020-11-05. url: https://igfej.justica.gov.pt/Sobre-o-IGFEJ/Quem-somos.

[51] Portal Europeu da Justiça. "Identificador europeu da jurisprudência (ECLI)". Em: (out. de 2020). Online. Accessed: 2020-01-11. url: https://e-justice.europa.eu/content_european_case_law_identifier_ecli-175-pt.do.

[52] Benjamin Strickson e Beatriz De La Iglesia. "Legal Judgement Prediction for UK Courts". Em: *Proceedings of the 2020 The 3rd International Conference on Information Science and System*. ACM, mar. de 2020. doi: 10.1145/3388176.3388183. url: https://doi.org/10.1145/3388176.3388183.

[53] CONSELHO SUPERIOR DA MAGISTRATURA. "ECLI · European Case Law Identifier". Em: (2021). Online. Accessed: 2021-01-11. url: https://jurisprudencia.csm.org.pt/ecli.

[54] *Online Dispute Resolution*. https://ec.europa.eu/consumers/odr/main/?event=main.trader.register. Online. Accessed: 2020-11-15.

Apêndice A

Estrutura do texto integral da decisão

Figura A.1: Exemplo do texto integral da decisão de um acórdão do Tribunal da Relação de Coimbra, com o destaque das diferentes partes: relatório, fundamentação e decisão.

Apêndice B

Preparação dos dados

Listagem B.1: Expressão regular usada para filtrar as citaçaões de artigos legislativos.

```
regex = r"((artº|artigo|arts|art\.|art°).{1,50}?)(d[aeo]s?)(\ +((\b[A-Z][\.a-zçõóéèíãêáàõ]+
[nº°\"\. ]*?[0-9/-A-Z]+)|(\b[A-Z][A-Z]+ [nº°\"\. ]*?[0-9/-A-Z]+)|(\b[A-Z][A-Z]+)|(\b[A-Z]
[\.a-zçõóéèíãêáàõ\-\,]+([nº°\"\. ]*?[0-9/-A-Z]+)?)|(\bd[aeo]s? [A-Z][a-zçõóéêèíãõáà\. ]
+)\b)+)+"
```

Listagem B.2: Lista de palavras irrelevantes que foram removidas do texto.

```
stopwords = ['de', 'a', 'o', 'que', 'e', 'é', 'do', 'da', 'em', 'um', 'para', 'com', 'não',
'uma', 'os', 'no', 'se', 'na', 'por', 'mais', 'as', 'dos', 'como','mas', 'ao',
'ele', 'das', 'à', 'seu', 'sua', 'ou', 'quando', 'muito', 'nos', 'já', 'eu', 'também',
'só', 'pelo', 'pela', 'até', 'isso', 'ela', 'sem', 'mesmo', 'aos', 'seus', 'quem',
'nas', 'me', 'esse', 'eles','você', 'essa', 'num', 'nem', 'suas', 'meu', 'às',
'minha', 'numa', 'pelos', 'elas', 'qual', 'nós', 'lhe', 'deles', 'essas', 'esses',
'pelas', 'este', 'dele', 'tu', 'te', 'vocês', 'vos', 'lhes', 'meus', 'minhas',
'teu', 'tua', 'teus', 'tuas', 'nosso', 'nossa', 'nossos', 'nossas', 'dela',
'delas', 'esta', 'estes', 'estas', 'aquele', 'aquela', 'aqueles', 'aquelas', 'isto',
'aquilo','estou', 'está', 'estamos', 'estão', 'estive', 'esteve', 'estivemos', 'estiveram',
'estava', 'estávamos', 'estavam', 'estivera', 'estivéramos','esteja',
'estiver', 'estivermos', 'estiverem','hei', 'há', 'havemos', 'hão', 'houve',
'houvemos', 'houveram', 'houvera', 'houvéramos','haja', 'hajamos', 'hajam',
'houvesse', 'houvéssemos', 'houvessem', 'houver','houvermos', 'houverem', 'houverei',
'houverá', 'houveremos', 'houverão', 'houveria', 'houveríamos','houveriam',
'sou', 'somos', 'são', 'era', 'éramos', 'eram', 'fui', 'foi', 'fomos', 'foram',
'fora', 'fôramos', 'seja','sejamos', 'sejam', 'fosse', 'fôssemos', 'fossem',
'for', 'formos', 'forem', 'serei', 'será', 'seremos', 'serão', 'seria', 'seríamos',
'seriam', 'tenho', 'tem', 'temos', 'tém', 'tinha', 'tínhamos', 'tinham', 'tive', 'teve',
'tivemos', 'tiveram', 'tivera', 'tivéramos', 'tenha','tenhamos', 'tenham',
'tivesse', 'tivéssemos', 'tivessem', 'tiver','depois', 'entre', 'terei',
'artigo', 'dl', 'terá', 'teremos', 'terão', 'teria', 'teríamos', 'teriam' 'ii',
'iii', 'iv', 'ix', 'nº', 'artº', 'art', 'aa', 'rr', 'ré']
```

Tabela B.1: Lista de legislação portuguesa inserida na base de dados.

Lista de legislação		
1. Código Civil	2. Lei do Contrato de Seguro	3. Código da Estrada
4. Código Penal	5. Regime Jurídico do Processo de Inventário	6. Código Cooperativo
7. Código de Processo Civil	8. Código do Procedimento Administrativo	9. Lei Tutelar Educativa
10. Código de Processo Penal	11. Código da Publicidade	12. Estatuto dos Benefícios Fiscais
13. Constituição da República Portuguesa	14. Regime Jurídico de Empreitadas de Obras Públicas	15. Regime do IVA nas Transações Intracomunitárias
16. Convenção de Bruxelas	17. Código do Registo Predial	18. Regime Geral das Infrações Tributárias
19. Convenção relativa ao contrato de transporte internacional de mercadorias por estrada	20. Lei de Organização e Funcionamento dos Tribunais Judiciais	21. Código do Imposto Municipal sobre as Transmissões Onerosas de Imóveis
22. Tratado sobre o Funcionamento da União Europeia	23. Código de Justiça Militar	24. Regime Geral das Contra-Ordenações
25. Tratado de Roma	26. Código do Notariado	27. Supremo Tribunal de Justiça
28. Código do Trabalho	29. Lei da Organização do Sistema Judiciário	30. Estatuto dos Tribunais Administrativos e Fiscais
31. Código de Processo do Trabalho	32. Código do Direito de Autor e dos Direitos Conexos	33. Código Fiscal do Investimento
34. Código das Sociedades Comerciais	35. Código da Insolvência e da Recuperação de Empresas	36. Regime Jurídico da Arbitragem em Matéria Tributária
37. Código dos Valores Mobiliários	38. Estatuto do Administrador Judicial	39. Organização Tutelar de Menores
40. Regulamento de Controlo Metrológico dos Alcoolímetros	41. Código do Imposto sobre o Rendimento das Pessoas Singulares	42. Contribuição sobre o setor bancário
43. Código do Registo Civil	44. Código das Custas Judiciais	45. Lei Geral Tributária
46. Código das Expropriações	47. Regulamento das Custas Processuais	48. Convenção sobre os Direitos da Criança
49. Código das Expropriações	50. Código Comercial	51. Declaração Universal dos Direitos Humanos
52. Código da Propriedade Industrial	53. Código do Imposto sobre o Valor Acrescentado	54. Regime de comunicação de informações financeiras
55. Regime Geral do Processo Tutelar Cível	56. Código dos Impostos Especiais de Consumo	57. Regime Jurídico do Contrato Individual de Trabalho
58. Regulamento de Fiscalização da Condução sob Influência do Álcool ou de Substâncias Psicotrópicas	59. Regime Complementar do Procedimento de Inspeção Tributária e Aduaneira	60. Estatuto da Ordem dos Solicitadores e dos Agentes de Execução

61. Código dos Processos Especiais de Recuperação da Empresa e de Falência	62. Código dos Regimes Contributivos do Sistema Previdencial de Segurança Social	63. Regime que cria a contribuição extraordinária sobre o setor energético
64. Base Instrutória	65. Código do Imposto do Selo	66. Lei do Apoio Judiciário
67. Convenção Europeia dos Direitos do Homem	68. Código do Imposto Municipal sobre Imóveis	69. Regime Jurídico das Infrações Fiscais Não Aduaneiras
70. Código de Procedimento e Processo Tributário	71. Código do Imposto sobre Veículos	72. Lei de proteção de crianças e jovens em perigo
73. Carta dos Direitos Fundamentais da União Europeia	74. Código do Imposto Único de Circulação	75. Regime Geral das Instituições de Crédito e Sociedades Financeiras
76. Regulamento Geral das Capitanias	77. Código de Processo nos Tribunais Administrativos	78. Lei de Cessação do Contrato de Trabalho
79. Código da Execução das Penas e Medidas Privativas da Liberdade	80. Código das Associações Mutualistas	81. Regime Jurídico das Armas e Munições
82. Código de Execução de Penas	83. Estatuto da Ordem dos Advogados	84. Regras e Usos Uniformes
85. Código do Imposto sobre o Rendimento das Pessoas Coletivas	86. Lei das Cláusulas Contratuais Gerais	87. Código Deontológico
88. Novo Regime do Arrendamento Urbano	89. Lei da Arbitragem Voluntária	90. Serviço Nacional de Saúde
91. Regime Jurídico da Urbanização e Edificação	92. Lei Uniforme Relativa às Letras e Livranças	93. Pacto Internacional sobre os Direitos Civis e Políticos
94. Regulamento Geral das Edificações Urbanas	95. Lei de Acidentes de Trabalho	96. Convenção coletiva de trabalho
97. Registo Nacional de Pessoas Coletivas	98. Lei Uniforme Relativa ao Cheque	99. Tratado sobre o Funcionamento da União Europeia

Apêndice C

Relação entre os *clusters* de legislação e a decisão

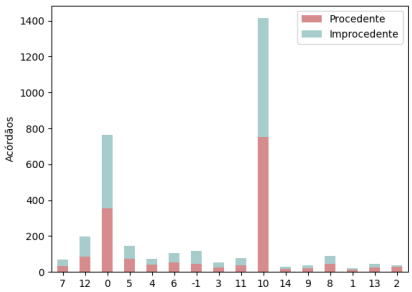


Figura C.1: Distribuição dos acórdãos pelos diferentes *clusters* de legislação e pelas duas categorias de decisão no tópico 0.

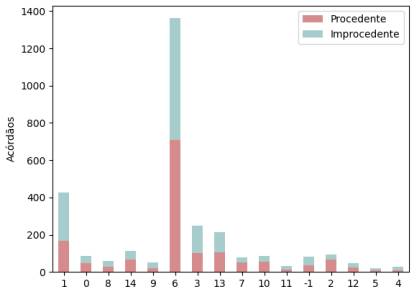


Figura C.2: Distribuição dos acórdãos pelos diferentes *clusters* de legislação e pelas duas categorias de decisão no tópico 2.

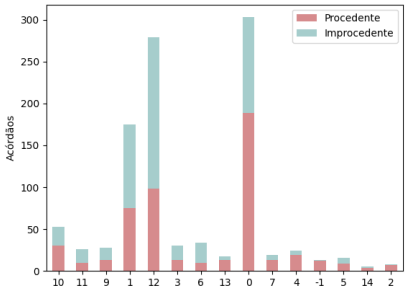


Figura C.3: Distribuição dos acórdãos pelos diferentes *clusters* de legislação e pelas duas categorias de decisão no tópico 3.

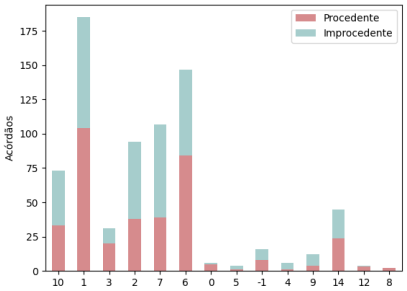


Figura C.4: Distribuição dos acórdãos pelos diferentes *clusters* de legislação e pelas duas categorias de decisão no tópico 4.

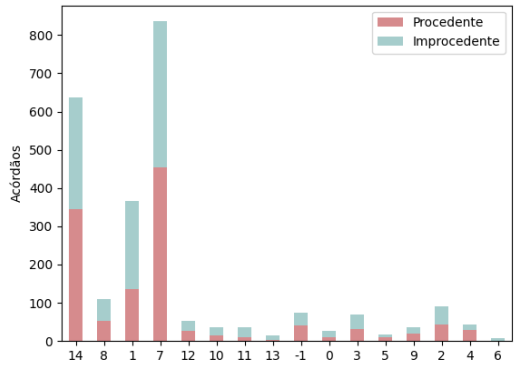


Figura C.5: Distribuição dos acórdãos pelos diferentes *clusters* de legislação e pelas duas categorias de decisão no tópico 5.

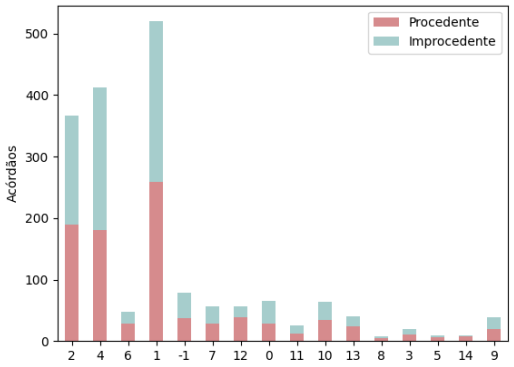


Figura C.6: Distribuição dos acórdãos pelos diferentes *clusters* de legislação e pelas duas categorias de decisão no tópico 6.

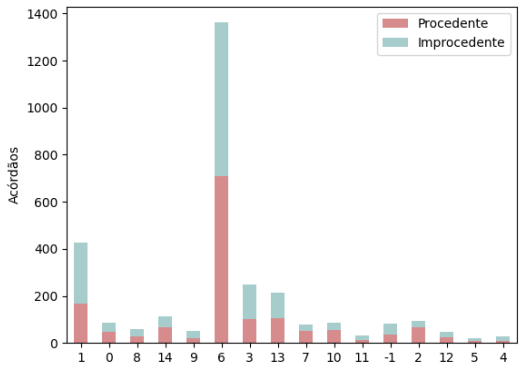


Figura C.7: Distribuição dos acórdãos pelos diferentes *clusters* de legislação e pelas duas categorias de decisão no tópico 7.

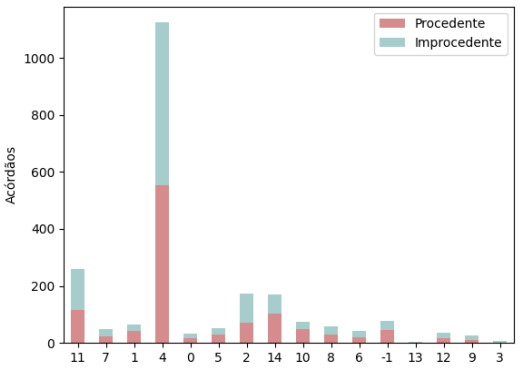


Figura C.8: Distribuição dos acórdãos pelos diferentes *clusters* de legislação e pelas duas categorias de decisão no tópico 8.

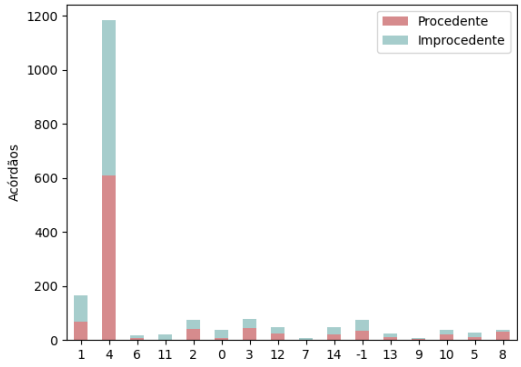


Figura C.9: Distribuição dos acórdãos pelos diferentes *clusters* de legislação e pelas duas categorias de decisão no tópico 11.

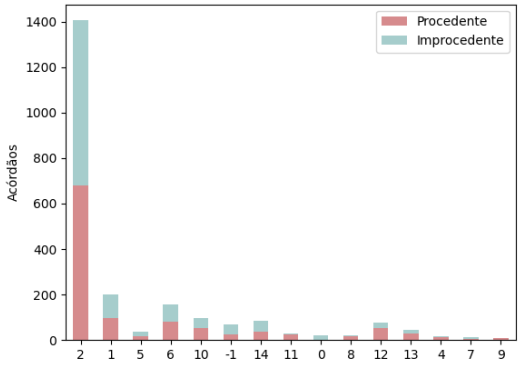


Figura C.10: Distribuição dos acórdãos pelos diferentes *clusters* de legislação e pelas duas categorias de decisão no tópico 12.

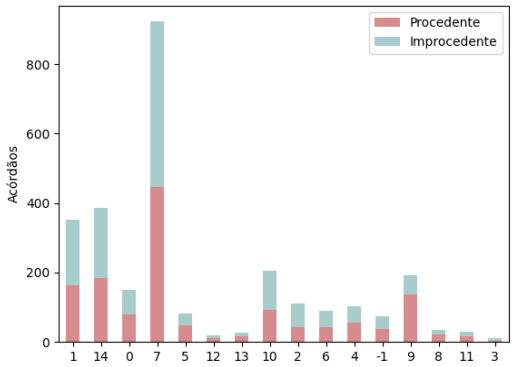


Figura C.11: Distribuição dos acórdãos pelos diferentes *clusters* de legislação e pelas duas categorias de decisão no tópico 13.

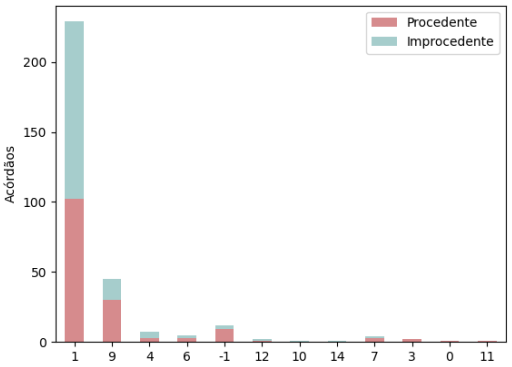


Figura C.12: Distribuição dos acórdãos pelos diferentes *clusters* de legislação e pelas duas categorias de decisão no tópico 14.

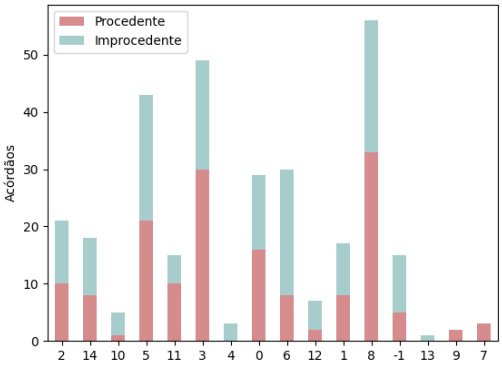


Figura C.13: Distribuição dos acórdãos pelos diferentes *clusters* de legislação e pelas duas categorias de decisão no tópico 15.

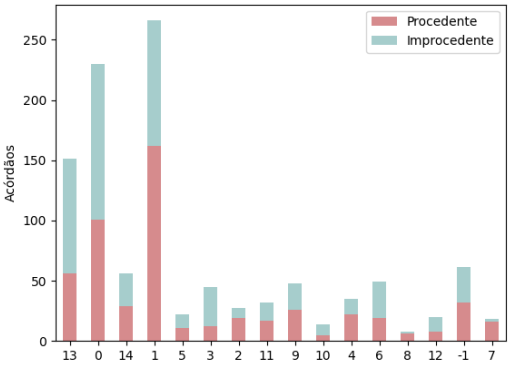


Figura C.14: Distribuição dos acórdãos pelos diferentes *clusters* de legislação e pelas duas categorias de decisão no tópico 16.

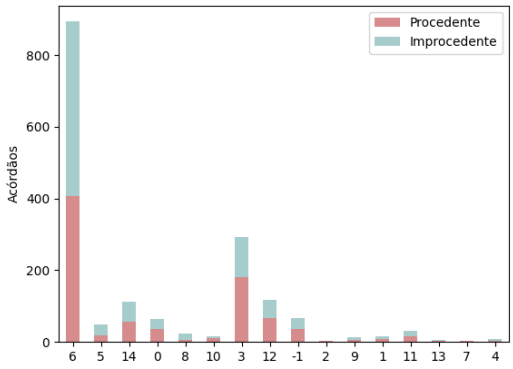


Figura C.15: Distribuição dos acórdãos pelos diferentes *clusters* de legislação e pelas duas categorias de decisão no tópico 17.

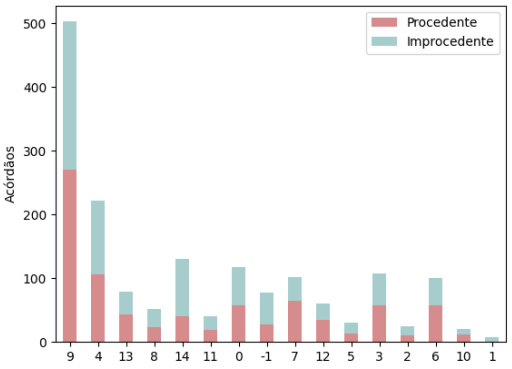


Figura C.16: Distribuição dos acórdãos pelos diferentes *clusters* de legislação e pelas duas categorias de decisão no tópico 19.

Apêndice D

Dashboard

A figura D.1 apresenta um dos gráficos apresentado na página inicial da aplicação. Contém informação do número de processos entrados, findos e pendentes nos tribunais da relação portugueses. De referir que estes são dados provenientes da fonte externa *PORDATA*.

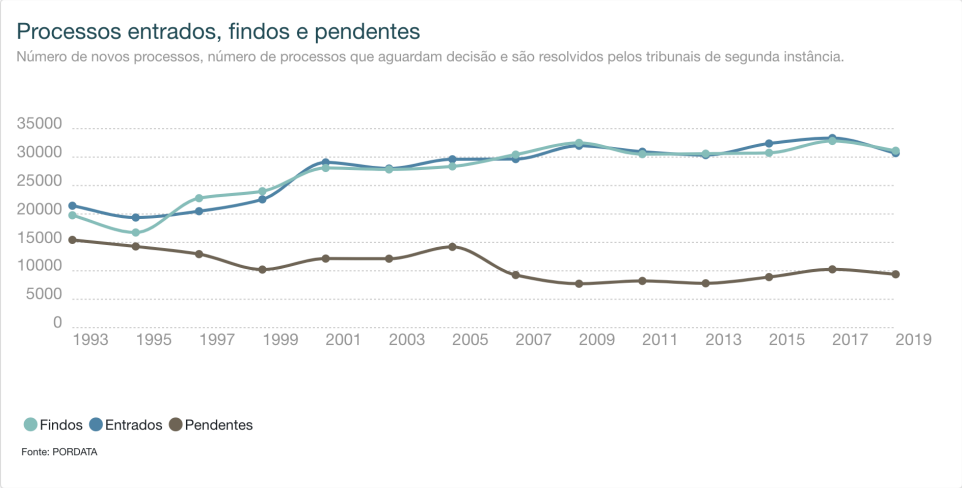


Figura D.1: Segundo excerto da página inicial da aplicação.

A figura D.1 apresenta os gráficos finais apresentados na página inicial da aplicação. O primeiro contém informação do número total de cada tipo de decisão tomada nos diferentes tribunais da relação, enquanto que o segundo é mais genérico pois apresenta a percentagem de cada tipo de decisão presente na base de dados.

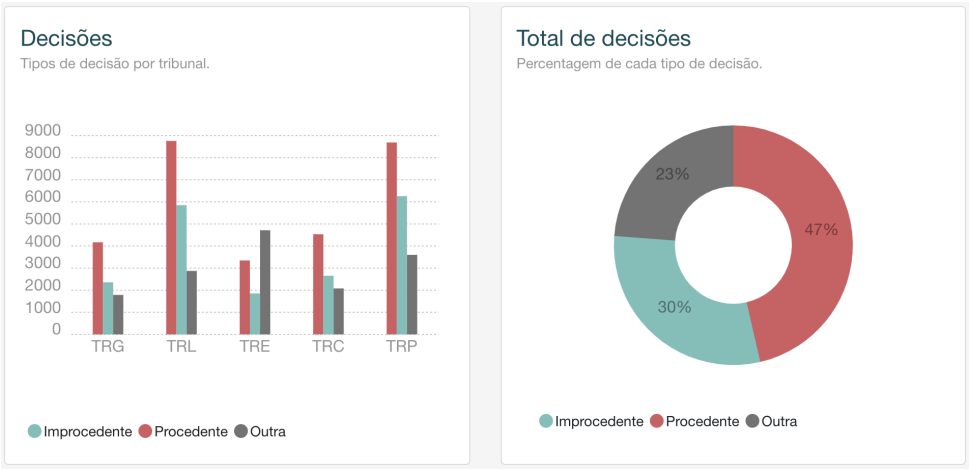


Figura D.2: Terceiro excerto da página inicial da aplicação.

A figura D.3 apresenta dois gráficos que resultam de uma pesquisa avançada na aplicação. Enquanto que o primeiro corresponde ao número total de cada tipo de decisão nos acórdãos filtrados, o segundo apresenta os 10 descritores mais frequentes e o número de acórdãos em que aparecem.

Figura D.3: Segundo excerto da página correspondente à funcionalidade de pesquisa avançada.