Alina Yanchuk

Classificação automática de artigos estigmatizantes de doenças mentais em jornais de notícias portugueses *online*

Automatic classification of stigmatizing articles of mental illness in Portuguese online newspapers

Alina Yanchuk

Classificação automática de artigos estigmatizantes de doenças mentais em jornais de notícias portugueses *online*

Automatic classification of stigmatizing articles of mental illness in Portuguese online newspapers

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Informática, realizada sob a orientação científica da Doutora Alina Liliana Trifan, Investigadora do Instituto de Engenharia Eletrónica e Informática de Aveiro, e do Doutor José Luís Guimarães Oliveira, Professor Catedrático do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro.

o júri / the jury

presidente / president

Doutor António Joaquim da Silva Teixeira
Professor Associado C/ Agregação da Universidade de Aveiro

vogais / examiners committee

Doutora Cátia Luísa Santana Calisto Pesquita
Professora Auxiliar da Faculdade de Ciências da Universidade de Lisboa

Doutor José Luis Guimarães Oliveira
Professor Catedrático da Universidade de Aveiro

agradecimentos

Agradeço aos meus pais, que me apoiaram e fizeram com que este percurso académico tenha sido possível.

Aos meus amigos e colegas universitários, que percorreram comigo os momentos mais difíceis e, ao mesmo tempo, permitiram criar memórias inesquecíveis da cidade de Aveiro.

Aos meus orientadores, que se mostraram sempre presentes e ajudaram-me em todas as etapas.

E a todos aqueles que, de alguma forma, participaram na realização deste trabalho.

Palavras-chave

Classificação de texto, Classificação automática, Inteligência Artificial, Jornais de notícias, Processamento de Linguagem Natural, *Machine learning*, *Deep learning*, *Topic modeling*

Resumo

Os meios de comunicação social, nomeadamente os jornais de notícias presentes na Internet, são os principais responsáveis pelo fornecimento de informação ao público e possuem uma grande influência na modelação da nossa sociedade. A presença de estigma associado à saúde mental continua a ser frequente nos artigos publicados nos mesmos, onde, muitas vezes, as doenças mentais são utilizadas de forma metafórica para se referir a entidades ou situações fora do contexto clínico da saúde mental.

Tendo em conta que a análise manual deste problema requer um grande esforço humano e tempo, este projeto explora a implementação de técnicas de Inteligência Artificial e de Processamento de Linguagem Natural para a tarefa de classificação automática de artigos estigmatizantes dos transtornos mentais da esquizofrenia e psicose, presentes em jornais de notícias portugueses *online* e recolhidos do repositório público Arquivo.pt. Foram implementados dez algoritmos de *machine learning* e *deep learning* para a realização desta tarefa, sendo que 45% dos modelos permitiram obter resultados com exatidão acima dos 90%. Além disso, foi também realizada a deteção automática de tópicos presentes nos artigos, através de *topic modeling* com o modelo *top2vec*, que permitiu concluir que a estigmatização da saúde mental ocorre, essencialmente, nas temáticas da Economia e Política. Os resultados experimentais confirmam a existência de estigma nos jornais de notícias portugueses (52% dos 978 artigos recolhidos) e a eficácia da utilização de modelos computacionais para a sua deteção. Adicionalmente, é criado e disponibilizado um conjunto de 978 artigos recolhidos e manualmente anotados com as classes "estigmatizante" e "literal".

Keywords Text classification, Automatic classification, Artificial Intelligence, Newspapers, Natural Language Processing, Machine learning, Deep learning, Topic modeling

Abstract The media, namely the written newspapers available on the Internet are primarily responsible for providing information to the public and have a great influence on shaping our society. The presence of stigma related to mental health remains frequent in the articles published online, where mental diseases are often used metaphorically to refer to entities or situations outside the clinical context of mental health.

Considering that the manual analysis of this problem requires a great deal of human effort and time, this project explores the implementation of Artificial Intelligence and Natural Language Processing techniques for the task of automatically classifying stigmatizing articles on the mental disorders of schizophrenia and psychosis, present in Portuguese online newspapers and collected from the public repository Arquivo.pt. Ten machine learning and deep learning algorithms were implemented to perform this task, and 45% of the models led to results with accuracy above 90%. In addition, the automatic detection of the articles topics was also performed, through topic modeling with the *top2vec* model, which allowed to conclude that the stigmatization of mental health occurs, essentially, in the topics of Economics and Politics.

The experimental results confirm the existence of stigma in Portuguese online newspapers (52% of the 978 articles collected) and the effectiveness of the use of Artificial Intelligence to detect it. Additionally, a set of 978 articles collected and manually annotated with the classes "stigmatizing" and "literal" is created and made available.

Índice

Lista de figuras

Lista de tabelas

Lista de abreviações

API . . . . . *Application Programming Interface*

BERT . . . . . *Bidirectional Encoder Representations from Transformers*
Bi-LSTM . . . . . *Bidirectional Long Short Term Memory*

CNN . . . . . *Convolutional Neural Network*
CSV . . . . . *Comma-separated values*

HTML . . . . . *HyperText Markup Language*
HTTP . . . . . *Hypertext Transfer Protocol*

IA . . . . . Inteligência Artificial

JSON . . . . . *JavaScript Object Notation*

KNN . . . . . *K-Nearest Neighbors*

LDA . . . . . *Linear Discriminant Analysis*
LIWC . . . . . *Linguistic Inquiry Word Count*
LSA . . . . . *Latent Semantic Analysis*
LSTM . . . . . *Long Short Term Memory*

NER . . . . . *Named-entity recognition*
NLTK . . . . . *Natural Language Toolkit*
NMF . . . . . *Non-negative Matrix Factorization*

PCA . . . . . *Principal Component Analysis*
PLN . . . . . Processamento de Linguagem Natural
PLSA . . . . . *Probabilistic Latent Semantic Analysis*
POS . . . . . *Part-of-Speech*

SVD . . . . . *Singular Value Decomposition*
SVM . . . . . *Support Vector Machine*

TF-IDF . . . . . *Term Frequency - Inverse Document Frequency*

URL . . . . . *Uniform Resource Locator*

WWW . . . . . *World Wide Web*

XML . . . . . *Extensible Markup Language*

Capítulo 1

Introdução

Neste capítulo, é introduzido o tema da presente dissertação, sendo o mesmo a classificação automática de artigos publicados em jornais de notícias portugueses *online*, e arquivados no repositório público Arquivo.pt, que utilizam expressões estigmatizantes de doenças mentais. São também explorados os objetivos do desenvolvimento do projeto e discutida a contribuição do mesmo para o tema proposto. No final, é apresentada a estrutura que este documento segue, bem como os principais tópicos discutidos em cada secção.

1.1 Contexto e Motivação

A presença de estigma na nossa sociedade é uma realidade bastante frequente. Quando o mesmo é associado às doenças mentais, tem implicações negativas nos doentes, nos seus tratamentos e nos próprios profissionais de saúde. A estigmatização ocorre, geralmente, a dois níveis. O primeiro é caracterizado pela utilização de termos médicos, num sentido figurado ou metafórico, para descrever entidades ou situações fora do contexto clínico da saúde mental, o que contribui para os maiores níveis de negatividade e desvalorização da doença. O segundo é caracterizado pela utilização dos termos para se referir aos seus portadores em situações onde este cenário é desnecessário, o que pode criar desconforto ao limitá-los apenas à sua doença. Neste âmbito, surge a necessidade de combater o estigma presente na comunicação social, nomeadamente nos jornais de notícias, onde a utilização de expressões estigmatizantes é ainda bastante comum, tanto por parte dos próprios autores como dos indivíduos que os mesmos entrevistam ou citam.

Por outro lado, a análise de notícias jornalísticas tem apresentado um grande crescimento na área da investigação. De acordo com o portal *Scopus*[1], a maior base de dados *online* de literatura revista por pares, o número de publicações relativas à temática de classificação automática de notícias apresenta um grande aumento nos últimos anos, e principalmente a partir do ano de 2015 (Figura 1.1). Além disso, cada vez mais têm sido adotadas abordagens computacionais para a realizar, em contraste com os tradicio-

[1] https://www.scopus.com/

nais métodos manuais. Os métodos manuais caracterizam-se pela anotação, por humanos, dos textos a classificar, enquanto que os métodos computacionais utilizam Inteligência Artificial (IA). Os subcampos da IA mais relevantes para este processo são a aprendizagem automática ou *machine learning*, o Processamento de Linguagem Natural (PLN) e a Mineração de Texto. Apesar de as duas metodologias apresentarem diferenças na sua implementação e precisão de resultados, partilham ambas o mesmo objetivo que é a sensibilização do público para questões importantes que devem ser tratadas na área da comunicação social.

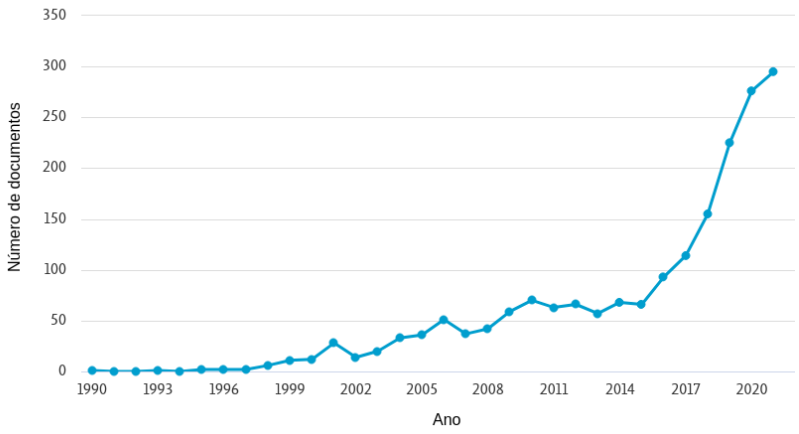


Figura 1.1: Comparação do número de publicações por ano com os termos "*news*", "*classification*" e "*automatic*", presentes no título, no resumo ou nas palavras-chave. Adaptado do *Scopus*.

Por fim, tendo em conta a necessidade de aceder a um grande volume de artigos de jornais de notícias digitais, recentes e mais antigos, surge em auxílio o repositório público Arquivo.pt[2], um repositório *online* de páginas web portuguesas arquivadas. Esta plataforma é também responsável pela realização do concurso "Prémio Arquivo.pt", que tem sido lançado anualmente desde 2018 com o objetivo de promover a utilidade do repositório e das suas funcionalidades [1], podendo o projeto realizado ser submetido ao concurso.

Assim, é de grande interesse explorar os níveis de estigmatização na comunicação social portuguesa, e realizá-lo recorrendo às funcionalidades do repositório Arquivo.pt e a processos computacionais, que cada vez mais têm ganho destaque no nosso quotidiano e apresentam grandes vantagens, nomeadamente a eficiência.

1.2 Objetivos

O principal objetivo deste projeto é a exploração de técnicas de *machine learning*, de PLN e de mineração de texto para a tarefa de classificação automática de artigos

[2] https://arquivo.pt/

estigmatizantes de doenças mentais, presentes em jornais de notícias *online* e recolhidos do repositório Arquivo.pt. É explorada a classificação de artigos que utilizam as doenças mentais num sentido metafórico, para expressar ideias que vão além do significado literal e básico dos termos. Este processo pode, assim, ser dividido em cinco etapas principais:

- Estudo de métodos e tecnologias para o processo de recolha de informação e para o processo de classificação automática de dados textuais;
- Exploração do repositório Arquivo.pt e utilização das suas funcionalidades para a recolha de páginas web, e consequente extração de dados relevantes das mesmas;
- Aplicação de técnicas de pré-processamentos dos dados, com recurso a *frameworks* e bibliotecas existentes para o efeito;
- Desenvolvimento e implementação de modelos computacionais para a classificação automática de artigos estigmatizantes, e sua avaliação e comparação;
- Deteção automática de tópicos presentes nos artigos, através de *topic modeling*;
- Análise e apresentação dos resultados obtidos, bem como submissão para o concurso "Prémio Arquivo.pt 2022".

1.3 Contribuições

O estigma associado às doenças mentais continua a estar presente na imprensa portuguesa, um dos meios que mais contribui para o fornecimento de informação e formação de opinião. Assim, é importante continuar a explorar este problema e a proporcionar possíveis ferramentas para o combater. Tendo em conta que uma grande parte das tarefas de classificação de texto é realizada manualmente, este projeto apresenta um nível inovador ao explorar a estigmatização de doenças mentais na comunicação social portuguesa através de IA. É também de referir que os dados obtidos e os modelos desenvolvidos contribuem para a formação de conhecimento na área de PLN em português, e podem ser estendidos e ajustados a outros problemas similares. Por fim, o projeto utiliza os dados e funcionalidades do repositório Arquivo.pt, demonstrando a utilidade e benefícios do mesmo, e servindo como exemplo para futuros projetos neste âmbito.

1.4 Estrutura do Documento

Este documento é composto por sete capítulos que dizem respeito às diferentes etapas necessárias para a elaboração do projeto. No Capítulo 1, da Introdução, é efetuada uma explicação introdutória do projeto, contextualizando-se a utilidade da mesmo para o tema proposto e especificando-se os seus objetivos. No Capítulo 2, do Estado da Arte, é apresentada uma reflexão sobre o atual conhecimento científico referente aos principais tópicos

do projeto e é realizado um levantamento de projetos similares. No Capítulo 3, da Recolha dos dados, é descrita a metodologia adotada para a recolha dos dados, nomeadamente dos artigos de jornais de notícias *online*, do repositório Arquivo.pt. No Capítulo 4, da Anotação manual dos artigos, é descrito o modo como os artigos recolhidos foram manualmente anotados com os sentidos "estigmatizante" e "literal". No Capítulo 5, da Classificação automática de texto, são retratadas as principais etapas da implementação de modelos de IA para a classificação do sentido dos artigos e para a deteção de tópicos presentes nos mesmos, bem como o desenvolvimento do *website* para a apresentação dos resultados obtidos das mesmas. No Capítulo 6, dos Resultados e Avaliação, são apresentados e discutidos os resultados obtidos nas diferentes fases de implementação e divulgação do projeto. Por fim, o documento termina com o Capítulo 7, da Conclusão, onde são apresentadas as principais conclusões retiradas.

Capítulo 2

Estado da Arte

Neste capítulo, são apresentados os três tópicos mais importantes para o desenvolvimento desta dissertação, bem como o atual conhecimento científico que representa cada um. O primeiro tópico aborda as doenças mentais e a presença de estigma a elas associado. São apresentadas algumas estatísticas sobre a saúde mental em Portugal e as estratégias que têm sido adotadas para combater o estigma presente na população portuguesa. É também discutida a perceção pública dos portadores de perturbações mentais e a influência da comunicação social na mesma. O segundo tópico apresenta o repositório Arquivo.pt, a fonte de dados para a elaboração do projeto desta dissertação, bem como os métodos de recolha de dados que podem ser adotados. O último tópico foca-se na classificação automática de artigos estigmatizantes. São apresentados os métodos que são estado da arte nos campos de *machine learning*, de PLN e de mineração de texto, bem como as novas abordagens que têm surgido recentemente em cada área. Além disso, são apresentadas as ferramentas que poderão ser utilizadas para auxiliar na implementação destes métodos. De forma complementar, ao longo do capítulo é também analisado o nível inovador deste projeto tendo em conta trabalhos similares já realizados.

O capítulo termina com um sumário dos principais aspetos referentes aos tópicos apresentados e com as conclusões retiradas neste âmbito.

2.1 Doenças mentais e estigma

As doenças mentais são condições de saúde diagnosticadas que podem envolver alterações de pensamento, de emoções e de comportamento [2]. Os casos mais leves têm uma interferência menor no quotidiano dos doentes e apresentam uma maior frequência na população, sendo as mais comuns as perturbações de depressão mais leves e as perturbações de ansiedade [3]. Por outro lado, as doenças mais graves, que afetam significativamente a vida dos doentes e podem levá-los à necessidade de cuidados hospitalares, constituem, essencialmente, as perturbações depressivas graves, a esquizofrenia e o transtorno bipolar [2].

Para além dos desafios físicos e psicológicos que estas doenças trazem, os seus por-

tadores sofrem também com o estigma a elas associado. A palavra "estigma", segundo os dicionários portugueses [4], é frequentemente utilizada no seu sentido figurado e tem como significado a "perceção negativa associada a certo comportamento, característica, grupo, etc". Por sua vez, a Organização Mundial da Saúde define o estigma como uma marca distintiva que, ao aliar-se às perturbações mentais, cria um ambiente de exclusão social e discriminação perante a pessoa estigmatizada [5]. É um conceito muito associado a estereótipos negativos e, na maioria das vezes, forma-se com base em informações falsas e sem qualquer fundamento científico.

Em resultado disso, pessoas portadoras de alguma doença mental não só vivem rodeadas de comentários e comportamentos violentos e desrespeitosos acerca da sua doença e pessoa, como também são prejudicadas ao nível da sua qualidade de vida. Frequentemente, deparam-se com situações em que lhes são negados empregos, serviços, integração social e oportunidades [6].

2.1.1 A situação em Portugal

Um documento da Ordem dos Psicólogos Portugueses, lançado em 2021, revela que as doenças mentais afetam um em cada cinco portugueses (23%), sendo a pandemia da COVID-19 um fator que tem contribuído para o aumento deste número [7]. Em Portugal, tal como em outros países, a realidade do estigma existe e grande parte da sociedade ainda tende a estigmatizar comportamentos que não entende e que, do seu ponto de vista, diferem do senso comum. Quanto aos meios de comunicação social informativos, que são os principais transmissores de informação ao público, estes raramente produzem conteúdos sobre doenças mentais. Quando o fazem, tendem a dar destaque àqueles que terão mais visibilidade e impacto na população, apresentando muito deles de forma exagerada e negativa [8, 9].

Para combater esta situação, ao longo dos últimos anos tem sido aplicado algum esforço tanto ao nível da legislação como por parte de organizações. Um exemplo disso é o Plano Nacional de Saúde Mental 2007-2016 [10], que foi aprovado em 24 de janeiro de 2008 e tem vindo a promover e a avaliar projetos de combate ao estigma presente na população portuguesa. Um projeto neste âmbito é o INFORMEMENTE [1], lançado pela Sociedade Portuguesa de Psiquiatria e Saúde Mental, que apresenta um manual prático designado por "Guia essencial para jornalistas sobre saúde mental" [11], e cujo objetivo é combater o estigma existente nos meios de comunicação social.

2.1.2 Perceção pública das doenças mentais

Ao longo dos últimos anos, foram efetuados vários estudos sobre atitudes estigmatizantes das pessoas perante portadores de doenças mentais. Como resultados, grande parte relata a perceção negativa dos doentes, o medo dos seus comportamentos imprevisíveis, o desconforto na sua presença e a dificuldade na interação com os mesmos [12, 13]. Estas

[1] https://www.sppsm.org/informemente

opiniões não surgem nas pessoas de forma intrínseca mas formam-se como resultado da perceção individual. Fatores influenciadores podem manifestar-se através das experiências que temos durante a nossa vida, das partilhas de opiniões com os nossos conhecidos, de aspetos culturais e da informação externa e de fraca qualidade que recebemos no nosso dia a dia. Durante muitos anos, o maior fator era a crença religiosa e muitas pessoas consideravam as doenças mentais como pecado e castigo de Deus [13]. Os avanços da ciência conseguiram dar outra razão para a existência destas condições de saúde, porém este tipo de mentalidade ainda continua a existir, mesmo que de forma mais leve ou disfarçada. Além disso, o humor é um fator presente no nosso quotidiano e, apesar de transmitir geralmente uma ideia positiva, também contribui para o aumento de estigma. As doenças mentais continuam a ser utilizadas em anedotas, situações cómicas e também de forma metafórica para se referir, simbolicamente, a outras situações que nada têm a ver com a área da Saúde.

Os meios de comunicação social também reforçam o estigma, na medida em que retratam os doentes como personagens estereotipadas no campo do Entretenimento ou os dramatizam na imprensa, visando captar a atenção dos leitores [13]. No Entretenimento, não apenas os doentes são retratados de forma preconceituosa como também os tratamentos que estes recebem. Durante as décadas de 1960 e 1970, a eletroconvulsoterapia foi apresentada de forma negativa em muitos livros, filmes e peças de teatro [11], e até hoje recebe destaque em filmes de terror, onde as clínicas psiquiátricas servem como um cenário que introduz medo e violência. Um estudo experimental [14], realizado para perceber o impacto dos meios de comunicação em indivíduos diagnosticados com episódios de depressão, verificou a presença de efeitos negativos e um aumento dos níveis de estigma no grupo testado após estes visualizarem um filme sobre um evento negativo referente ao transtorno da depressão. Deste modo, foi possível perceber que vários meios influenciam as nossas opiniões e que as doenças mentais são mal compreendidas tanto por pessoas saudáveis como pelos seus portadores.

Aliando-se a isto, um estudo realizado na Grã-Bretanha [12] também concluiu que, em muitos casos, as opiniões estigmatizantes não são apenas resultado da falta de conhecimento acerca das doenças mentais mas também derivam da forma negativa e dramática com que alguns dos doentes são retratados pelos meios de comunicação social, dando destaque às notícias jornalísticas.

2.1.3 A contribuição da comunicação social no desenvolvimento de estigma

Os meios de comunicação social, como formadores da opinião pública, devem ser responsáveis por contribuir para a construção de uma sociedade mais inclusiva e justa. No entanto, frequentemente deparámo-nos com textos sensacionais que dramatizam alguns factos ou até mesmo os falsificam. Muitas notícias que relatam crimes ou episódios violentos dão ênfase ao estado de saúde mental dos intervenientes, destacando a doença nos

títulos e conferindo um tom negativo ao texto. Além disso, as doenças mentais continuam a ser utilizadas de forma metafórica e em contextos que não se relacionam com o campo da Saúde. Os termos "esquizofrénico", "bipolar", "depressivo" e outros são utilizados como adjetivos para se referir, no sentido figurado, a situações ou pessoas de forma negativa, como por exemplo quando a palavra "esquizofrénico" é utilizada para se referir a uma situação ridícula ou contraditória.

Inúmeros estudos já realizados confirmam que os jornalistas contribuem para o desenvolvimento do estigma relativo à saúde mental. A metodologia mais utilizada caracteriza-se pela análise manual de notícias que fazem referência à saúde mental e pela procura de aspetos estigmatizantes nas mesmas, sendo para esta dissertação relevantes apenas os estudos que se focam no estigma resultante da utilização das doenças mentais no sentido metafórico. Na Europa, um estudo [15] analisou 695 notícias, que apresentavam termos relacionados com a saúde mental, de 20 jornais populares em Espanha no ano de 2010, e verificou a presença de 47.9% notícias estigmatizantes que utilizavam doenças mentais como metáforas. Na Grécia, analisaram-se 150 notícias, referentes apenas à esquizofrenia, e verificou-se a presença de 34% de notícias com estigma no sentido metafórico [16]. No Reino Unido, esse número constituiu 11% de um total de 600 notícias analisadas [17]. Nos Estados Unidos da América, foram analisados 1740 artigos e a percentagem dos que utilizavam a esquizofrenia como metáfora constituiu 28% [18]. No Brasil, um estudo [19], que também se focou apenas na esquizofrenia, verificou uma percentagem de 34%, num total de 229 notícias avaliadas, de estigma no sentido metafórico e concluiu que o mesmo está mais presente nos campos da Política, Economia e Entretenimento, onde desempenha o papel de caracterizar algo como "incoerente" e "absurdo". Foi observada a utilização não médica da doença de forma negativa tanto por parte de jornalistas como de indivíduos por eles entrevistados.

Estes resultados remetem para o facto de os níveis de estigma não serem iguais de país para país, o que vai ao encontro do estudo de Nawková et al. [20], que mostrou que o estigma na imprensa, associado às doenças mentais, não é representado de igual modo em todos os países e culturas. Assim, a situação em Portugal não pode ser replicada apenas se baseando nas conclusões efetuadas noutros países.

Atualmente, com base nas pesquisas efetuadas, apenas foram encontrados dois estudos portugueses neste âmbito. O primeiro [21] conduziu uma análise de conteúdo de notícias portuguesas sobre a saúde mental, publicadas entre janeiro e junho de 2015, e revelou que a depressão e esquizofrenia tendem a ser as doenças mais estigmatizadas na imprensa portuguesa. O segundo [22] avaliou a utilização da palavra "esquizofrenia" num total de 1058 notícias portuguesas, publicadas entre 2007 e 2013, e verificou que 40% das notícias eram estigmatizantes, sendo a área de destaque a Política. No entanto, estes estudos possuem algumas limitações, tais como um intervalo de tempo curto e antigo, e o facto de estas terem sido classificadas manualmente, apenas por dois anotadores no último. Seria de grande interesse avaliar um maior número de artigos recentes e verificar se os níveis de estigma aumentaram ou diminuíram, procurar outros aspetos pertinentes, aplicar outras

técnicas, nomeadamente computacionais, e também analisar outras fontes de informação.

2.2 Arquivo.pt

Praticamente toda a informação com que nos deparamos no dia a dia é disponibilizada na Internet. No entanto, grande parte dos dados que existiram no passado já não se encontra disponível, e mesmo os atuais acabam por ser perdidos passado algum tempo. O repositório Arquivo.pt surge para resolver este problema, ao preservar dados históricos da Web portuguesa e os disponibilizar, publicamente, às gerações presentes e futuras. O acesso e tratamento destes dados é, agora, mais rápido e organizado, existindo várias técnicas que auxiliam nestes processos.

2.2.1 Arquivo.pt como fonte de dados

O Arquivo.pt é um repositório de páginas web portuguesas arquivadas desde 1996 até hoje. Foi criado com o objetivo de preservar, em formato digital, uma grande quantidade de informação que poderia ser perdida, e permitir a sua utilização para fins de investigação [23]. Este repositório possui armazenadas várias categorias de páginas, que se encontram sob o domínio .PT ou têm interesse para a comunidade portuguesa [24], permite aceder a páginas que já não se encontram disponíveis *online* e apresenta também as funcionalidades de pesquisa e de acesso aos conteúdos através de uma *Application Programming Interface* (API). Atualmente, possui 13 158 milhões de ficheiros preservados, correspondendo 28 milhões deles a *websites* [25].

Como mencionado na Secção 1.1, o Arquivo.pt realiza, anualmente, o concurso "Prémio Arquivo.pt" [2]. No ano de 2018, o primeiro classificado foi o projeto "Conta-me histórias", que utiliza diferentes fontes jornalísticas para criar uma narrativa temporal simples sobre um dado tópico, excluindo a necessidade de os utilizadores realizarem uma procura e análise integral e individual dos dados. Em 2019, foi o projeto "meuParlamento.pt", uma aplicação móvel que permite aos utilizadores conhecer propostas legislativas apresentadas no Parlamento português e "votar" nas mesmas, de modo a perceber que partidos se aproximam mais do seu ponto de vista. O primeiro classificado em 2020 foi o projeto "Desarquivo", que permite a exploração de relações entre notícias portuguesas através de um grafo de ligações. Finalmente, em 2021, obteve o primeiro lugar o projeto "Major Minors", que consiste na primeira base de dados ontológica portuguesa que permite explorar a representação de minorias em notícias portuguesas.

Estes projetos são de código aberto e apresentam diferentes propostas e algoritmos, contribuindo, assim, para a comunidade científica e auxiliando no desenvolvimento de futuros trabalhos. No âmbito desta dissertação, foi feita uma análise mais focada nos mecanismos de recolha de informação mencionados e já testados para perceber quais as melhores abordagens a adotar.

[2] https://arquivo.pt/premios

2.2.2 Métodos de recolha de dados disponíveis

Tendo em conta a organização do Arquivo.pt e os mecanismos de acesso aos dados que este apresenta, existem dois métodos que podem ser adotados para recolher os dados, sendo eles o web *crawling* e a utilização de uma API.

Web *crawling*

Um *crawler* é um programa que percorre, analisa e armazena páginas alojadas na Internet. Estas páginas constituem a chamada *World Wide Web* (WWW) e estão organizadas numa estrutura em grafo, em que as páginas web são os vértices do grafo e as ligações entre elas as arestas. A tarefa dos *crawlers* é percorrer este grafo orientado, de modo a encontrar, sucessivamente, novas páginas a partir das anteriores e das interligações nelas contidas. Os web *crawlers* têm sido bastante utilizados em processos de extração e indexação automática de informação *online*, sendo a sua principal aplicação em motores de busca [26]. O repositório Arquivo.pt é preenchido através de processos de *crawling*, que ocorrem três a quatro vezes por ano [24].

A arquitetura geral de um web *crawler* [26] (Figura 2.1), tem como principais componentes:

- um *Frontier*, que armazena os *Uniform Resource Locator* (URL) das páginas ainda não visitadas;
- um *Page Downloader*, que transfere as páginas da WWW através de pedidos *Hypertext Transfer Protocol* (HTTP);
- um Repositório, que armazena as páginas extraídas numa base de dados.

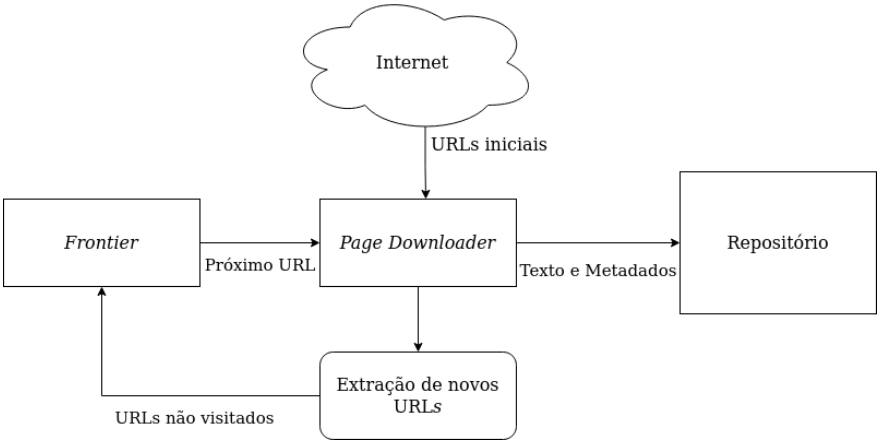


Figura 2.1: Arquitetura geral de um web *crawler*. Adaptado de [26].

Face à existência de uma ampla quantidade de páginas e de recursos finitos, como tempo, tráfego e memória, surgiu uma variação do *crawler* clássico que utiliza menos

recursos de *hardware* e é mais precisa. Designa-se por *Focused Crawler* e tem como objetivo a continuação da extração de novas páginas apenas a partir de páginas relevantes segundo uma certo critério, como por exemplo tópico, excluindo todas as restantes que não apresentam interesse. Este é o mecanismo de *crawling* que tem sido mais explorado na comunidade científica [27].

O processo de *crawling* implica também a posterior ocorrência de um processo de *scraping*. Web *scraping* é o processo que irá permitir extrair informação estruturada das páginas obtidas no processo de *crawling*. Esta informação pode ser o título do artigo contido na página, o conteúdo, a data de publicação ou outros dados relevantes, sendo que os mesmos podem ser, posteriormente, organizados e armazenados em formatos tabulares. Estes dados são recolhidos através da análise de identificadores de *HyperText Markup Language* (HTML), e uma das ferramentas mais utilizadas para este processo é a biblioteca de *Python BeautifulSoup*[3].

Martins et al. [28], primeiros classificados no "Prémio Arquivo.pt 2021", referem que a API pública do repositório Arquivo.pt mostrou-se ineficiente na recolha de páginas web para o seu projeto e, por isso, eles próprios construíram um *crawler* personalizado. O *crawler* por eles desenvolvido não é focado em tópicos, sendo a filtragem feita numa fase posterior. Esta estratégia pode ser adotada visto que o Arquivo.pt permite pesquisar e recolher várias versões de páginas iniciais dos jornais, que poderão servir como ponto de partida para o processo de *crawling*.

API do Arquivo.pt

API é um conjunto de protocolos HTTP que permitem a duas aplicações comunicarem e transferirem dados entre si. Uma aplicação que possui uma API funciona como servidor e permite que outras aplicações, que são seus clientes, solicitem-lhe conteúdos. Estes conteúdos são devolvidos, normalmente, em formatos já estruturados como *JavaScript Object Notation* (JSON) ou *Extensible Markup Language* (XML). A *Arquivo.pt API*[4] é a API pública do repositório Arquivo.pt, que permite recolher conteúdos preservados da Web portuguesa e seus metadados através de pesquisas de texto. Os dados são retornados em formato JSON e os pedidos podem ser feitos com diferentes parâmetros.

Tal como mencionado no ponto anterior, a API pública do Arquivo.pt não se mostrou útil para os objetivos dos primeiros classificados no concurso de 2021. No entanto, uma das limitações que eles apresentaram foi o facto de as pesquisas por termos apenas retornarem resultados com correspondência exata desses termos, ignorando as restantes páginas que tivessem esses termos implícitos ou descritos de outra forma. Isto não é um problema para o projeto desta dissertação, já que a pesquisa será feita com os exatos termos referentes às doenças mentais. Além disso, o processo de *crawling* é bastante extenso e apresenta algumas limitações como a complexidade de algumas páginas, a existência de mecanismos de bloqueio e a necessidade de grande pós-processamento [28]. Assim, é mais eficiente

[3] https://www.crummy.com/software/BeautifulSoup
[4] https://github.com/arquivo/pwa-technologies/wiki/Arquivo.pt-API

utilizar a API do repositório, recurso principal oferecido pelo Arquivo.pt e desenvolvido para tornar o processo de recolha de dados mais rápido e organizado.

2.3 Classificação automática de texto estigmatizante

Está comprovado que a incorporação de métodos de análise linguística na deteção de estigma *online* permite descobrir padrões associados, quais as doenças mais estigmatizadas e o público que as estigmatiza [29, 30]. Por sua vez, estes resultados permitem melhorar a atuação das campanhas de combate ao estigma, que funcionam melhor quando orientadas [30, 31].

Tendo em conta a quantidade massiva de dados existentes em formato digital, o seu processamento e extração de informação manuais iriam exigir um grande esforço humano e tempo. Visando automatizar estes processos, várias técnicas computacionais foram desenvolvidas e têm sido melhoradas ao longo dos anos, nomeadamente as técnicas de *machine learning*.

Artigos jornalísticos são dados textuais não estruturados e escritos em linguagem natural, facilmente percetível por humanos mas dificilmente percetível por máquinas. Neste âmbito, a tarefa de classificação automática torna-se mais complexa e surge a necessidade de aplicar técnicas adicionais, nomeadamente as técnicas de PLN e de mineração de texto, para que os algoritmos computacionais consigam interpretar o que está escrito e produzir melhores resultados.

2.3.1 *Machine learning* e o problema da classificação de texto

Machine learning é um tipo de IA cujos algoritmos permitem aos computadores extrair informação a partir de um conjunto de dados e aprender a tomar decisões de forma automática e independente, com intervenção mínima por parte dos programadores. Quanto à forma como o processo de aprendizagem é realizado, existem quatro tipos:

- Aprendizagem supervisionada, onde existe um conjunto de dados de treino já anotado ou classificado;
- Aprendizagem não supervisionada, onde não existem dados de treino já classificados e é o modelo o responsável por agrupar os dados e construir conhecimento;
- Aprendizagem semi-supervisionada, onde podem existir alguns dados de treino classificados e estes são utilizados apenas como base para guiar o modelo;
- Aprendizagem por reforço, onde o modelo aprende por experiência própria num processo de tentativa e erro.

A classificação de texto é um problema que pertence à categoria de algoritmos que utilizam aprendizagem supervisionada e consiste num processo automático de associar dados

textuais a uma dada classe. Pode ser do tipo binário, onde existem duas classes possíveis de classificação, ou multiclasse, onde existem mais do que duas classes de classificação. As classes são variáveis discretas e cada objeto apenas pode ser assinalado a uma.

O processo de classificação (Figura 2.2) depende muitas vezes da extração e seleção de atributos relevantes dos dados [32]. Os dados são conjuntos de objetos com os mesmos atributos, onde os objetos são independentes e possuem diferentes valores para cada atributo. Na classificação de texto, os objetos correspondem a documentos de texto e os atributos a palavras que podem estar presentes nesses documentos. A extração de atributos refere-se à representação dos atributos de cada objeto numa forma numérica, sendo a representação mais comum a vetorial [33]. A representação vetorial traduz-se, frequentemente, no modelo de *bag-of-words*, uma representação numérica que tem em conta apenas a frequência das palavras e não a ordem pela qual elas aparecem. Além disso, este modelo gera vetores esparsos, visto que, geralmente, o tamanho dos vetores, constituídos por todas as palavras únicas na coleção dos documentos, é bastante superior ao tamanho dos respetivos documentos. Outros dois modelos, que utilizam *word embeddings* e têm mostrado melhores resultados na representação dos objetos [33], são o *word2vec* [34] e o *GloVe* [35]. *Word embeddings* é uma representação que mapeia as palavras para um número limitado de certas categorias, produzindo vetores densos e de pequena dimensão, e que permite captar melhor a semântica das palavras, fazendo com que palavras similares tenham vetores também similares. A seleção de atributos refere-se à determinação dos atributos mais importantes e que melhor representam os objetos. Os métodos mais utilizados são a remoção de *stop words* e o processo de *stemming* [32]. Além disso, após o processo de seleção, pode também ser necessário projetar os atributos selecionados num espaço de menor dimensão, de modo a conseguir a representação que melhor caracteriza os dados e melhorar o desempenho do classificador. Os algoritmos mais utilizados na projeção de atributos são o *Principal Component Analysis* (PCA), *Singular Value Decomposition* (SVD) e *Linear Discriminant Analysis* (LDA) [33].

Os algoritmos de aprendizagem supervisionada mais utilizados no âmbito da classificação de texto são árvores de decisão, *Naive Bayes*, *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN) e *Logistic Regression* [32, 36–39]. SVM com *kernels* lineares e *Naive Bayes* são os algoritmos que têm apresentado melhores resultados neste problema [37, 39]. Gottipati et al. [40] referem também o bom desempenho do algoritmo *XGBoost*, um algoritmo baseado em árvores de decisão com *gradient boosting*. Para além destes modelos, a recente e popular abordagem assente em redes neuronais, o *deep learning*, tem se mostrado mais benéfica na captura do contexto dos dados e, consequentemente, tem gerado melhores resultados ao nível da classificação de texto [41].

Deep learning também se revela bastante eficaz no processo de deteção de metáforas. Gao et al. [42] demonstraram que arquiteturas baseadas em *Bidirectional Long Short Term Memory* (Bi-LSTM), conjugadas com a representação de *word embeddings*, apresentam resultados estado da arte na identificação e classificação de texto com expressões metafóricas. Por outro lado, Chen et al. [43] referem que os modelos neuronais não têm em

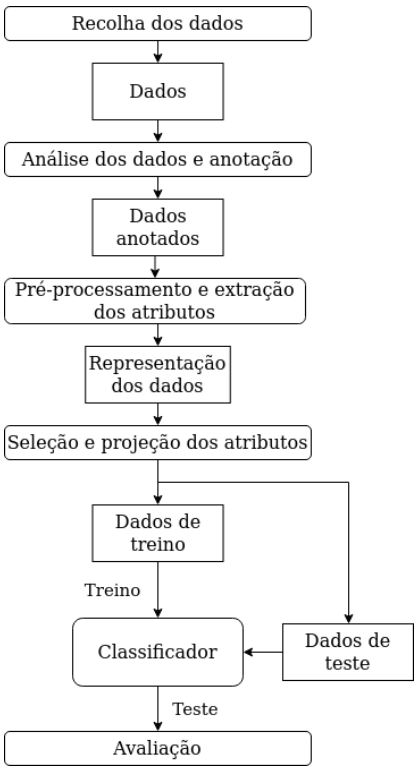


Figura 2.2: Processo geral de um sistema de classificação. Adaptado de [33].

atenção um aspeto bastante importante que é a inconsistência contextual entre os termos metafóricos e o resto dos termos que constituem uma frase. Segundo as suas observações, a inconsistência contextual pode ser medida e quanto maior for maior é a probabilidade de os termos alvo serem realmente metafóricos.

O processo da avaliação é a fase final da classificação de texto, e existem várias métricas e técnicas que podem ser aplicadas. As mais comuns para a classificação binária são a exatidão, a precisão, o *recall* e a *F1-score/F1* [33].

Não foi encontrado nenhum trabalho publicado no âmbito da classificação automática de texto estigmatizante em Portugal, sendo que todos os trabalhos encontrados realizavam a classificação manualmente. Mirończuk e Protasiewicz [33] realizaram um estudo sobre a quantidade de artigos escritos na área da classificação de texto e concluíram que os países China e Estados Unidos da América são os que possuem um maior número de artigos, com uma percentagem de 24.78% e 12.32% respetivamente, e que Portugal representa apenas 0.29% do total.

2.3.2 Processamento de linguagem natural

Não existe uma definição única para o PLN, podendo este ser descrito como um conjunto de métodos computacionais, pertencentes ao ramo da IA, que permitem aos computadores analisar e entender textos escritos por humanos, interpretar e gerar falas e identificar sentimentos expressos. Deste modo, a classificação automática de texto estigmatizante é

um problema de processamento de linguagem natural.

A linguagem natural pode ser processada ao nível fonético, morfológico, léxico, sintático, semântico, discursivo e pragmático. Estes níveis são dependentes e devem ser analisados simultaneamente num único processo [44], no entanto, uma solução única ainda não foi concebida e nem todos os problemas necessitam de percorrer todas as etapas. O processo de classificação automática de texto faz parte das tarefas de PLN mas, ao mesmo tempo, utiliza e aplica as restantes, nomeadamente na fase de pré-processamento dos dados. Algumas das mais relevantes são:

- *Tokenization* — repartição do texto do documento numa sequência de *tokens* ou termos, normalmente palavras;
- Conversão para letras minúsculas de todas as palavras do texto — normalização que permite diminuir o número de termos únicos, mas que pode provocar a perda do significado semântico de algumas palavras;
- *Part-of-Speech* (POS) *tagging* — categorização das palavras em classes gramaticais, como advérbios, adjetivos, pronomes, verbos e outros;
- Remoção de *stop words* — remoção de palavras com alta frequência e sem importância semântica para o texto;
- *Dependency Parsing* — identificação da palavra raiz das frases e a sua relação com as restantes;
- *Lemmatization* e *stemming* — transformação de palavras derivadas à sua forma raiz, tendo em conta o contexto e ignorando-o, respetivamente. No processo de *lemmatization*, as palavras são reduzidas ao seu lema, e no processo de *stemming* ao seu radical. Melo e Figueiredo [45] referem que os processos de *lemmatization* e *stemming* para a língua portuguesa são pouco precisos, face à escassez de ferramentas próprias para o português;
- *Named-entity recognition* (NER) — localização e classificação de entidades presentes no texto, tais como nomes de pessoas, organizações, localizações, doenças e outras.

Estas técnicas de pré-processamento, bem como a etapa de seleção e extração de atributos, são bastante importantes nos problemas de classificação de texto, estando já demonstrado o seu impacto no aumento da precisão dos classificadores [46]. Oliveira e Mershmann [46] realizaram um estudo sobre a conjugação de diferentes combinações de técnicas de pré-processamento com diferentes algoritmos de classificação, para a língua portuguesa, e concluíram que o conjunto das melhores combinações varia quando varia o algoritmo de classificação utilizado. Assim, é importante escolher as técnicas de pré-processamento tendo em conta também a natureza dos classificadores a utilizar. Para além das técnicas referidas, outras técnicas de PLN mais complexas, e que podem ser relevantes para esta

dissertação, são o *topic modeling* e análise de sentimento.

Topic modeling

É um processo de aprendizagem não supervisionada que consiste na descoberta de tópicos existentes num texto, através da análise da similaridade e da ocorrência de termos. Tem por base as ideias de que cada documento de texto apresenta vários tópicos e que cada tópico é representado por uma coleção de palavras. Os algoritmos básicos mais utilizados [47, 48] são:

- *Latent Semantic Analysis* (LSA) — é um dos algoritmos fundamentais de *topic modeling* e, em termos gerais, tem por base a decomposição da matriz inicial documento--termos em duas matrizes separadas, documento-tópicos e tópico-termos, recorrendo à utilização do algoritmo SVD;

- *Probabilistic Latent Semantic Analysis* (PLSA) — utiliza o método probabilístico, em vez de SVD, para encontrar o modelo probabilístico de tópicos que conseguem gerar a matriz inicial documento-termos;

- LDA — é uma versão bayesiana do PLSA, que utiliza processos *Dirichlet* nas distribuições. É a técnica mais utilizada e que apresenta, tipicamente, melhores resultados. No entanto, possui a limitação da captura de contexto, existindo várias extensões deste modelo que tentam resolver este problema;

- *Non-negative Matrix Factorization* (NMF) — é um modelo que utiliza álgebra linear e atua melhor em textos muito curtos.

Além da limitação da captura de contexto, estes modelos também apresentam dificuldade na captura da semântica das palavras e necessidade de predefinição do número de tópicos a descobrir, de modo a serem produzidos melhores resultados. Angelov [49] apresenta o algoritmo *top2vec*, que deteta automaticamente tópicos presentes num documento, sem a necessidade de pré-processamento, e gera representações que têm em conta o conteúdo semântico do texto. Este algoritmo tende a gerar melhores resultados que os tradicionais modelos, PLSA e LDA, baseados nas distribuições de palavras.

Recentemente, também têm sido utilizadas abordagens de *deep learning*, que apresentam melhores resultados em captar contextos. Zhang et al. [50] propõem um modelo melhorado do *attention-based Long Short Term Memory* (LSTM), que captura características contextuais de sequências de texto. Este modelo obtém tópicos semânticos com base numa camada de *topic modeling* e na restrição de similaridade, e gera representações semânticas de documentos usando *tree-structured* LSTM. A representação de documentos gerada demonstrou melhor performance do que modelos que são estado da arte nos campos de *topic modeling*, classificação de texto e recuperação de informação.

A identificação de tópicos pode ser particularmente relevante para a identificação de

metáforas, na medida em que a presença de metáfora cria um vocabulário incoerente no tópico ou contexto geral do texto. Neste âmbito, Jang et al. [51] exploraram uma abordagem mais complexa que utiliza, como atributos para o treino de um classificador SVM, a similaridade de tópicos entre as frases do texto, medida com recurso ao modelo *Sentence-LDA* [52], e a presença de palavras que expressam emoções. O *topic modeling* não é feito ao nível das palavras mas ao nível das frases e está assente nas ideias de que cada frase tem um tópico principal, que todas as palavras que constituem uma frase pertencem ao mesmo tópico e que frases com expressões metafóricas apresentam tópicos distintos dos tópicos das restantes frases literais. A exploração de emoções no texto também surge como um critério relevante visto que, na maioria das vezes, as metáforas são utilizadas para expressar as emoções do comunicador e não o sentido literal dos termos [53].

Análise de sentimento

Análise de sentimento refere-se à classificação de sentimento humano presente em dados textuais. Os sentimentos são, geralmente, classificados como positivos, negativos ou neutros, e dependem do contexto onde estão inseridos, podendo conceitos iguais ter classificações diferentes em diferentes domínios. A classificação pode ser feita ao nível do documento, considerado como uma unidade única, ao nível das frases presentes no documento ou ao nível dos aspetos presentes em entidades referenciadas no documento [54]. Os métodos mais utilizados para a classificação de sentimento são os baseados em léxicos emocionais e os que constituem o estado da arte na área de *machine learning*.

Quanto aos métodos baseados em léxicos, o processo geral segue a ideia de que as palavras têm uma pontuação de sentimento a elas associada, expressa em dicionários. Uma abordagem clássica passa por realizar a soma das pontuações de todas as palavras que existem, em simultâneo, no texto alvo e no dicionário, refletindo a pontuação final o sentimento geral conferido ao texto. Existem vários dicionários de sentimento para a língua inglesa, nomeadamente o *WordNet Affect*[5], o *SentiWordNet*[6], o *Affective Norms for English Words* ou *ANEW*[7] e o *NRC Word-Emotion Association Lexicon* ou *EmoLex*[8] [55]. No entanto, Pereira [56] refere que os recursos existentes para processar outras línguas são escassos. Para a língua portuguesa existem muito poucos dicionários, sendo eles o *opLexicon*[9] e o *Linguistic Inquiry Word Count* (LIWC) para o português do Brasil (Brazilian Portuguese LIWC 2007 Dictionary[10]), e o *SentiLex-PT*[11] para o português de Portugal. É de referir também que o LIWC é um dicionário que possui não só mapeamento de palavras para sentimentos expressos como também para muitas outras categorias. Estas categorias estão, essencialmente, agrupadas em quatro tipos: processos linguísticos básicos,

[5] https://wndomains.fbk.eu/wnaffect.html
[6] https://github.com/aesuli/SentiWordNet
[7] https://csea.phhp.ufl.edu/media/anewmessage.html
[8] https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
[9] https://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/oplexicon/
[10] http://143.107.183.175:21380/portlex/index.php/pt/projetos/liwc
[11] https://b2share.eudat.eu/records/93ab120efdaa4662baec6adee8e7585f

processos psicológicos, expressões relativas e preocupações pessoais [57]. Onan e Tocoglu [58] utilizaram atributos extraídos a partir do LIWC para identificar sátira em notícias turcas e concluíram que os mesmos geram melhores resultados de classificação do que modelos de *bag-of-words*.

Outra alternativa consiste na tradução do texto português para a língua inglesa e na utilização das ferramentas automáticas específicas para o inglês para realizar a classificação. Araújo et al. [59] analisaram esta abordagem, ao nível de frases individuais, e concluíram que a mesma apresenta melhores resultados que a utilização de léxicos desenvolvidos especificamente para a linguagem alvo. Algumas ferramentas que apresentaram bons resultados foram *VADER* [12], *SentiStrength*[13] e *SO-CAL*[14]. No entanto, deve-se ter em conta que a tradução pode captar apenas aspetos gerais e similares entre as duas linguagens, podendo ignorar outros mais específicos mas com grande impacto. Por outro lado, Tavares et al. [60] concluíram que a tradução de notícias portuguesas para inglês e a utilização de ferramentas de análise de sentimento específicas para o inglês não permitem a obtenção de resultados precisos no domínio da Economia. Em alternativa, adotaram uma abordagem baseada em regras, que demonstrou melhores resultados.

Melville et al. [61] utilizaram uma metodologia híbrida para a análise de sentimento, que combinou a utilização de dicionários e de algoritmos de *machine learning*. As experiências realizadas mostraram que a combinação de informações lexicais com algoritmos de *machine learning* produz melhores resultados que a utilização separada das duas metodologias. Oliveira e Mershmann [46] mostraram que o classificador SVM apresenta bom desempenho quando conjugado com diferentes combinações de técnicas de pré-processamento e que deve ser considerado para a análise sentimental de textos portugueses.

Face à existência de limitações, como a necessidade de dicionários suficientemente amplos ou de um grande número de dados de treino para os modelos de classificação, outras técnicas alternativas têm sido desenvolvidas neste âmbito [36]. Uma abordagem que surgiu recentemente é a aprendizagem por transferência, ou *transfer learning*, onde o conhecimento aprendido num domínio é aplicado noutro domínio similar. Chintalapudi et al. [62] apresentam a utilização do mais recente modelo de *transfer* e *deep learning* para análise de sentimento em dados textuais, designado por *Bidirectional Encoder Representations from Transformers* (BERT), que demonstra melhores performances de classificação em contraste com os algoritmos clássicos de *machine learning*. BERT foi lançado pelos investigadores da *Google AI*, em 2018, e é apresentado pela primeira vez no artigo [63], onde é descrito como um modelo pré-treinado de PLN que é capaz de entender melhor o significado e relações das palavras numa frase, bem como o contexto onde estão inseridas, ao realizar a leitura da frase toda de uma só vez. Está pré-treinado num grande corpus de texto e pode ser adaptado para atuar em outros domínios sem grandes alterações na sua arquitetura base. Este modelo pode ser usado em várias tarefas de PLN e os seus re-

[12] https://github.com/cjhutto/vaderSentiment
[13] http://sentistrength.wlv.ac.uk/
[14] https://github.com/sfu-discourse-lab/SO-CAL

sultados já ultrapassam os que constituem estado da arte neste campo. *BERTimbau* [64] é o modelo BERT treinado na língua portuguesa do Brasil. Existe também uma versão mais robusta e otimizada do BERT, proposta pelos investigadores da *Facebook AI* e designada por *RoBERTa* [65], que também tem apresentado melhores resultados em tarefas de classificação do que modelos clássicos que usam *bag-of-words* [66].

A análise de sentimento é uma técnica que se está a desenvolver cada vez mais, porém ainda apresenta alguns problemas e desafios. Para além das limitações gramaticais mais comuns e da existência de grande diversidade linguística, é também necessário ter em conta aspetos que podem alterar totalmente o sentido dos textos, como a presença de ironia, sarcasmo, metáforas, negações e expressões idiomáticas cujo significado e sentimento não pode ser identificado através da análise individual de cada termo.

2.3.3 Mineração de texto

Mineração de texto refere-se ao processo de extração de informação útil de texto não estruturado através da identificação e exploração de padrões. As suas técnicas são aplicadas em textos escritos em linguagem natural humana e os mesmos são convertidos em dados estruturados, que podem ser mais facilmente armazenados em formatos tabulares e analisados para formar conhecimento. Técnicas de mineração de texto têm sido cada vez mais estudadas e são principalmente aplicadas na análise de *Big Data*, uma área em grande crescimento [44].

O processo clássico de mineração de texto [67] é descrito por um modelo geral constituído por quatro fases, onde as primeiras duas são as principais:

- Pré-processamento, onde os dados não estruturados são convertidos em estruturados através da aplicação de diferentes métodos que podem incluir métodos de PLN e de *machine learning*. As técnicas que mais se destacam nesta etapa são a classificação de texto, *clustering* e recuperação de informação;
- Aplicação de operações de mineração, em que são descobertos padrões, tendências e conhecimento, sendo habitualmente utilizadas técnicas de distribuição e associação;
- Apresentação de resultados, através de ferramentas de visualização;
- Refinamento, onde técnicas de otimização e pós-processamento podem ser aplicadas, tais como filtragem, *clustering*, ordenação e outras.

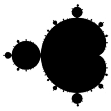
Deste modo, mineração de texto pode ser vista como um processo que, quando utiliza PLN e modelos de *machine learning*, consegue compreender a linguagem humana e automatizar tarefas de análise de texto.

2.3.4 Ferramentas existentes

Tendo em conta a existência e relevância de ferramentas para o PLN e mineração de texto, e para a implementação de modelos de *machine* e *deep learning* para a classificação de texto, as *frameworks* e bibliotecas mais utilizadas [56, 68] podem ser comparadas na Tabela 2.1:

Tabela 2.1: Comparação de *frameworks* e bibliotecas existentes para processamento de linguagem natural, mineração de texto e implementação de modelos de classificação de texto.

***Frameworks* e bibliotecas**	**Linguagem**	**Licença**	**Características mais relevantes**
 *Natural Language Toolkit* (NLTK) [69]	*Python*	*Apache 2.0*	Classificação de texto (Árvores de decisão, *Naive Bayes* e outros); *Tokenization*, POS *tagging*, *stemming*, NER e outros; Análise de sentimento (*VADER*); Processamento para a língua portuguesa; Métricas de avaliação.
 *scikit-learn* [70]	*Python*	*BSD*	Extração de atributos; Redução de dimensão; *Topic modeling* (LDA, NMF, *clustering*); Grande variedade de algoritmos clássicos de classificação (*Random Forest*, KNN, SVM, redes neuronais e outros); Métricas de seleção e avaliação de modelos.
 *spaCy* [71]	*Python*	*MIT*	Classificação de texto, *lemmatization*, análise sintática, *tokenization*, NER, *dependency parsing* e outros; Processamento para a língua portuguesa; Modelos neuronais pré-treinados (BERT e outros).
nlpnet [72]	*Python*	*MIT*	Pré-processamento com *deep learning*; POS *tagging*, análise semântica e *dependency parsing*; *Word embeddings*; Processamento para a língua portuguesa.
 *Gensim* [73]	*Python*	*GNU LGPLv2.1*	*Topic modeling* (LDA, LSA, NMF); Modelos de representação (*bag-of-words*, *tf-idf*, *word2vec*, *doc2vec*, *fastText*).

TensorFlow *TensorFlow* [74]	*Python*, *C++*, *JavaScript*, *Java*	*Apache 2.0*	Pré-processamento e construção de modelos com *deep learning*; Repositório com modelos neuronais pré-treinados(BERT e outros); *Word embeddings*; Tradução de texto.
 *Transformers* [74]	*Python*	*Apache 2.0*	Proporciona APIs para transferir e treinar modelos pré-treinados de *deep learning*; Integração com *Jax*, *PyTorch* e *TensorFlow*.
Stanza *Stanza* [75]	*Python*	*Apache 2.0*	Pré-processamento com *deep learning*; *Tokenization*, análise sintática, *dependency parsing*, NER e outros; Possui interface *Python* para o pacote *CoreNLP Java*; Análise de sentimento; Processamento para a língua portuguesa; Modelos neuronais pré-treinados.
 TextBlob *TextBlob* [76]	*Python*	*MIT*	POS *tagging*, correção ortográfica e outros; Tradução de texto; Análise de sentimento; Classificação de texto (Árvores de decisão, *Naive Bayes*).

Grande parte das ferramentas apresentadas depende da utilização das outras e não consegue responder ao problema desta dissertação por si só. Assim, é necessário conjugá-las, procurando perceber quais delas melhor se adequam a cada etapa do processo. Além disso, é possível concluir que a linguagem predominante é o *Python*, o que era expetável visto que, atualmente, o *Python* é uma das linguagens mais utilizadas e adequadas para as tarefas referidas.

2.4 Sumário

As doenças mentais afetam uma grande parte da população portuguesa e nem sempre as mesmas são retratadas de forma rigorosa e objetiva nos meios de comunicação social. Os jornais de notícias, que são agora também acedidos através da Internet, revelam estigmatização de doenças mentais ao utilizar as mesmas num sentido figurado e fora de contextos adequados. Vários estudos confirmam esta realidade, e existem apenas dois realizados no cenário português, tendo ambos efetuado a deteção de estigma manualmente.

O Arquivo.pt, além da sua ampla base de dados de páginas web portuguesas, proporciona ferramentas que permitem a pesquisa e recolha das mesmas, podendo estes processos

ser realizados através da implementação de web *crawling* ou através da utilização da API fornecida. A utilização da API mostra-se mais benéfica para o presente projeto ao permitir recolher os dados de forma mais eficiente e organizada.

No processo de classificação automática de texto, alguns dos algoritmos que têm apresentado melhores resultados são SVM e *Naive Bayes*. No entanto, estes resultados variam com o contexto onde são aplicados e existem ainda muito poucos trabalhos realizados no âmbito de deteção de estigma. A subárea do *deep learning* também tem ganho destaque e apresentado resultados mais precisos, ao captar melhor contextos e semântica dos textos. Por outro lado, a área de PLN cresceu bastante nos últimos anos e apresenta, atualmente, vários benefícios a nível do pré-processamento de dados e em tarefas como *topic modeling* e análise de sentimento. No âmbito do *topic modeling*, os algoritmos fundamentais correspondem, essencialmente, aos algoritmos LDA e PLSA. No ramo da análise de sentimento, as principais abordagens passam pela utilização de léxicos emocionais e de algoritmos de *machine learning*. O *deep learning* também tem sido conjugado com tarefas de PLN, tanto a nível da representação dos textos como nos processos de classificação, destacando-se o uso de modelos pré-treinados como BERT. No entanto, apesar dos grandes avanços, a área de PLN ainda apresenta várias lacunas e limitações, nomeadamente no processamento de textos onde se verifica a presença de ironia, metáforas, expressões idiomáticas e ambiguidade de palavras.

Para auxiliar na implementação destas técnicas de IA, existem várias ferramentas disponíveis, sendo a linguagem *Python* a que mais se destaca neste âmbito. No entanto, a área de classificação automática de texto encontra-se muito pouco desenvolvida em Portugal, o que se revela também na falta de ferramentas próprias para a língua portuguesa. Este aspeto tem impulsionado a tradução dos textos portugueses para a língua inglesa, de modo a permitir o uso de mais ferramentas, mas também incentiva a exploração da área, ao abrir espaço para a geração de novo conhecimento.

Capítulo 3

Recolha dos dados

Neste capítulo, são apresentadas as principais etapas do processo de recolha dos dados do repositório Arquivo.pt. Esta é a primeira fase de implementação do projeto desta dissertação, sendo que todas as restantes podem ser visualizadas no Apêndice A. Primeiramente, é descrito o funcionamento da *Arquivo.pt API* e a natureza dos dados que a mesma devolve. De seguida, são apresentados os parâmetros de pesquisa que foram utilizados, sendo feito um especial foco nos jornais de notícias *online* considerados. Por fim, é apresentada a metodologia geral do processo de recolha e os resultados obtidos do mesmo, que se traduzem nos dados/artigos recolhidos.

O capítulo termina com um breve sumário dos pontos descritos.

3.1 Fonte e natureza dos dados

Tal como mencionado anteriormente, a fonte de dados para o projeto desta dissertação é o repositório Arquivo.pt, e os mesmos foram recolhidos através da API pública *Arquivo.pt API.* Esta API permite efetuar pesquisas de texto, onde são retornadas todas as páginas web preservadas que contêm, no seu conteúdo, os termos de pesquisa, ou pesquisas de URL, onde são retornadas todas as versões preservadas do URL pesquisado. Para a recolha de dados, foi utilizada apenas a pesquisa de texto.

A pesquisa é efetuada através da construção de uma *query* de pesquisa, onde podem ser utilizados vários parâmetros. Para cada *query* de pesquisa, a resposta é retornada em formato JSON. A resposta é constituída por um conjunto de campos descritivos e pelo conjunto dos elementos retornados. Cada elemento retornado corresponde a uma página web arquivada e seus metadados. Uma limitação da funcionalidade de pesquisa de texto é o facto de, para cada pesquisa, apenas ser devolvida uma resposta com no máximo 2000 elementos. Esta limitação verifica-se tanto na pesquisa através da API como através do campo de pesquisa original no *website* do Arquivo.pt. Os elementos retornados encontram-se ordenados pelo critério de relevância, dos mais relevantes para os menos relevantes [77]. No entanto, após a realização de uma análise inicial da API, verificou-se que o número de elementos retornados, para as pesquisas a efetuar no âmbito desta dissertação, nunca

excede o valor máximo de 2000 resultados. Além disso, caso surgisse uma resposta que excedesse esse valor, a pesquisa poderia ser segmentada ao nível de certos parâmetros, como a data inicial e final de arquivamento, de modo a que fossem retornados todos os resultados existentes mas de uma forma repartida. Mais informação sobre a descrição da API pode ser encontrada no Apêndice B.

Existem duas hipóteses para aceder ao conteúdo de uma página web preservada. A primeira hipótese é através do campo "linkToExtractedText", que disponibiliza um URL onde se encontra todo o conteúdo textual extraído da mesma. Este conteúdo encontra-se não estruturado e apresenta todos os elementos textuais misturados, sendo bastante complicado encontrar as partes relevantes do texto e as agrupar. A segunda hipótese é através do campo "linkToOriginalFile", que disponibiliza o URL que dá acesso ao HTML original da página preservada. Para extrair o seu conteúdo é necessário realizar um processo de web *scraping*. Além disso, é preciso ter em conta que as páginas pertencentes a diferentes domínios apresentam diferentes estruturas no seu HTML, sendo essencial uma definição de regras para um processamento adequado.

3.2 Objetivo e parâmetros de pesquisa

Tendo em conta a finalidade de classificar automaticamente artigos estigmatizantes de doenças mentais publicados nos jornais digitais portugueses ao longo dos anos, o objetivo do processo de recolha de dados é recolher os dados estruturados que poderão ser mais relevantes para o processo, sendo eles o título do artigo, o seu conteúdo, a data de publicação, a data de arquivamento, o nome do jornal e os URLs para a versão original e arquivada. Após a análise do método de acesso aos dados e da natureza dos mesmos, prosseguiu-se à determinação dos parâmetros de pesquisa a utilizar.

Começando pelos termos de pesquisa, foi decidido focar-se em notícias que estigmatizam a doença mental da esquizofrenia, visto estudos anteriores apresentarem-na como uma das doenças mais utilizadas, pela imprensa, num sentido metafórico. Esta doença faz parte das perturbações menos comuns mas, ao mesmo tempo, das perturbações que mais aparecem no nosso vocabulário de termos utilizados fora do seu contexto original. Além disso, para aumentar o número de artigos recolhidos foram também tidos em conta termos referentes à psicose, visto esta ser uma condição que faz parte dos sintomas da doença da esquizofrenia e ambos os transtornos serem, muitas vezes, utilizados de forma relacionada.

Tendo em conta todas as palavras que é possível derivar das palavras "esquizofrenia" e "psicose", através do uso de sufixos de derivação e sem perder o significado das mesmas, e a probabilidade de encontrar essas palavras nos textos, foram estabelecidos os termos de pesquisa ["esquizofrenia", "esquizofrénico", "esquizofrenico", "esquizofrénica", "esquizofrenica", "esquizofrénicas", "esquizofrenicas", "esquizofrénicos", "esquizofrenicos", "esquizofrenicamente", "esquizofrenizar", "psicose", "psicótica", "psicotica", "psicóticas", "psicoticas", "psicótico", "psicotico", "psicóticos", "psicoticos"]. A API de pesquisa é *case insensitive*, não havendo necessidade de distinguir entre os termos que começam por letra

minúscula e maiúscula, e é *accent sensitive*. Assim, foram recolhidas todas as páginas web de jornais de notícias que contenham pelo menos um dos termos do conjunto.

Um grande número de dados é essencial, no entanto, para um tratamento eficiente dos mesmos, é necessário ter em conta que nem todos eles serão relevantes, podendo alguns apenas adicionar complexidade ao processo. Dada a grande quantidade de jornais portugueses e o facto de nem todos eles apresentarem forte probabilidade de utilização de termos referentes à esquizofrenia e psicose num sentido metafórico, foram selecionados apenas nove jornais digitais. Os critérios de seleção foram a popularidade do jornal *online*, a sua relevância no âmbito do projeto e a sua longevidade. Quanto à popularidade, um estudo desenvolvido pela Entidade Reguladora para a Comunicação Social revelou, com base num inquérito nacional realizado em 2014, as preferências dos utilizadores nas fontes de notícias portuguesas *online* [78]. Todas as fontes citadas no estudo foram exploradas, com recurso à API do Arquivo.pt, de modo a obter uma primeira visão sobre os resultados expectáveis. Os jornais que apresentaram melhores resultados foram analisados quanto à sua longevidade, bem como quantidade e histórico de versões no Arquivo.pt. Por fim, tendo em conta as observações efetuadas, concluiu-se que os jornais que poderão vir a ser mais relevantes para o projeto são:

- **Público**[1] — um jornal diário português fundado em 1990 e propriedade da empresa *PÚBLICO Comunicação Social*, com sede em Maia. A partir de 1995 passou a estar disponível na Internet[2], possuindo atualmente um dos *websites* de notícias mais acessados em Portugal.

  O endereço do *website* do Público, e de todos os restantes jornais, não foi sempre constante ao longo dos anos, podendo estes terem pertencido a outros domínios e apresentarem vários subdomínios. Assim, foi necessário descobrir todos os endereços dos *websites* arquivados no Arquivo.pt para cada jornal. Para isto, foi utilizada a informação presente no relatório técnico de Cunha [79], disponibilizado publicamente na página do Arquivo.pt. No entanto, o relatório apenas possui dados referentes a alguns jornais e ao período de tempo entre 1996 e 2016, podendo também estar desatualizado. Assim, foi também necessário utilizar o sistema de pesquisa de URL do Arquivo.pt, para verificar a veracidade dos dados do relatório e as diferentes versões do endereço arquivadas depois de 2016. É de salientar que da leitura do relatório e das descrições dos projetos premiados nos anos anteriores do concurso, verificou-se que o repositório possui bastantes problemas e limitações, sendo que alguns dos endereços podem apresentar problemas de preservação ou nem serem retornados na pesquisa. A descoberta dos endereços do jornal Público, parceiro de comunicação oficial do "Prémio Arquivo.pt 2022", foi facilitada, devido à disponibilização, no *website* do Arquivo.pt, de uma lista[3] de seus domínios e subdomínios entre 1996 e 2019.

[1] https://www.publico.pt/

[2] https://www.publico.pt/2005/09/22/portugal/noticia/publicopt-um-jornal-no-ciberespaco-desde-1995-1233488

[3] https://sobre.arquivo.pt/pt/colabore/premios-arquivo-pt/premio-arquivo-pt-2022/

A Tabela 3.1 apresenta os endereços mais relevantes, referentes ao jornal Público, ao longo dos anos.

Tabela 3.1: Diferentes endereços do jornal Público e intervalos de tempo que possuem suas versões no Arquivo.pt.

Endereço	**Intervalo de tempo**
publico.pt / www.publico.pt	1996 – 2022
ultimahora.publico.pt	1999 – 2009
jornal.publico.pt	2000 – 2016
dossiers.publico.pt	2001 – 2011
desporto.publico.pt	2001 – 2012
www.publico.clix.pt	2005 – 2009
digital.publico.pt	2006 – 2011
economia.publico.pt	2007 – 2012
m.publico.pt	2011 – 2013
blogues.publico.pt	2011 – 2021

- **Observador**[4] — um jornal diário português com sede em Lisboa e fundado em 2014 por um grupo de jornalistas e investidores insatisfeitos com a informação produzida pela comunicação social em Portugal[5]. É o único jornal português inteiramente digital.

  Dada a natureza recente do jornal, o único endereço encontrado foi "observador.pt", possuindo o Arquivo.pt material preservado de 2014 a 2022.

- **Diário de Notícias**[6] — um jornal diário português fundado em 1864 e propriedade da empresa *Global Media Group*, com sede em Lisboa. Passou a estar disponível na Internet a partir de 1995 [79].

  A Tabela 3.2 apresenta os endereços mais relevantes, referentes ao jornal Diário de Notícias, ao longo dos anos.

Tabela 3.2: Diferentes endereços do jornal Diário de Notícias e intervalos de tempo que possuem suas versões no Arquivo.pt.

Endereço	**Intervalo de tempo**
www.dn.pt	1996 – 2022
dn.sapo.pt / www.dn.sapo.pt	2001 – 2012

[4] https://observador.pt/
[5] https://observador.pt/explicadores/tudo-o-que-precisa-de-saber-sobre-o-observador/
[6] https://www.dn.pt/

As pesquisas com o endereço "www.dn.pt" apresentam vários elementos correspondentes a páginas de outros jornais. No entanto, este endereço também foi tido em conta visto que alguns dos jornais pertencem ao conjunto de dados e esta pesquisa pode retornar resultados relevantes. Foi realizado um posterior processo de filtragem para não ocorrerem duplicações de conteúdo.

- **Expresso**[7] — um jornal semanal português fundado em 1973 e propriedade da empresa *Impresa*, com sede em Lisboa. Passou a estar disponível na Internet a partir de 1997 [79].

  A Tabela 3.3 apresenta os endereços mais relevantes, referentes ao jornal Expresso, ao longo dos anos.

Tabela 3.3: Diferentes endereços do jornal Expresso e intervalos de tempo que possuem suas versões no Arquivo.pt.

Endereço	**Intervalo de tempo**
expresso.pt	1998 – 2022
aeiou.expresso.pt	2008 – 2012
expresso.sapo.pt	2012 – 2015

- **Correio da Manhã**[8] — é um jornal diário português fundado em 1979, e propriedade da empresa *Cofina*, com sede em Lisboa. Passou a estar disponível na Internet a partir de 1995 [79].

  A Tabela 3.4 apresenta os endereços mais relevantes, referentes ao jornal Correio da Manhã, ao longo dos anos.

Tabela 3.4: Diferentes endereços do jornal Correio da Manhã e intervalos de tempo que possuem suas versões no Arquivo.pt.

Endereço	**Intervalo de tempo**
www.correiomanha.pt	1996 – 2015
www.correiodamanha.pt	2001 – 2009
www.cmjornal.xl.pt	2010 – 2016
www.cmjornal.pt	2010 – 2022

- **Jornal de Notícias**[9] — é um jornal diário português fundado em 1888 e propriedade da empresa *Global Media Group*, com sede no Porto. Passou a estar disponível na

[7] https://expresso.pt/
[8] https://www.cmjornal.pt/
[9] https://www.jn.pt/

Internet a partir de 1995[10].

A Tabela 3.5 apresenta os endereços mais relevantes, referentes ao jornal Jornal de Notícias, ao longo dos anos.

Tabela 3.5: Diferentes endereços do jornal Jornal de Notícias e intervalos de tempo que possuem suas versões no Arquivo.pt.

Endereço	**Intervalo de tempo**
jn.pt / www.jn.pt	1998 – 2022
jn.sapo.pt	2002 – 2011

- **Sábado**[11] — é uma revista semanal portuguesa fundada em 2004 e propriedade da empresa *Cofina*, com sede em Lisboa.

  A Tabela 3.6 apresenta os endereços mais relevantes, referentes à revista Sábado, ao longo dos anos.

Tabela 3.6: Diferentes endereços da revista Sábado e intervalos de tempo que possuem suas versões no Arquivo.pt.

Endereço	**Intervalo de tempo**
www.sabado.xl.pt:80	2006 – 2007
www.sabado.xl.pt	2008
sabado.pt / www.sabado.pt	2009 - 2022

- **Visão**[12] — é uma revista semanal portuguesa fundada em 1993 e propriedade da empresa *Trust in News*, com sede em Lisboa. Passou a estar disponível na Internet a partir de 2001[13].

  A Tabela 3.7 apresenta os endereços mais relevantes, referentes à revista Visão, ao longo dos anos.

[10] https://www.jn.pt/nacional/media/jn-online-nasceu-ha-20-anos--4702293.html
[11] https://www.sabado.pt/
[12] https://visao.sapo.pt/
[13] https://www.publico.pt/2001/03/29/portugal/noticia/visao-online-arrancou-hoje-16921

Tabela 3.7: Diferentes endereços da revista Visão e intervalos de tempo que possuem suas versões no Arquivo.pt.

Endereço	**Intervalo de tempo**
aeiou.visao.pt	2009 – 2012
visao.sapo.pt	2012 - 2022

- **A Bola**[14] — é um jornal desportivo português fundado em 1945 e propriedade da empresa *Sociedade Vicra Desportiva*, com sede em Lisboa. Passou a estar disponível na Internet a partir de 2000[15].

  A Tabela 3.8 apresenta os endereços mais relevantes, referentes ao jornal A Bola, ao longo dos anos.

Tabela 3.8: Diferentes endereços do jornal A Bola e intervalos de tempo que possuem suas versões no Arquivo.pt.

Endereço	**Intervalo de tempo**
abola.pt / www.abola.pt	2000 – 2007
abola.pt:80	2008 - 2022

Dado que o Arquivo.pt permite a pesquisa apenas a partir do ano de 1996 e tem um período de embargo correspondente a um ano [24], definiu-se o intervalo de tempo de pesquisa entre 1996 e 2021. Foi também definido que apenas deveriam ser retornadas páginas HTML, um máximo de 2000 resultados, o conjunto de campos {title, tstamp, originalURL, linkToOriginalFile, linkToArchive} a incluir em cada elemento da resposta e o valor "false" para o parâmetro "prettyPrint".

3.3 Metodologia e resultados

As etapas que descrevem a metodologia geral adotada para o processo de recolha dos dados do Arquivo.pt (Figura 3.1) são:

- Utilização da *Arquivo.pt API*, através da realização de pedidos HTTP para cada *query* de pesquisa. As *queries* de pesquisa foram repartidas por termo do conjunto de termos de pesquisa e por jornal do conjunto de jornais de notícias a utilizar. A repartição das *queries* por termos de pesquisa é obrigatória, tendo em conta os objetivos pretendidos e a documentação da API. No entanto, é possível colocar todos os jornais de notícias numa mesma *query*. Porém, após a realização de algumas

[14] https://www.abola.pt/
[15] https://www.infopedia.pt/apoio/artigos/$a-bola

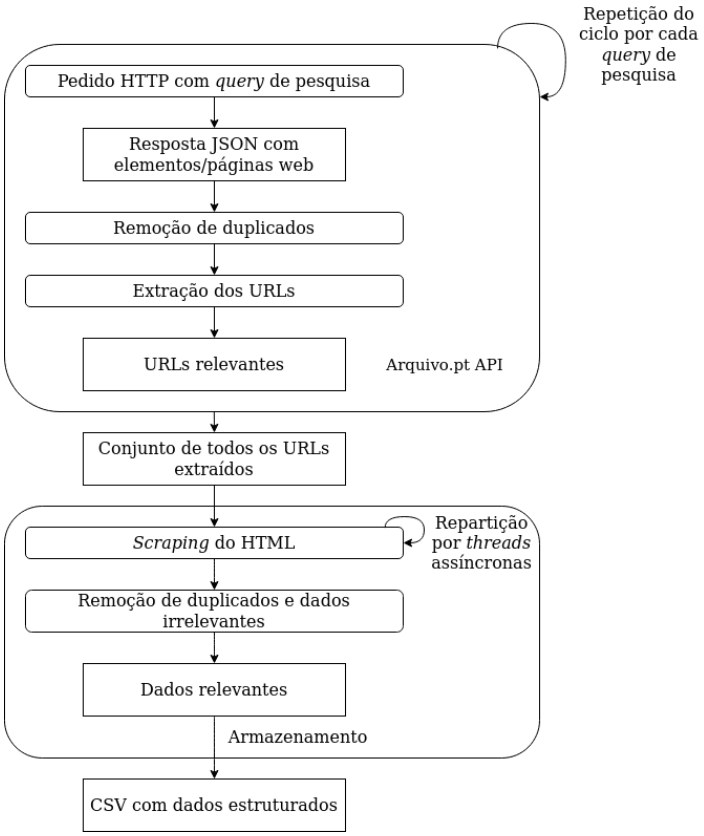


Figura 3.1: Diagrama da metodologia da recolha de dados do Arquivo.pt.

tentativas de pesquisa para perceber o funcionamento da mesma, foi verificado que, por vezes, os resultados apresentavam algumas falhas e diferenças no número de elementos retornados. Assim, preferiu-se realizar a divisão das *queries* também por jornal. Para cada *query* de pesquisa, repetiu-se o fluxo desde a realização do pedido até à extração dos URLs das páginas retornadas. No total, foram realizados 680 ciclos, ou seja utilizadas 680 *queries* de pesquisa;

- Dentro de cada ciclo, foi realizado um processo de remoção de artigos duplicados. Esta remoção foi realizada através da comparação dos campos "linkToArchive", visto terem sido retornadas muitas páginas com o mesmo URL;

- Após a extração de todos os URLs relevantes, procedeu-se ao processo de web *scraping* do HTML de cada página. O número total de páginas a analisar foi de 8235. O processo de web *scraping* foi realizado recorrendo à biblioteca *newspaper* [80], visto que a mesma conseguiu realizar o processo, de forma rápida e eficaz, para todas as páginas retornadas, mesmo as mais antigas. Tendo em conta que o número de páginas a processar é grande, decidiu-se utilizar *multithreading* na tarefa de web *scraping*, ou seja, dividir o processamento das páginas por várias *threads*. Cada *thread* é independente e responsável por realizar o *scraping* de um dos URLs, de forma assíncrona. Além disso, foram aqui também removidos todos os artigos que

não continham no seu título ou conteúdo pelo menos um dos termos do conjunto de termos de pesquisa (apesar de terem sido retornados pela API, nem todos possuíam os termos situados no próprio artigo), e todos os artigos duplicados, sendo a remoção efetuada através da comparação dos conteúdos extraídos;

- Após o processo de web *scraping*, que permitiu obter um total de 1111 artigos e seus metadados, os mesmos foram armazenados num ficheiro *Comma-separated values* (CSV).

3.4 Sumário

Os dados para este projeto, que consistiram no conteúdo e metadados de artigos presentes em jornais de notícias digitais arquivados no Arquivo.pt, foram recolhidos com recurso à *API Arquivo.pt*. Foram recolhidos todos os artigos que possuíam, no seu conteúdo, termos referentes aos transtornos mentais da esquizofrenia e psicose, e foram utilizados nove jornais digitais, sendo eles o Público, o Observador, o Diário de Notícias, o Expresso, o Correio da Manhã, o Jornal de Notícias, o Sábado, o Visão e o A Bola. Tendo em conta que o endereço dos jornais não foi sempre constante ao longo dos anos, foram descobertos e utilizados, na pesquisa, diferentes versões de endereços de cada jornal. Após o processo de obtenção dos URLs das páginas web mais relevantes, foi realizado um processo de filtragem e de web *scraping*, para obter dados estruturados. No final, foram obtidos 1111 dados a utilizar nas subsequentes etapas.

Capítulo 4

Anotação manual dos artigos

Neste capítulo, é descrito o processo realizado para anotar/classificar manualmente os artigos recolhidos, bem como apresentados os resultados finais obtidos. O capítulo termina com um breve sumário dos pontos descritos.

4.1 Metodologia

Tendo em conta que a classificação automática de texto implica a existência de dados já corretamente classificados, para treinar e testar os modelos, foi realizada a anotação manual de todos os artigos. A mesma foi dividida por um conjunto de 15 anotadores humanos, composto por estudantes, investigadores e docentes da Universidade de Aveiro e pertencentes aos ramos da Informática (N=14) e Biotecnologia (N=1). Cada anotador recebeu um conjunto de dados não anotados, sendo que cada elemento era constituído por:

- ID do artigo;
- Título do artigo;
- Conteúdo do artigo, sendo que este conteúdo constituía apenas um excerto do artigo, onde se encontrava pelo menos um dos termos do conjunto dos termos de pesquisa (Secção 3.2). Estes excertos foram gerados de forma automática, ao ser realizada a concatenação da frase onde se encontrava o termo e de algumas frases anteriores e posteriores à mesma (dependendo a quantidade do tamanho do respetivo artigo);
- URL da página com o artigo original arquivado;
- Categoria a que o artigo pertence, sendo este o campo a ser preenchido pelos anotadores.

É de salientar que durante o processo de preparação do conjunto de dados para esta fase, alguns artigos foram descartados por apresentarem problemas estruturais, serem duplicados ou não serem relevantes para o problema. Assim, foi pedido a cada anotador

para classificar o sentido de cada excerto como pertencente a uma das seguintes categorias:

- Estigmatizante — o excerto do artigo é estigmatizante, ou seja, utiliza a doença no sentido metafórico e dentro de um contexto inadequado, para revelar uma ideia que vai além do sentido literal do termo;
- Literal — o excerto do artigo não é estigmatizante, utiliza a doença apenas no seu sentido literal e dentro de um contexto adequado;
- Indefinido — o anotador não consegue decidir a que categoria pertence o excerto do artigo.

Foi disponibilizada uma mesma instrução a cada anotador, referindo em que circunstâncias, baseadas nas apresentadas em estudos anteriores, cada uma das categorias deve ser atribuída. Cada artigo foi classificado por pelo menos dois anotadores diferentes. Após todos os artigos terem sido classificados pelo menos duas vezes, prosseguiu-se à comparação das categorias atribuídas, sendo que foram aprovadas todas as anotações de artigos que possuíam ambas as categorias atribuídas iguais. Nos casos em que o artigo possuía duas categorias distintas, o mesmo passou por uma terceira etapa de anotação, em que uma terceira pessoa (que não foi responsável por classificar o artigo em nenhuma etapa anterior) decidiu a categoria final. Nos casos em que a terceira pessoa não conseguiu decidir a categoria, ou a categoria do artigo não conseguiu o consenso de dois anotadores nas sucessivas etapas de classificação, o artigo foi descartado. A metodologia descrita pode ser visualizada na Figura 4.1.

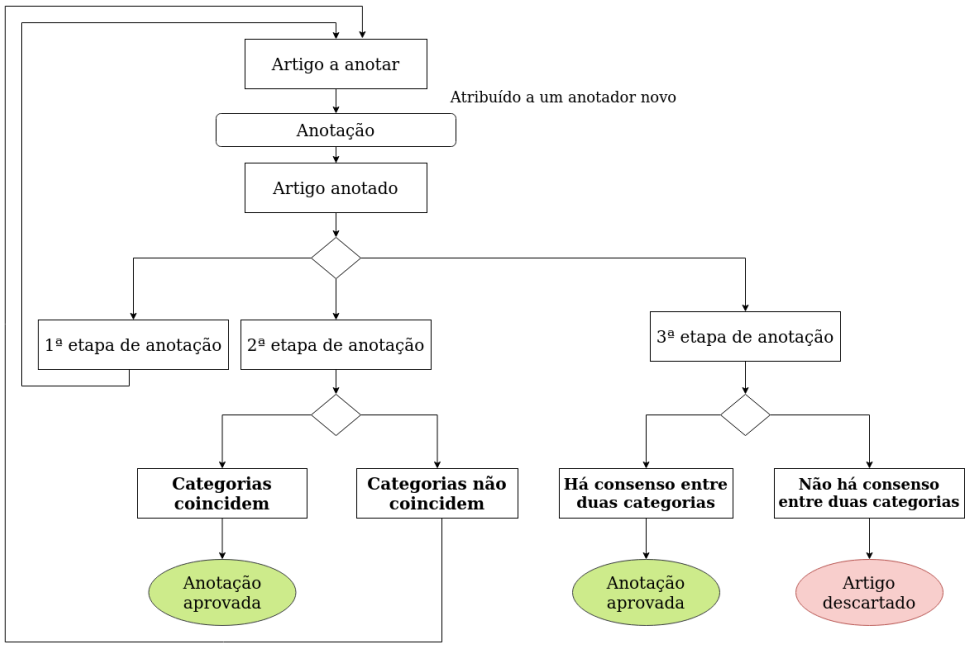


Figura 4.1: Metodologia adotada na etapa de anotação manual dos artigos.

4.2 Resultados

No final, foram obtidos 978 artigos manualmente anotados com as classes ["estigmatizante", "literal"] e, assim, elegíveis a serem utilizados no subsequente processo de classificação automático. Um exemplo de excertos de artigos classificados pode ser visualizado na Tabela 4.1.

Tabela 4.1: Exemplo de excertos de artigos manualmente classificados.

Excerto de artigo	**Sentido**
Os adeptos do Sporting estão a viver uma espécie de **"esquizofrenia"** coletiva. E o que a próxima semana vai trazer, das duas uma, ou a agudiza, ou a resolve.	Estigmatizante
Em Fevereiro e Março de 1996, a Jihad Islâmica e o Hamas levaram a cabo uma série de ataques suicidas, forçando-o ao papel **esquizofrénico** de "guardião de Israel e carcereiro dos palestinianos."	Estigmatizante
Quando Moses Herzog enlouquece, desata a escrever cartas **psicóticas** a toda a gente, incluindo ao presidente dos Estados Unidos.	Estigmatizante
Os internados na clínica psiquiátrica do Hospital Prisional São João de Deus são, em cerca de três quartos dos casos, doentes mentais profundos - **esquizofrénicos**, **psicóticos** maníaco-depressivos - e, nos restantes casos, pessoas com distúrbios de personalidade graves.	Literal
Mais tarde tiveram dois filhos, Isaiah e Eli. Eli, que tem agora 19 anos e está preso, sofre de **esquizofrenia** desde os 14 anos.	Literal
O jovem de 20 anos que foi morto esta terça-feira após sequestrar 37 pessoas num autocarro no Estado brasileiro do Rio de Janeiro estava em "surto **psicótico**", segundo a psicóloga que acompanhou a missão no local.	Literal

4.3 Sumário

A classificação automática de texto implica a existência de dados já classificados, para treinar e testar os modelos computacionais. Neste âmbito, foram manualmente classificados todos os artigos recolhidos e considerados relevantes para o problema. Cada artigo foi classificado por pelo menos dois anotadores humanos, e os artigos que não conseguiram o consenso de pelo menos dois anotadores foram descartados. No final, foi obtido um conjunto de 978 artigos manualmente anotados com as classes ["estigmatizante", "literal"], elegíveis a serem utilizados nos processos de classificação automática de texto.

Capítulo 5

Classificação automática de texto

Neste capítulo, são descritas todas as etapas realizadas no processo de classificação automática dos artigos recolhidos (N=978). Estas etapas consistiram no pré-processamento dos documentos, na implementação dos modelos para a classificação automática do sentido de cada documento, como estigmatizante ou literal, e na deteção automática e classificação dos tópicos presentes, através de um processo de *topic modeling*. É também apresentada a implementação de um *website* como interface de apresentação do projeto e dos principais resultados obtidos pelos modelos de classificação.

O capítulo termina com um breve sumário dos pontos descritos.

5.1 Pré-processamento

Durante a etapa de pré-processamento, foi realizada uma limpeza dos textos e utilizadas técnicas de PLN para preparar os documentos para os subsequentes processos de extração de atributos, classificação e *topic modeling*. Cada documento corresponde ao texto obtido da concatenação do título e conteúdo de cada artigo.

Esta fase é bastante importante e tem o intuito de permitir aos modelos computacionais compreender melhor os textos e gerar resultados mais precisos e consistentes. As técnicas aplicadas foram:

- *Tokenization*: repartição do texto de cada documento numa sequência de termos;
- Conversão para letras minúsculas de todas as palavras do texto;
- Remoção de *stop words*[1], obtidas do NLTK;
- Remoção de todos os URLs, de texto dentro de parêntesis e parêntesis retos, de todos os sinais de pontuação, de todos os termos que contenham números, de todos os termos com tamanho menor que três carateres, de alguns termos irrelevantes sem relação ao texto dos artigos e de todos os pronomes pessoais conectados a verbos.

[1] https://www.nltk.org/howto/portuguese_en.html

Outras técnicas, como *lemmatization* e *stemming*, não foram aplicadas face à escassez de ferramentas precisas para a língua portuguesa e também com o objetivo de não reduzir muito mais o vocabulário dos documentos, o que poderia intervir de forma negativa no desempenho dos modelos computacionais. É de referir que foi também realizada uma simples e introdutória análise exploratória dos dados, que permitiu verificar que o conjunto de dados encontra-se equilibrado, com 52% dos dados pertences à classe positiva ("estigmatizante"), retirando a necessidade de aplicação de quaisquer técnicas para aproximar as quantidades de dados de cada classe.

5.2 Classificação automática do sentido dos artigos

A etapa da classificação automática do sentido dos artigos implicou a ocorrência de duas fases: extração de atributos, onde os documentos foram transformados em representações numéricas vetoriais, e implementação dos modelos de classificação, que foram treinados e testados para a tarefa de classificação automática do sentido presente.

5.2.1 Modelos de representação

Na extração de atributos, foram utilizados quatro modelos diferentes para gerar representações numéricas dos documentos, sendo eles:

- o modelo de *bag-of-words*, que permitiu gerar vetores de dimensão 44522 (número de termos únicos no conjunto dos documentos de treino) para cada documento. Cada vetor representa a contagem, normalizada para o intervalo entre [0, 1], dos termos no respetivo documento. Foi utilizada, para o efeito, a classe "CountVectorizer", do *scikit-learn*;

- o modelo de *Term Frequency - Inverse Document Frequency* (TF-IDF), um modelo similar ao modelo de *bag-of-words* mas que tem adicionalmente em conta a importância dos termos do documento em relação à coleção de todos os documentos. Permitiu gerar vetores de dimensão 44522 para cada documento, tendo sido utilizada, para o efeito, a classe "TfidfVectorizer", do *scikit-learn*;

- o modelo de *word embeddings*, utilizando vetores de 300 dimensões pré-treinados, na língua portuguesa do Brasil e de Portugal, com o algoritmo GloVe e obtidos do repositório NILC-Embeddings[2]. Foi utilizado o tamanho de 300 dimensões visto o menor tamanho disponível (50 dimensões) poder gerar perda de informação e *underfitting*, e o maior tamanho disponível (1000 dimensões) poder gerar modelos demasiado complexos e *overfitting*. Os termos de cada documentos foram mapeados para os correspondentes vetores GloVe, e foi, assim, criada uma matriz de tamanho (51 285, 300), correspondente a (Número de termos únicos em toda a coleção de documen-

[2] http://nilc.icmc.usp.br/embeddings

tos, Dimensão dos vetores GloVe). Estes pesos foram posteriormente utilizados na camada de *embedding* das redes neuronais implementadas;

- o mapeamento dos termos dos textos para as 464 categorias do dicionário *Brazilian Portuguese LIWC 2007 Dictionary*. Foram gerados, para cada documento, vetores de dimensão 464, com cada elemento do vetor a corresponder à contagem de termos (no respetivo documento) pertencentes à respetiva categoria, normalizada para o intervalo entre [0, 1].

5.2.2 Modelos de classificação

O processo de classificação consistiu na implementação dos modelos, no seu treino, utilizando os dados de treino, e na posterior avaliação e comparação dos resultados obtidos, usando os dados de teste. Os dados de treino correspondem a 80% (N=782) dos dados totais (documentos e suas classes) e os dados de teste a 20% (N=196), tendo a repartição sido feita de forma aleatória. Foram utilizados seis algoritmos tradicionais de *machine learning* e quatro algoritmos de *deep learning*.

Os modelos de *machine learning* foram implementados utilizando as bibliotecas *scikit-learn* e *xgboost* [81]. Os principais hiperparâmetros, parâmetros definidos antes do treino dos modelos e responsáveis por definir a sua estrutura e configuração, utilizados em cada um dos modelos foram obtidos através de um processo de otimização usando a estratégia de *5-Fold Cross Validation*. Este processo foi implementado recorrendo à biblioteca *scikit-optimize*, que utiliza um algoritmo de otimização baseado num modelo sequencial (usando processos gaussianos) para encontrar soluções ótimas em menos tempo. Visto que não existem valores oficialmente definidos para este processo (e adequados para qualquer tipo de problema), os intervalos utilizados para o ajuste de cada hiperparâmetro foram definidos tendo em conta a natureza dos mesmos e os valores utilizados em problemas similares encontrados. Os valores finais foram escolhidos considerando as posteriores experiências realizadas com os intervalos encontrados, que permitiram averiguar os conjuntos mais adequados e que iriam requerer um menor consumo de tempo e memória.

Os seis modelos consistiram, assim, nos algoritmos:

- *Logistic Regression* — algoritmo que utiliza uma função logística para modelar a probabilidade das dadas classes. É usado quando os dados são linearmente separáveis e o resultado é de natureza binária. No processo de otimização, foi ajustado o valor do parâmetro "C", utilizando o intervalo de números reais entre [0, 32];

- SVM — algoritmo que tem como objetivo encontrar um hiperplano num espaço de X dimensões (X - número de atributos) que distinga as dadas classes. Foi utilizada a classe "LinearSVC" (*kernel* linear). No processo de otimização, foi ajustado o valor do parâmetro "C", utilizando o intervalo de números reais entre [0, 32];

- *Naive Bayes* — algoritmo probabilístico baseado no Teorema de Bayes e na suposição

de independência condicional dos atributos dada uma classe. Foi utilizada a classe "MultinomialNB", específica para a classificação de atributos discretos. No processo de otimização, foi ajustado o valor do parâmetro "alpha", utilizando o intervalo de números reais entre [0, 10];

- KNN — algoritmo que procura encontrar um número predefinido de amostras de treino mais próximas, em distância, do novo ponto e prever a classe a partir dos mesmos. No processo de otimização, foram ajustados os valores do parâmetro "n_neighbors", utilizando o intervalo de números inteiros entre [1, 31], e do parâmetro "metric", utilizando as categorias ["euclidean", "manhattan"];

- *Random Forest* — algoritmo que ajusta um número de classificadores de árvore de decisão em várias subamostras aleatórias do conjunto de dados de treino e que combina os resultados de cada classificador para determinar a classe final. No processo de otimização, foram ajustados os valores do parâmetro "n_estimators", utilizando o intervalo de números inteiros entre [10, 200], e do parâmetro "max_depth", utilizando o intervalo de números inteiros entre [3, 100];

- *XGBoost* — algoritmo baseado em árvores de decisão mas que utiliza uma estrutura de *gradient boosting*, em que as árvores são construídas sequencialmente de forma a que cada subsequente árvore tenha como objetivo reduzir os erros da árvore anterior. No processo de otimização, foram ajustados os valores do parâmetro "max_depth", utilizando o intervalo de números inteiros entre [3, 10], do parâmetro "tree_method", utilizando as categorias ["exact", "approx", "hist"], do parâmetro "min_child_weight", utilizando o intervalo de números reais entre [1, 6], do parâmetro "learning_rate", utilizando o intervalo de números reais entre [0, 1], e do parâmetro "gamma", utilizando o intervalo de números reais entre [0, 10].

Todos estes algoritmos foram conjugados e treinados com as representações de *bag-of-words*, TF-IDF e a gerada pelo dicionário português do LIWC.

Os modelos de *deep learning* foram implementados utilizando a biblioteca *TensorFlow*, a *API Keras* [82], a *framework PyTorch* [83] e a biblioteca *transformers* [84], e consistiram na construção de três redes neuronais e na implementação de uma rede já pré-treinada. A otimização dos principais hiperparâmetros dos primeiros três modelos foi efetuada recorrendo à biblioteca *Keras Tuner* [85], com o *tuner* da classe "Hyperband", que utiliza o algoritmo de *random search* e procura acelerá-lo através da alocação adaptativa de recursos e paragem antecipada. Além disso, as quatro redes neuronais implementadas possuem a condição de todos os documentos apresentarem o mesmo tamanho, e, por isso, todos os textos sofreram um processo de *padding* ou *truncation* para um número máximo de 512 *tokens*. Foi definido este número por ser o tamanho máximo aceite pelo modelo BERT e também por ser a potência de dois mais próxima do tamanho médio de todos os documentos.

Os quatro modelos consistiram, assim, nos algoritmos:

- *Convolutional Neural Network* (CNN) — tipo de rede neuronal tipicamente utilizado no reconhecimento de imagens mas que também tem sido usado em tarefas de PLN. A CNN (Figura 5.1) foi implementada, sequencialmente, com uma camada de *embedding*, que gera vetores de 300 dimensões usando a matriz de *word embeddings* anteriormente criada a partir dos termos dos documentos e dos vetores *GloVe*, uma camada de *dropout*, duas camadas convolucionais 1D (com função de ativação "relu") seguidas da camada de *max-pooling*, uma camada de *flatten*, uma camada densa (com função de ativação "relu"), uma camada de *dropout* e outra camada densa (com função de ativação "sigmoid", visto o *output* ser de natureza binária).

  No processo de otimização dos hiperparâmetros, foram ajustados os valores do parâmetro "rate", utilizando o conjunto de valores [0.2, 0.3, 0.4, 0.5], do parâmetro "filters", utilizando o intervalo de números inteiros entre [10, 60], do parâmetro "kernel_size", utilizando o intervalo de números inteiros [3, 15], e do parâmetro "units", utilizando o conjunto de valores [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]. A rede foi compilada com o otimizador "Adam" (e função *loss* "binary_crossentropy"), um algoritmo com *learning rate* automaticamente adaptável e mais frequentemente utilizado em problemas de classificação, e treinada usando um tamanho de *batch* de 32 e valor de *epochs* de 10;

- LSTM — tipo de rede neuronal recorrente que consegue manter a longo prazo as informações necessárias ou úteis para previsão. A rede LSTM (Figura 5.2) foi implementada, sequencialmente, com uma camada de *embedding*, que gera vetores de 300 dimensões usando a matriz de *word embeddings* anteriormente criada, uma camada de LSTM e uma camada densa (com função de ativação "sigmoid").

  No processo de otimização, foi ajustado o valor do parâmetro "units", utilizando os valores inteiros pertencentes ao intervalo [32, 512] com um valor de *step* igual 32. A rede foi compilada com o otimizador "Adam" (e função *loss* "binary_crossentropy") e treinada usando um tamanho de *batch* de 32 e valor de *epochs* de 10;

- Bi-LSTM — tipo de rede neuronal recorrente similar à rede LSTM mas que processa a informação nas duas direções (de trás para a frente e de frente para trás). A rede Bi-LSTM (Figura 5.3) foi implementada, sequencialmente, com uma camada de *embedding*, que gera vetores de 300 dimensões usando a matriz de *word embeddings* anteriormente criada, uma camada de LSTM (inserida numa camada bidirecional) e uma camada densa (com função de ativação "sigmoid").

  No processo de otimização, foi ajustado o valor do parâmetro "units", utilizando os valores inteiros pertencentes ao intervalo [32, 512] com um valor de *step* igual 32. A rede foi compilada com o otimizador "Adam" (e função *loss* "binary_crossentropy") e treinada usando um tamanho de *batch* de 32 e valor de *epochs* de 10;

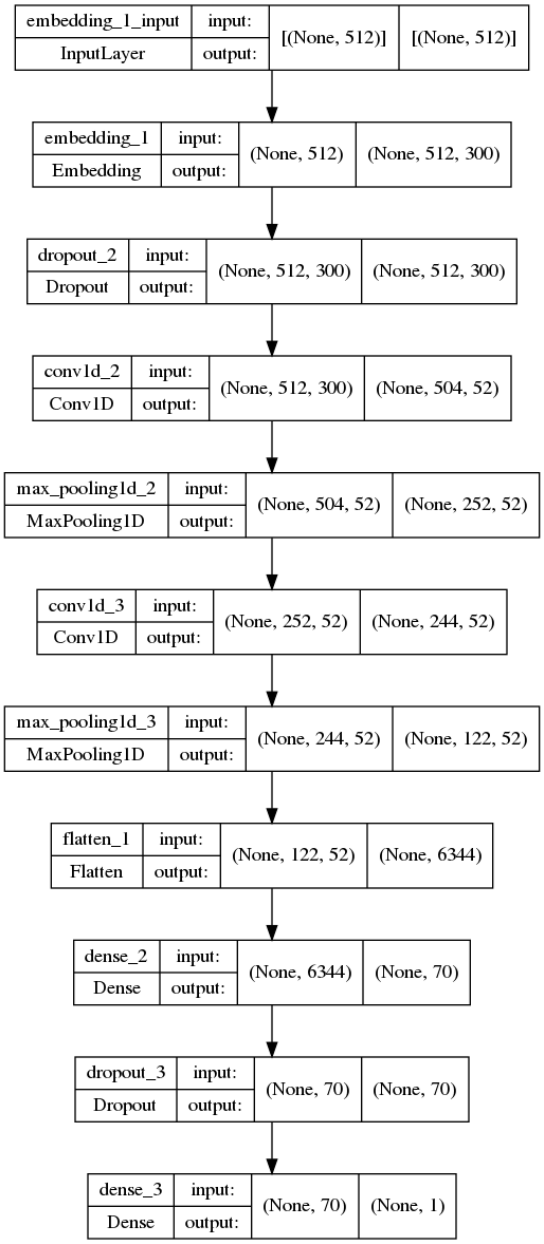


Figura 5.1: Arquitetura da CNN implementada.

- *BERTimbau* — modelo BERT treinado para a língua portuguesa. Foi utilizado o modelo pré-treinado *BERTimbau* no tamanho "Base" (que possui 12 camadas/blocos de *Transformers*, 12 *attention heads* e 110 milhões de parâmetros), retornado através da classe "AutoModelForSequenceClassification", que já possui uma camada de classificação implementada no topo.

  A rede foi compilada com o otimizador "AdamW", uma variante melhorada do otimizador "Adam", e treinada usando um tamanho de *batch* de 8 e valor de *epochs* de 4.

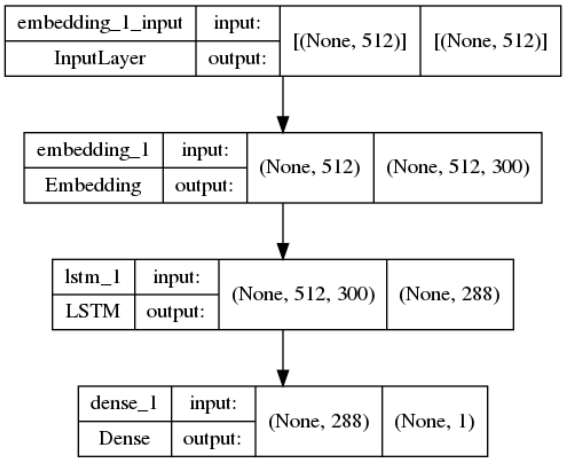


Figura 5.2: Arquitetura da rede LSTM implementada.

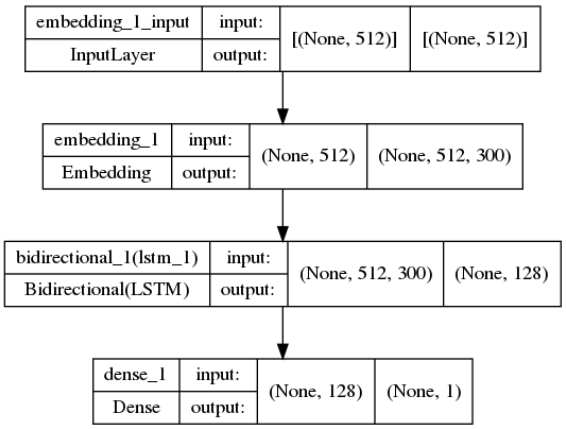


Figura 5.3: Arquitetura da rede Bi-LSTM implementada.

5.3 *Topic modeling*

A deteção automática de tópicos foi realizada usando os algoritmos de LDA e *top2vec*. O algoritmo LDA foi implementado através da construção de uma matriz documento--termos, construída com os 978 documentos, e utilizando a classe "LDAMulticore" da biblioteca *Gensim*. Permitiu obter os termos que mais descrevem e contribuem para cada um dos tópicos (tendo sido predefinidos sete tópicos a descobrir), bem como os documentos onde cada tópico predomina.

O algoritmo *top2vec* foi treinado nos 978 documentos (usando apenas os documentos e ignorando as suas classes) e permitiu obter um conjunto de 50 termos descritivos dos tópicos descobertos (sem predefinição do número de tópicos a descobrir), pontuações da sua similaridade ao tópico e os documentos semanticamente mais similares a cada tópico. Este algoritmo cria um *joint embedding* dos vetores que representam os documentos e dos vetores que representam as palavras contidas nos mesmos, colocando-os no mesmo espaço

vetorial para, de seguida, encontrar *clusters* densos de documentos (tópicos) e identificar quais as palavras que atraíram esses documentos para os respetivos *clusters*, sendo então essas palavras as descritivas do tópico. Para criar o *joint embedding* dos vetores foi utilizado o algoritmo *default* de *doc2vec*.

Como os modelos de *topic modeling* não permitem obter uma classificação geral dos tópicos detetados, e sim os termos mais relevantes para a descrição dos mesmos, a designação geral do tópico foi atribuída manualmente, através da análise da temática presente no conjunto dos termos mais descritivos. É também de referir que este processo permitiu obter resultados bastante satisfatórios e sem necessidade de intervenções complexas, ao contrário do também referido anteriormente processo de PLN de análise de sentimento, onde as ferramentas mencionadas (Secção 2.3.2) e testadas (numa fase inicial exploratória) mostraram-se bastante imprecisas e pouco eficazes para os textos portugueses utilizados e para o tema do presente trabalho.

5.4 Interface de apresentação

A fase final do projeto consistiu na exploração e análise de todos os resultados obtidos dos processos de classificação e *topic modeling*, através de técnicas de visualização implementadas com recurso às bibliotecas *matplotlib* [86] e *seaborne* [87]. Além disso, tendo em conta que a submissão para o concurso "Prémio Arquivo.pt 2022" iria beneficiar do uso de elementos visuais, foi também criado um *website*[3] usando a biblioteca *React*. O seu principal objetivo é apresentar, não só aos júris do concurso mas também aos utilizados comuns, o projeto realizado e os principais resultados obtidos de uma forma mais simples, interativa e intuitiva. De modo a demonstrar a eficácia da utilização de IA para as tarefas de análise de texto, os resultados apresentados no *website* são os referentes ao modelo de *top2vec* (que demonstrou melhores resultados), para a deteção de tópicos, e ao modelo de classificação automática implementado que demonstrou melhor exatidão e precisão, para a classificação do sentido dos artigos. É também de referir que no *website* são apresentados resultados de métricas de avaliação ligeiramente diferentes dos atuais, visto os modelos terem sido posteriormente melhorados, e que não é apresentado o uso do algoritmo *XGBoost*, visto que, durante a submissão para o concurso, este algoritmo ainda não tinha sido implementado.

O *website* está dividido em seis secções, sendo elas:

- Início (Figura 5.4) — página de apresentação;
- Sobre o projeto — página com descrição e arquitetura globais do projeto;
- Exemplo (Figura 5.5) — página com um exemplo interativo onde o utilizador pode experimentar classificar o sentido dos artigos, sendo-lhe disponibilizada a classificação atribuída pelo modelo de classificação desenvolvido mais exato;

[3] https://alina-yanchuk02.github.io/estigma/

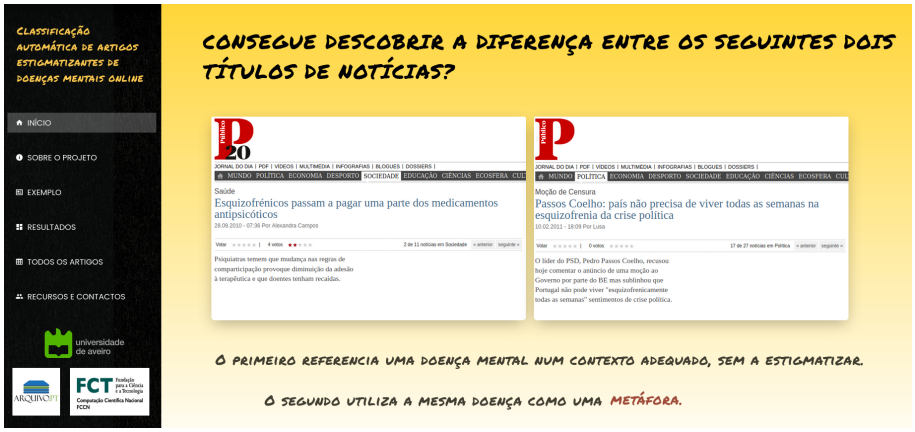


Figura 5.4: Secção "Início" do *website*, com a página inicial.

Figura 5.5: Secção "Exemplo" do *website*, com um exemplo interativo de classificação.

- Resultados (Figura 5.6) — página ao estilo *dashboard* com os principais resultados obtidos da classificação automática (usando os resultados do modelo de classificação mais exato) e do *topic modeling*;
- Todos os artigos (Figura 5.7) — página com uma tabela onde estão agrupados todos os artigos por tópico automaticamente detetado, por ano de arquivamento no Arquivo.pt, por jornal a que pertencem e por classificação automaticamente atribuída ao seu sentido, podendo também ser efetuadas pesquisas por título;
- Recursos e Contactos — página com os principais recursos públicos do projeto (incluindo o conjunto de dados anotados criado) e contactos dos autores/orientadores.

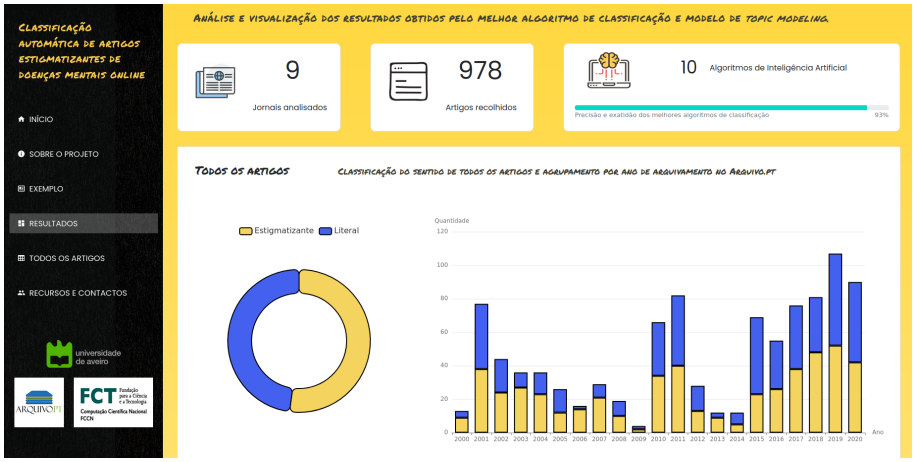


Figura 5.6: Secção "Resultados" do *website*, com a *dashboard* dos principais resultados obtidos.

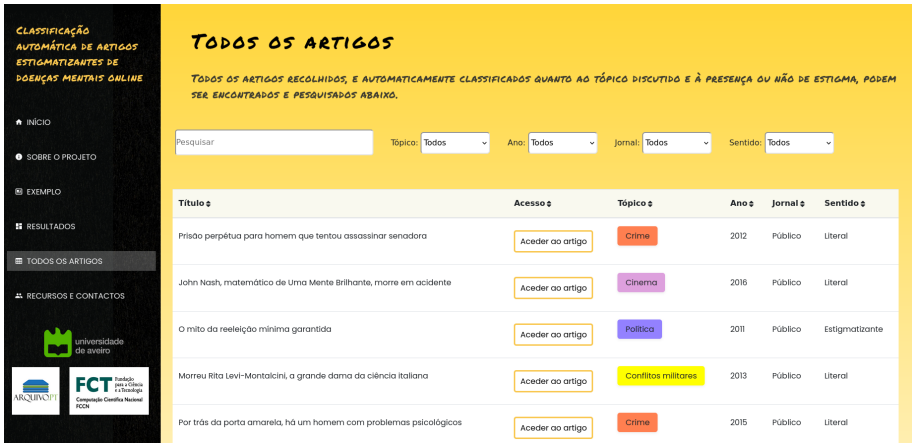


Figura 5.7: Secção "Todos os artigos" do *website*, com a tabela dos artigos agrupados.

5.5 Sumário

A classificação automática dos artigos constituiu as etapas de pré-processamento dos documentos, da implementação dos modelos de representação e de classificação do sentido dos mesmos, e da deteção automática de tópicos presentes através de *topic modeling*. Na etapa de pré-processamento, foram utilizadas várias técnicas de PLN para limpar os documentos e os preparar para os modelos computacionais. Os modelos de representação utilizados foram os modelos de *bag-of-words*, TF-IDF, *word embeddings* e o de mapeamento dos termos dos documentos para as categorias do dicionário *Brazilian Portuguese LIWC 2007 Dictionary*. Os modelos de classificação utilizados foram seis de *machine learning*, sendo eles os modelos de *Logistic Regression*, SVM, *Naive Bayes*, KNN, *Random Forest* e *XGBoost*, e quatro de *deep learning*, sendo eles os modelos de CNN, LSTM, Bi-LSTM e *BERTimbau*. A deteção de tópicos foi realizada utilizando os algoritmos de LDA e *top2vec*,

tendo ambos permitido obter os termos mais descritivos de cada tópico descoberto e os documentos onde cada tópico predomina.

Além disso, foi construído um *website* com o objetivo de apresentar este projeto, e principais resultados obtidos, de uma forma mais simples e intuitiva. Esta interface serviu também para submeter ao concurso "Prémio Arquivo.pt 2022" um elemento visual do projeto, de modo a demonstrar aos utilizadores os benefícios da utilização da IA nas tarefas de análise de textos escritos por humanos.

Capítulo 6

Resultados e Avaliação

Neste capítulo, são apresentados e discutidos os resultados obtidos dos processos de classificação manual e automática do sentido dos artigos e de deteção automática de tópicos presentes, bem como apresentado o processo de divulgação dos mesmos. São também apresentadas métricas de avaliação para avaliar e comparar o desempenho dos modelos implementados. O capítulo termina com um breve sumário dos pontos descritos.

De um modo geral e mais prático, este projeto permitiu obter:

- um conjunto de 978 artigos de jornais portugueses *online*, que fazem referência aos transtornos mentais da esquizofrenia e psicose, manualmente anotados como detentores de um sentido estigmatizante ou literal;
- um conjunto de dez modelos de *machine learning* e *deep learning* que realizam a classificação automaticamente;
- um conjunto de dez tópicos extraídos automaticamente dos artigos, sendo todos estes resultados de acesso público[1].

6.1 Resultados da classificação manual

Quanto aos resultados da anotação manual dos artigos, foi verificado que 52% dos artigos (N=509) possuem um sentido estigmatizante e 48% (N=469) um sentido literal. Esta quantidade de artigos estigmatizantes é um valor bastante significativo e demonstra que os jornais de notícias portugueses *online* retratam, em mais de metade dos casos, os transtornos da esquizofrenia e psicose em contextos fora da temática da saúde mental.

O agrupamento destes resultados por ano de arquivamento no Arquivo.pt pode ser visualizado na Figura 6.1. Pode-se verificar que o maior número de artigos recolhidos foi no ano de 2019, onde foi também verificada a maior quantidade de artigos estigmatizantes. Nos últimos anos (região entre 2015 e 2020) verifica-se um número relativamente maior de artigos que referem transtornos mentais e, em simultâneo, dos que os estigmatizam. Os

[1] https://github.com/alina-yanchuk02/news_classification_stigma

anos que obtiveram a maior diferença entre número de artigos com sentido estigmatizante e com sentido literal foram os de 2003 e 2018, enquanto que o ano que obteve a maior diferença entre número de artigos com sentido literal e com sentido estigmatizante foi o de 2015. É possível também observar que não foram obtidos, do repositório Arquivo.pt, artigos para os anos de 1996, 1997, 1998, 1999 e 2021.

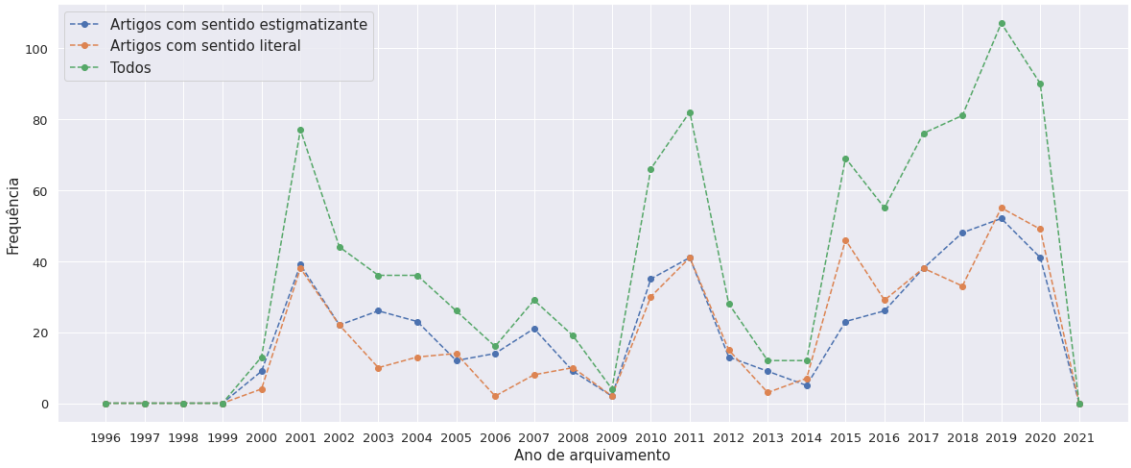


Figura 6.1: Agrupamento dos artigos, manualmente classificados, por ano de arquivamento no Arquivo.pt.

O agrupamento da quantidade de artigos com sentido estigmatizante e literal por jornal de notícias pode ser visualizado na Tabela 6.1. Pode-se observar que foi encontrado conteúdo estigmatizante em todos os jornais em estudo. O jornal Público é o que apresenta maior quantidade de artigos recolhidos, e também a maior quantidade de artigos estigmatizantes. Os jornais que possuem a maior diferença entre o número de artigos com sentido estigmatizante e com sentido literal são o jornal Público e o jornal Expresso, com mais 20 artigos estigmatizantes do que não estigmatizantes. Por outro lado, o jornal Correio da Manhã é o que possui maior diferença entre o número de artigos com sentido literal e com sentido estigmatizante, com mais 16 artigos não estigmatizantes do que estigmatizantes.

6.2 Resultados da classificação automática e avaliação dos modelos

Quanto aos resultados obtidos pelos modelos de classificação automática desenvolvidos, foram utilizadas quatro métricas tipicamente usadas para avaliar modelos de classificação, sendo elas a exatidão, a precisão, o *recall* e o *F1*. Os valores registados podem ser visualizados e comparados na Tabela 6.2. Estão presentes, na mesma, todas as combinações de algoritmo de classificação e modelo de representação dos atributos implementadas.

A maioria dos modelos apresenta bons resultados, com dez em 22 modelos (45%) a apresentar exatidão acima dos 90%, o que significa que mais de 90% dos artigos foram, por eles, corretamente classificados. Destacam-se, no topo, os algoritmos de *Naive Bayes*

Tabela 6.1: Agrupamento dos artigos, manualmente classificados, por jornal de notícias.

Jornal de notícias	**Estigmatizante (N=509)**	**Literal (N=469)**
Público	147 (28.9%)	127 (27.1%)
Observador	113 (22.2%)	114 (24.3%)
Diário de Notícias	50 (9.8%)	39 (8.3%)
Expresso	118 (23.2%)	98 (20.9%)
Correio da Manhã	15 (2.9%)	31 (6.6%)
Jornal de Notícias	30 (5.9%)	31 (6.6%)
Sábado	8 (1.6%)	1 (0.2%)
Visão	16 (3.1%)	23 (4.9%)
A Bola	12 (2.4%)	5 (1.1%)

Tabela 6.2: Valores das métricas de avaliação para cada combinação de modelo de classificação e representação dos atributos implementada.

Modelo de classificação	**Modelo de representação**	**Exatidão (%)**	**Precisão**	***Recall***	***F1***
Logistic Regression	*Bag-of-words*	91.84	0.92	0.92	0.92
	TF-IDF	93.37	0.93	0.94	0.94
	LIWC	77.55	0.80	0.76	0.78
SVM	*Bag-of-words*	90.31	0.92	0.90	0.91
	TF-IDF	90.82	0.93	0.90	0.91
	LIWC	82.14	0.81	0.84	0.83
Naive Bayes	*Bag-of-words*	88.78	0.84	0.97	0.90
	TF-IDF	**94.39**	0.93	0.97	**0.95**
	LIWC	52.04	0.52	**1.00**	0.69
KNN	*Bag-of-words*	65.82	0.89	0.40	0.54
	TF-IDF	88.78	0.90	0.88	0.90
	LIWC	69.90	0.70	0.76	0.72
Random Forest	*Bag-of-words*	90.82	0.88	0.96	0.92
	TF-IDF	92.86	0.90	0.97	0.93
	LIWC	76.53	0.75	0.82	0.89
XGBoost	*Bag-of-words*	84.69	0.83	0.88	0.86
	TF-IDF	84.69	0.85	0.86	0.85
	LIWC	78.57	0.79	0.79	0.79
CNN	*Word embeddings*	90.82	**0.97**	0.85	0.91
LSTM	*Word embeddings*	91.33	0.89	0.95	0.92
Bi-LSTM	*Word embeddings*	89.29	0.93	0.86	0.90
BERTimbau	*Tokenizer* do *BERTimbau*	91.33	0.93	0.91	0.92

(94.39%) e *Logistic Regression* (93.37%), ambos conjugados com a representação de TF-IDF. Na representação de atributos, destacam-se os modelos de TF-IDF, *bag-of-words* e *word embeddings*, sendo que o modelo do LIWC português é o que apresenta os piores resultados, com diferenças bastante significativas. No campo do *deep learning*, os modelos com melhor exatidão foram o *BERTimbau* (91.33%) e o LSTM (91.33%). A rede CNN obteve também o melhor valor de precisão (97%) de entre todos, o que significa que de todos os artigos que a mesma classificou como estigmatizantes, 97% eram realmente estigmatizantes. Quanto ao *recall*, que calcula quantos dos artigos estigmatizantes foram classificados como tais, o modelo que apresentou melhor resultado (100%) foi o *Naive Bayes* conjugado com LIWC, apesar da baixa exatidão (52.04%) e precisão (52%).

Apesar dos bons resultados que os algoritmos de *deep learning* conjugados com *word embeddings* conseguiram obter, estes não conseguiram superar os resultados obtidos por alguns dos tradicionais algoritmos de *machine learning* conjugados com representações de atributos mais simples, como é o caso dos dois modelos que obtiveram a melhor exatidão. Isto pode sugerir a necessidade de experimentar com outros algoritmos de *word embeddings*, ou dimensões dos seus vetores, ou gerar novos, através do treino com maior volume de textos portugueses. Além disso, outras configurações de redes podem ser exploradas ou até mesmo os seus resultados podem ser conjugados através de *ensemble learning*. Apesar de o *deep learning* ser, atualmente, a abordagem preferida em muitos problemas de PLN, o conjunto de dados utilizados neste projeto é bastante pequeno e não permite averiguar, de forma totalmente certa, a sua eficácia. Muitas vezes, as redes neuronais apresentam grandes variações quando utilizadas em novos conjuntos de dados e com um tamanho mais significativo, o que apela para a necessidade do treino e teste dos modelos em mais dados.

Quanto aos baixos resultados apresentados pelo modelos onde se utilizou a representação criada com o dicionário LIWC português, estes podem indicar a necessidade de estudar melhor as diferentes categorias presentes e realizar a sua filtragem, de modo a utilizar apenas as mais relevantes. No entanto, isto iria também implicar um estudo e/ou conhecimento mais aprofundado nas áreas dos processos psicológicos e linguísticos.

6.3 Resultados do *topic modeling*

O algoritmo *top2vec* gerou resultados significativamente melhores que o algoritmo LDA. O algoritmo LDA obteve uma coerência, retornada pela classe "CoherenceModel", de apenas 0.26, e palavras mais relevantes para cada tópico muito pouco descritivas. Para a obtenção de resultados mais precisos, mostrou-se evidente ser necessário um pré-processamento dos textos mais complexo e que poderia implicar várias rondas de implementação e avaliação. Por outro lado, o algoritmo *top2vec* apresentou resultados muito precisos e sem aparente necessidade de mais limpeza ou qualquer outro pré-processamento dos documentos. Este algoritmo detetou automaticamente os tópicos presentes sem a necessidade de predefinir o número de tópicos a descobrir, e os termos descritivos mostraram-se bastante coerentes. São, assim, aqui referidos, bem como utilizados como resultados no

website implementado, os resultados obtidos pelo algoritmo *top2vec*.

Foram automaticamente detetados dez tópicos, cada um definido por um conjunto de 50 termos mais descritivos do mesmo. Na Tabela 6.3 podem ser visualizados os 20 termos mais descritivos retornados, ordenados por ordem decrescente de similaridade semântica ao tópico, a classificação geral atribuída, manualmente, a cada tópico, e o número de artigos pertencentes aos mesmos. É possível verificar que as doenças mentais são, essencialmente, retratadas nas temáticas da Saúde e quando associadas a ações criminais, e que a maior percentagem de artigos estigmatizantes, relativamente ao total de artigos nesse tópico, está presente nos tópicos da Economia (97%) e da Política (96%).

6.4 Concurso e divulgação do trabalho

Tal como referido, os resultados do presente trabalho foram submetidos ao concurso "Prémio Arquivo.pt 2022", em maio de 2022. Os conteúdos submetidos consistiram no código aberto do projeto, no *website* implementado, num relatório técnico com as informações mais importantes e relevantes, e no conjunto dos dados manualmente anotados. Como resultado, o trabalho foi premiado e obteve o segundo lugar no concurso[2]. Além disso, um artigo científico produzido com base na dissertação foi submetido, em junho de 2022, ao *workshop "DOING: Intelligent Data – from data to knowledge"* a realizar-se na conferência *ADBIS 2022 (26th European Conference on Advances in Databases and Information Systems)*, tendo sido aceite e a ser publicado na série *Communications in Computer and Information Science* da *Springer*.

6.5 Sumário

Do processo de classificação manual do sentido dos artigos, foi verificado que 52% dos 978 artigos recolhidos eram estigmatizantes, sendo que ocorre a utilização de expressões referentes aos transtornos mentais da esquizofrenia e psicose num sentido metafórico em todos os jornais de notícias em estudo. Os modelos de classificação desenvolvidos também conseguiram realizar a deteção de estigma, com 45% a apresentar exatidão acima dos 90%. Destacam-se os modelos de *Naive Bayes* e *Logistic Regression*, conjugados com a representação de TF-IDF, que conseguiram obter um valor de exatidão de 94.39% e 93.37% respetivamente. No campo do *deep learning*, destacam-se a rede neuronal LSTM, conjugada com *word embeddings*, e o modelo pré-treinado *BERTimbau*, que apresentam uma exatidão de 91.33% e sugerem, adicionalmente, resultados promissores para configurações mais complexas das redes e uso de um maior conjunto de dados de treino e teste. Foram também automaticamente detetados dez tópicos diferentes, sendo que foi observado que os maiores níveis de estigmatização ocorrem nas temáticas da Economia e Política.

Os resultados obtidos foram divulgados e premiados no concurso "Prémio Arquivo.pt 2022" e aceites, como um artigo científico, na conferência *ADBIS 2022*.

[2] https://sobre.arquivo.pt/pt/conheca-os-vencedores-do-premio-arquivo-pt-2022/

Tabela 6.3: 20 termos mais descritivos retornados para cada tópico, classificação geral atribuída e número de artigos total e estigmatizantes (com percentagem em relação ao total de artigos nesse tópico).

Termos descritivos	**Tópico**	**Total Artigos**	**Artigos Estigmatizantes**
[doencas,estudo,doenca,medicamentos,ansiedade, sintomas,doentes,estudos,saude,tratamentos, tratamento,mental,mentais,pacientes,investigadores, existem,efeitos,utilizacao,genetica,comportamentos]	Saúde	232	13 (6%)
[homicidio,prisao,policia,crime,encontrado,crimes, inimputavel,tribunal,matou,sofre,psiquiatrica, vitima,arguido,psiquiatrico,internamento, internado,matar,acusacao,acusado,condenado]	Crime	158	13 (8%)
[filme,comedia,realizador,personagens,cinema,perso-nagem,actores,filmes,original,estreia,actor,hollywood, serie,americano,cena,peca,titulo,oscar,temporada,obra]	Cinema	112	61 (54%)
[europeia,austeridade,divida,euro,mercados, orcamental,uniao,europeu,economica,economia, economico,investimento,financas,europeias, bruxelas,defice,crescimento,crise,europa,financeira]	Economia	92	89 (97%)
[eua,russia,militar,armas,washington, forcas,americanos,nortemericana,guerra,militares, ataque,seguranca,conflito,putin,nortemericano, norte,ataques,estrangeiros,estados,presidente]	Conflitos militares	85	79 (93%)
[partido,governo,psd,parlamentar,mocao,parlamento, politico,socialista,cds,lider,coelho,partidos,oposicao, pcp,passos,socialistas,politica,socrates,eleitoral,voto]	Política	80	77 (96%)
[livros,escritor,literatura,escritores,escrita,escrever, romance,obra,escreve,livro,textos,escrevi,ler,escrito, personagens,nasceu,leitores,autor,paginas,irmao]	Literatura	70	44 (63%)
[banda,album,disco,pop,rock,musica,cancoes, musical,concerto,concertos,cancao,musico,bandas, palco,cantar,letras,editora,som,the,estreia]	Música	70	63 (90%)
[desporto,futebol,jogo,lideranca,dirigentes,jogos, valores,vitoria,clube,rio,liga,equipa,exercicio,etica, paixao,proprios,porto,gestao,caracteristicas,estilo]	Desporto	41	37 (90%)
[magistrados,justica,judicial,tribunais,ministerio, penal,processos,criminal,juizes,advogados,elina, fraga,corrupcao,gestao,cidadao,direito,politicos, codigo,judiciaria,segredo]	Justiça	38	34 (89%)

Capítulo 7

Conclusão

No projeto desta dissertação, foi realizada a recolha e classificação manual do sentido (estigmatizante ou literal) de artigos detentores de referências aos transtornos mentais da esquizofrenia e psicose, de jornais de notícias portugueses presentes na Internet e arquivados no repositório *online* Arquivo.pt, bem como a exploração de técnicas de IA para a realização automática das tarefas de classificação e *topic modeling*. Foram implementados dez diferentes modelos de *machine learning* e *deep learning* para a tarefa da classificação, e foi utilizado o algoritmo *top2vec* para a deteção de tópicos, tendo sido obtidos resultados bastante precisos e que permitiram averiguar a vantagem da utilização de modelos computacionais para a análise de textos na língua portuguesa. Tendo em conta que a análise de grandes volumes de texto implica um grande esforço humano e tempo, a metodologia automatizada retratada surge como uma vantagem ao permitir obter resultados muito próximos dos reais e explorar o problema de forma mais eficiente. Os modelos desenvolvidos podem também ser estendidos e ajustados a outros problemas e domínios similares, bem como melhorados para responder às presentes e outras questões do atual contexto.

Foi também obtido um conjunto de 978 artigos manualmente anotados com os sentidos "estigmatizante" e "literal", que permitem explorar como a saúde mental, e mais especificamente os transtornos da esquizofrenia e psicose, são retratados nos meios de comunicação social portuguesa. Foi, assim, confirmada a existência de estigma na imprensa *online* portuguesa, que se observou em 52% de todos os artigos e em todos os jornais de notícias em estudo e, em especial, nas temáticas da Economia e Política. Adicionalmente, o conjunto de dados desenvolvido é abertamente disponibilizado, podendo também ser utilizado para outros fins de investigação.

Este é o primeiro trabalho que explora a classificação de textos portugueses que contêm referências metafóricas através de IA, sendo que as grandes conclusões retiradas são que a maioria dos tradicionais algoritmos de *machine learning* permitem obter bons resultados, que o uso de redes neuronais sugere ser bastante promissor e que a tarefa de deteção de tópicos pode ser parcialmente automatizada usando *topic modeling* com o algoritmo *top2vec*, visto o mesmo ter permitido obter resultados bastante acertados. No entanto, o campo do PLN portuguesa encontra-se ainda muito pouco explorado, o que se revela

também na escassa quantidade de ferramentas e modelos treinados para o português de Portugal (como se verificou no caso da análise de sentimento), existindo também abordagens mais complexas que devem ser, futuramente, consideradas. Além disso, revela-se também uma necessidade de aprofundar as questões éticas relacionadas com a classificação automática do sentido dos artigos, bem como realizar um estudo mais aprofundado dos aspetos psicológicos e linguísticos subjacentes.

Referências

[1] *Prémios Arquivo.pt – sobre.arquivo.pt*. Fundação para a Ciência e Tecnologia. URL: https://sobre.arquivo.pt/pt/colabore/premios-arquivo-pt/ (acedido em 30/10/2021) (ver p. 2).

[2] *What Is Mental Illness?* American Psychiatric Association. URL: https://www.psychiatry.org/patients-families/what-is-mental-illness (acedido em 23/10/2021) (ver p. 5).

[3] «Depression and Other Common Mental Disorders: Global Health Estimates». Em: (2017). Geneva: World Health Organization. Licence: CC BY-NC-SA 3.0 IGO. (Ver p. 5).

[4] Infopédia. *estigma | Definição ou significado de estigma no Dicionário Infopédia da Língua Portuguesa*. Infopédia - Dicionários Porto Editora. URL: https://www.infopedia.pt/dicionarios/lingua-portuguesa/estigma (acedido em 28/10/2021) (ver p. 6).

[5] «Policies and practices for mental health in Europe: meeting the challenges». Em: World Health Organization Regional Office for Europe. 2008. ISBN: 978-92-890-4279-6. (Ver p. 6).

[6] Patrick W. Corrigan e Amy C. Watson. «Understanding the Impact of Stigma on People with Mental Illness». Em: *World Psychiatry* 1.1 (fev. de 2002), pp. 16–20. ISSN: 1723-8617. pmid: 16946807. (Ver p. 6).

[7] Ordem dos Psicólogos Portugueses. «Desenvolvimento Sustentável e Sustentabilidade dos Cuidados de Saúde Primários». Em: Lisboa, Portugal, 2021. ISBN: 978-989-53170-2-8. (Ver p. 6).

[8] F. Lopes, R. Duarte, G. B. Migliori e R. Araújo. «Tuberculosis in the News: How Do Portuguese Media Cover TB». Em: *Pulmonology* 24.2 (1 de mar. de 2018), pp. 69–72. ISSN: 2531-0437. DOI: 10.1016/j.pulmoe.2018.02.004. (Ver p. 6).

[9] David Dias Neto, Maria João Figueiras, Sónia Campos e Patrícia Tavares. «Impact of Economic Crisis on the Social Representation of Mental Health: Analysis of a Decade of Newspaper Coverage». Em: *The International Journal of Social Psychiatry* 63.8 (dez. de 2017), pp. 736–743. ISSN: 1741-2854. DOI: 10.1177/0020764017737102. pmid: 29058959. (Ver p. 6).

[10] Programa Nacional para a Saúde Mental. «Programa Nacional para a Saúde Mental 2017». Em: (2017). Ed. por Direção-Geral da Saúde.
URL: https://www.dgs.pt/em-destaque/relatorio-do-programa-nacional-para-a-saude-mental-2017.aspx (acedido em 25/10/2021) (ver p. 6).

[11] Sociedade Portuguesa de Psiquiatria e Saúde Mental. «Guia Essencial para Jornalistas». Em: (set. de 2016).
URL: https://www.sppsm.org/informemente/guia-essencial-para-jornalistas/ (acedido em 25/10/2021) (ver pp. 6, 7).

[12] A. H. Crisp, M. G. Gelder, S. Rix, H. I. Meltzer e O. J. Rowlands. «Stigmatisation of People with Mental Illnesses». Em: *The British Journal of Psychiatry: The Journal of Mental Science* 177 (jul. de 2000), pp. 4–7. ISSN: 0007-1250. DOI: 10.1192/bjp.177.1.4. pmid: 10945080. (Ver pp. 6, 7).

[13] Alastair Benbow. «Mental Illness, Stigma, and the Media». Em: *The Journal of Clinical Psychiatry* 68 Suppl 2 (2007), pp. 31–35. ISSN: 1555-2101. pmid: 17288505. (Ver pp. 6, 7).

[14] Nele Cornelia Goepfert, Steffen Conrad von Heydendorff, Harald Dreßing e Josef Bailer. «Effects of Stigmatizing Media Coverage on Stigma Measures, Self-Esteem, and Affectivity in Persons with Depression – an Experimental Controlled Trial». Em: *BMC Psychiatry* 19.1 (7 de mai. de 2019), p. 138. ISSN: 1471-244X. DOI: 10.1186/s12888-019-2123-6. (Ver p. 7).

[15] Enric Aragonès, Judit López-Muntaner, Santiago Ceruelo e Josep Basora. «Reinforcing Stigmatization: Coverage of Mental Illness in Spanish Newspapers». Em: *Journal of Health Communication* 19.11 (2014), pp. 1248–1258. ISSN: 1087-0415. DOI: 10.1080/10810730.2013.872726. pmid: 24708534. (Ver p. 8).

[16] Christina Athanasopoulou e Maritta Välimäki. «'Schizophrenia' as a Metaphor in Greek Newspaper Websites». Em: Studies in Health Technology and Informatics. Vol. 202. 2014, pp. 275–278. ISBN: 978-1-61499-422-0. DOI: 10.3233/978-1-61499-423-7-275. (Ver p. 8).

[17] Arun Chopra e Gillian Doody. «Schizophrenia, an Illness and a metaphor: Analysis of the use of the term 'schizophrenia' in the UK national newspapers». Em: *Journal of the Royal Society of Medicine* 100 (out. de 2007), pp. 423–6. DOI: 10.1258/jrsm.100.9.423. (Ver p. 8).

[18] Kenneth Duckworth, John H. Halpern, Russell K. Schutt e Christopher Gillespie. «Use of Schizophrenia as a Metaphor in US Newspapers». Em: *Psychiatric Services (Washington, D.C.)* 54.10 (out. de 2003), pp. 1402–1404. ISSN: 1075-2730. DOI: 10.1176/appi.ps.54.10.1402. pmid: 14557528. (Ver p. 8).

[19] Francisco Bevilacqua Guarniero, Ruth Helena Bellinghini e Wagner Farid Gattaz. «The Schizophrenia Stigma and Mass Media: A Search for News Published by Wide Circulation Media in Brazil». Em: *International Review of Psychiatry (Abingdon, England)* 29.3 (jun. de 2017), pp. 241–247. ISSN: 1369-1627. DOI: 10.1080/09540261.2017.1285976. pmid: 28492091. (Ver p. 8).

[20] Lucie Nawková, Alexander Nawka, Tereza Adámková, Tea Vukušić Rukavina, Petra Holcnerová, Martina Rojnić Kuzman, Nikolina Jovanović, Ognjen Brborović, Bibiána Bednárová, Svetlana Zuchová, Michal Miovský e Jiří Raboch. «The Picture of Mental Health/Illness in the Printed Media in Three Central European Countries». Em: *Journal of Health Communication* 17.1 (2012), pp. 22–40. ISSN: 1087-0415. DOI: 10.1080/10810730.2011.571341. pmid: 21707410. (Ver p. 8).

[21] *Os media e a saúde mental - Análise de conteúdo de notícias publicadas por meios de comunicação social portugueses*. Sociedade Portuguesa de Psiquiatria e Saúde Mental. Jun. de 2016.
URL: https://www.sppsm.org/informemente/apresentacao/ (acedido em 28/10/2021) (ver p. 8).

[22] Nuno Rodrigues-Silva, Telma Falcão de Almeida, Filipa Araújo, Andrew Molodynski, Ângela Venâncio e Jorge Bouça. «Use of the Word Schizophrenia in Portuguese Newspapers». Em: *Journal of Mental Health (Abingdon, England)* 26.5 (out. de 2017), pp. 426–430. ISSN: 1360-0567. DOI: 10.1080/09638237.2016.1207231. pmid: 27841067. (Ver p. 8).

[23] *Informações gerais – sobre.arquivo.pt*. Fundação para a Ciência e Tecnologia.
URL: https://sobre.arquivo.pt/pt/ajuda/o-que-e-o-arquivo-pt/ (acedido em 29/10/2021) (ver p. 9).

[24] *Recolha de conteúdos – sobre.arquivo.pt*. Fundação para a Ciência e Tecnologia.
URL: https://sobre.arquivo.pt/pt/ajuda/recolha-e-arquivo-de-conteudos/ (acedido em 29/10/2021) (ver pp. 9, 10, 29).

[25] *Arquivo.pt em números – sobre.arquivo.pt*. Fundação para a Ciência e Tecnologia.
URL: https://sobre.arquivo.pt/pt/imprensa/o-arquivo-pt-em-numeros/ (acedido em 29/10/2021) (ver p. 9).

[26] Trupti Udapure, Ravindra Kale e Rajesh Dharmik. «Study of Web Crawler and Its Different Types». Em: *IOSR Journal of Computer Engineering* 16 (1 de jan. de 2014), pp. 01–05. DOI: 10.9790/0661-16160105. (Ver p. 10).

[27] Manish Kumar, Rajesh Bhatia e Dhavleesh Rattan. «A Survey of Web Crawlers for Information Retrieval». Em: *WIREs Data Mining and Knowledge Discovery* 7.6 (2017), e1218. ISSN: 1942-4795. DOI: 10.1002/widm.1218. (Ver p. 11).

[28] Paulo Jorge Pereira Martins, Leandro José Abreu Dias Costa e José Carlos Ramalho. «Major Minors - Ontological Representation of Minorities by Newspapers». Em: *10th Symposium on Languages, Applications and Technologies (SLATE 2021)*. Ed. por Ricardo Queirós, Mário Pinto, Alberto Simões, Filipe Portela e Maria João Pereira. Vol. 94. Open Access Series in Informatics (OASIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021, 3:1–3:13. ISBN: 978-3-95977-202-0. DOI: 10.4230/OASIcs.SLATE.2021.3. (Ver p. 11).

[29] Emma E. McGinty, Elizabeth M. Stone, Alene Kennedy-Hendricks e Colleen L. Barry. «Stigmatizing Language in News Media Coverage of the Opioid Epidemic: Implications for Public Health». Em: *Preventive Medicine* 124 (jul. de 2019), pp. 110–114. ISSN: 1096-0260. DOI: 10.1016/j.ypmed.2019.03.018. pmid: 31122614. (Ver p. 12).

[30] Ang Li, Dongdong Jiao e Tingshao Zhu. «Detecting Depression Stigma on Social Media: A Linguistic Analysis». Em: *Journal of Affective Disorders* 232 (mai. de 2018), pp. 358–362. ISSN: 1573-2517. DOI: 10.1016/j.jad.2018.02.087. pmid: 29510353. (Ver p. 12).

[31] Ang Li, Xiaoxiao Huang, Dongdong Jiao, Bridianne O'Dea, Tingshao Zhu e Helen Christensen. «An Analysis of Stigma and Suicide Literacy in Responses to Suicides Broadcast on Social Media». Em: *Asia-Pacific Psychiatry: Official Journal of the Pacific Rim College of Psychiatrists* 10.1 (mar. de 2018). ISSN: 1758-5872. DOI: 10.1111/appy.12314. pmid: 29383880. (Ver p. 12).

[32] Charu C. Aggarwal e ChengXiang Zhai. «A Survey of Text Classification Algorithms». Em: *Mining Text Data*. Ed. por Charu C. Aggarwal e ChengXiang Zhai. Boston, MA: Springer US, 2012, pp. 163–222. ISBN: 978-1-4614-3223-4. DOI: 10.1007/978-1-4614-3223-4_6. (Ver p. 13).

[33] M.M. Mirończuk e J. Protasiewicz. «A Recent Overview of the State-of-the-Art Elements of Text Classification». Em: *Expert Systems with Applications* 106 (2018), pp. 36–54. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2018.03.058. (Ver pp. 13, 14).

[34] Tomas Mikolov, Kai Chen, G.s Corrado e Jeffrey Dean. «Efficient Estimation of Word Representations in Vector Space». Em: *Proceedings of Workshop at ICLR* 2013 (jan. de 2013). (Ver p. 13).

[35] Jeffrey Pennington, Richard Socher e Christopher Manning. «GloVe: Global Vectors for Word Representation». Em: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, out. de 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. (Ver p. 13).

[36] Shitao Zhang. «Sentiment Classification of News Text Data Using Intelligent Model». Em: *Frontiers in Psychology* 12 (2021), p. 4398. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.758967. (Ver pp. 13, 18).

[37] Jeelani Ahmed e Muqeem Ahmed. «Online news classification using machine learning techniques». Em: *IIUM Engineering Journal* 22.2 (2 4 de jul. de 2021), pp. 210–225. ISSN: 2289-7860. DOI: 10.31436/iiumej.v22i2.1662. (Ver p. 13).

[38] K.R. Reddy e S. Chaudhary. «Research Challenges in Text Mining and Empirical Research Directions». Em: *Indian Journal of Computer Science and Engineering* 12.3 (2021), pp. 752–764. ISSN: 0976-5166. DOI: 10.21817/indjcse/2021/v12i3/211203222. (Ver p. 13).

[39] Bi-Min Hsu. «Comparison of Supervised Classification Models on Textual Data». Em: *Mathematics* 8.5 (2020). ISSN: 2227-7390. DOI: 10.3390/math8050851. (Ver p. 13).

[40] Swapna Gottipati, Mark CHONG, Andrew Wei Kiat Lim e Benny Haryanto Kawidiredjo. «Exploring Media Portrayals of People with Mental Disorders Using NLP». Em: *Proceedings of the 14th International Conference on Health Informatics HEALTHINF 2021: Part of BIOSTEC 2021, Virtual, February 11-13* 5 (1 de fev. de 2021), pp. 708–715. DOI: 10.5220/0010380007080715. (Ver p. 13).

[41] M.N. Asim, M.U.G. Khan, M.I. Malik, A. Dengel e S. Ahmed. «A Robust Hybrid Approach for Textual Document Classification». Em: Proceedings of the International Conference on Document Analysis and Recognition, ICDAR. 2019, pp. 1390–1396. ISBN: 978-1-72812-861-0. DOI: 10.1109/ICDAR.2019.00224. (Ver p. 13).

[42] Ge Gao, Eunsol Choi, Yejin Choi e Luke Zettlemoyer. «Neural Metaphor Detection in Context». Em: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2018. Brussels, Belgium: Association for Computational Linguistics, out. de 2018, pp. 607–613. DOI: 10.18653/v1/D18-1060. (Ver p. 13).

[43] Xin Chen, Zhen Hai, Suge Wang, Deyu Li, Chao Wang e Huanbo Luan. «Metaphor Identification: A Contextual Inconsistency Based Neural Sequence Labeling Approach». Em: *Neurocomputing* 428 (7 de mar. de 2021), pp. 268–279. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2020.12.010. (Ver p. 13).

[44] Hossein Hassani, Christina Beneki, Stephan Unger, Maedeh Taj Mazinani e Mohammad Reza Yeganegi. «Text Mining in Big Data Analytics». Em: *Big Data and Cognitive Computing* 4.1 (1 mar. de 2020), p. 1. DOI: 10.3390/bdcc4010001. (Ver pp. 15, 19).

[45] Tiago de Melo e Carlos M. S. Figueiredo. «Comparing News Articles and Tweets About COVID-19 in Brazil: Sentiment Analysis and Topic Modeling Approach.» Em: *JMIR Public Health Surveill* (2021), e24585–e24585. DOI: 10.2196/24585. (Ver p. 15).

[46] Douglas Nunes de Oliveira e Luiz Henrique de Campos Merschmann. «Joint Evaluation of Preprocessing Tasks with Classifiers for Sentiment Analysis in Brazilian Portuguese Language». Em: *Multimedia Tools and Applications* (3 de fev. de 2021). DOI: 10.1007/s11042-020-10323-8. (Ver pp. 15, 18).

[47] Andreas Chandra. *Recent Works in Topic Modeling*. Data Folks Indonesia. 9 de out. de 2020.
URL: https://medium.com/data-folks-indonesia/recent-works-in-topic-modeling-56c38da8dfc4 (acedido em 12/11/2021) (ver p. 16).

[48] Miha Pavlinek e Vili Podgorelec. «Text Classification Method Based on Self-Training and LDA Topic Models». Em: *Expert Systems with Applications* 80 (1 de set. de 2017), pp. 83–93. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2017.03.020. (Ver p. 16).

[49] Dimitar Angelov. «Top2Vec: Distributed Representations of Topics». Em: *ArXiv* abs/2008.09470 (2020). (Ver p. 16).

[50] Wenyue Zhang, Yang Li e Suge Wang. «Learning Document Representation via Topic-Enhanced LSTM Model». Em: *Knowledge-Based Systems* 174 (15 de jun. de 2019), pp. 194–204. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2019.03.007. (Ver p. 16).

[51] Hyeju Jang, Yohan Jo, Qinlan Shen, Michael Miller, Seungwhan Moon e Carolyn Rosé. «Metaphor Detection with Topic Transition, Emotion and Cognition in Context». Em: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2016. Berlin, Germany: Association for Computational Linguistics, ago. de 2016, pp. 216–225. DOI: 10.18653/v1/P16-1021. (Ver p. 17).

[52] Yohan Jo e Alice Oh. «Aspect and sentiment unification model for online review analysis». Em: fev. de 2011, pp. 815–824. DOI: 10.1145/1935826.1935932. (Ver p. 17).

[53] Saif Mohammad, Ekaterina Shutova e Peter Turney. «Metaphor as a Medium for Emotion: An Empirical Study». Em: *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. Berlin, Germany: Association for Computational Linguistics, ago. de 2016, pp. 23–33. DOI: 10.18653/v1/S16-2003. (Ver p. 17).

[54] Walaa Medhat, Ahmed Hassan e Hoda Korashy. «Sentiment Analysis Algorithms and Applications: A Survey». Em: *Ain Shams Engineering Journal* 5.4 (1 de dez. de 2014), pp. 1093–1113. ISSN: 2090-4479. DOI: 10.1016/j.asej.2014.04.011. (Ver p. 17).

[55] Ricardo Martins, José João Almeida, Pedro Henriques e Paulo Novais. «A Sentiment Analysis Approach to Improve Authorship Identification». Em: *Expert Systems* 38.5 (2021), e12469. ISSN: 1468-0394. DOI: 10.1111/exsy.12469. (Ver p. 17).

[56] Denilson Alves Pereira. «A Survey of Sentiment Analysis in the Portuguese Language». Em: *Artificial Intelligence Review* 54.2 (1 de fev. de 2021), pp. 1087–1115. ISSN: 1573-7462. DOI: 10.1007/s10462-020-09870-1. (Ver pp. 17, 20).

[57] James Pennebaker, Martha Francis e Roger Booth. «Linguistic Inquiry and Word Count (LIWC)». Em: (1 de jan. de 1999). (Ver p. 18).

[58] Aytug Onan e Mansur Togoclu. «Satire Identification in Turkish News Articles Based on Ensemble of Classifiers». Em: *TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES* 28 (28 de mar. de 2020), pp. 1086–1106. DOI: 10.3906/elk-1907-11. (Ver p. 18).

[59] Matheus Araújo, Adriano Pereira e Fabrício Benevenuto. «A Comparative Study of Machine Translation for Multilingual Sentence-Level Sentiment Analysis». Em: *Information Sciences* 512 (1 de fev. de 2020), pp. 1078–1102. ISSN: 0020-0255. DOI: 10.1016/j.ins.2019.10.031. (Ver p. 18).

[60] Cátia Tavares, Ricardo Ribeiro e Fernando Batista. «Sentiment Analysis of Portuguese Economic News». Em: *10th Symposium on Languages, Applications and Technologies (SLATE 2021)*. Ed. por Ricardo Queirós, Mário Pinto, Alberto Simões, Filipe Portela e Maria João Pereira. Vol. 94. Open Access Series in Informatics (OASIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021, 17:1–17:13. ISBN: 978-3-95977-202-0. DOI: 10.4230/OASIcs.SLATE.2021.17. (Ver p. 18).

[61] Prem Melville, Wojciech Gryc e Richard D. Lawrence. «Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification». Em: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '09. New York, NY, USA: Association for Computing Machinery, 28 de jun. de 2009, pp. 1275–1284. ISBN: 978-1-60558-495-9. DOI: 10.1145/1557019.1557156. (Ver p. 18).

[62] Nalini Chintalapudi, Gopi Battineni e Francesco Amenta. «Sentimental Analysis of COVID-19 Tweets Using Deep Learning Models». Em: *Infectious Disease Reports* 13.2 (1 de abr. de 2021), pp. 329–339. ISSN: 2036-7430. DOI: 10.3390/idr13020032. pmid: 33916139. (Ver p. 18).

[63] Jacob Devlin, Ming-Wei Chang, Kenton Lee e Kristina Toutanova. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». Em: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. NAACL-HLT 2019. Minneapolis, Minnesota: Association for Computational Linguistics, jun. de 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. (Ver p. 18).

[64] Fábio Souza, Rodrigo Nogueira e Roberto Lotufo. «BERTimbau: pretrained BERT models for Brazilian Portuguese». Em: *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*. 2020. (Ver p. 19).

[65] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer e Veselin Stoyanov. «RoBERTa: A Robustly Optimized BERT Pretraining Approach». Em: *CoRR* abs/1907.11692 (2019). arXiv: 1907.11692. (Ver p. 19).

[66] P. Ghasiya e K. Okamura. «Investigating COVID-19 News across Four Nations: A Topic Modeling and Sentiment Analysis Approach». Em: *IEEE Access* 9 (2021), pp. 36645–36656. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3062875. (Ver p. 19).

[67] Ronen Feldman e James Sanger. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge; New York: Cambridge University Press, 2007. ISBN: 978-0-521-83657-9. (Ver p. 19).

[68] ODSC-Open Data Science. *10 Notable Frameworks for NLP*. Medium. 18 de mar. de 2020.
URL: https://medium.com/@ODSC/10-notable-frameworks-for-nlp-ce8c4196bfd6 (acedido em 03/11/2021) (ver p. 20).

[69] Edward Loper Bird Steven e Ewan Klein. *Natural Language Processing with Python*. Website: https://www.nltk.org/. O'Reilly Media Inc, 2009. ISBN: 978-0-596-51649-9. (Ver p. 20).

[70] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot e E. Duchesnay. «Scikit-learn: Machine Learning in Python». Em: *Journal of Machine Learning Research* 12 (2011). Website: https://scikit-learn.org, pp. 2825–2830. (Ver p. 20).

[71] Matthew Honnibal, Ines Montani, Sofie Van Landeghem e Adriane Boyd. «spaCy: Industrial-strength Natural Language Processing in Python». Em: (2020). Website: https://spacy.io/. DOI: 10.5281/zenodo.1212303. (Ver p. 20).

[72] Erick Fonseca. *Nlpnet: Neural Networks for NLP Tasks*. Versão 1.2.4.
URL: http://nilc.icmc.usp.br/nlpnet (acedido em 01/12/2021) (ver p. 20).

[73] Radim Řehůřek e Petr Sojka. «Software Framework for Topic Modelling with Large Corpora». Em: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Website: https://radimrehurek.com/gensim/. Valletta, Malta: ELRA, mai. de 2010, pp. 45–50. (Ver p. 20).

[74] Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu e Xiaoqiang Zheng. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. Website: http://tensorflow.org/. 2016. DOI: 10.48550/ARXIV.1603.04467. (Ver p. 21).

[75] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton e Christopher D. Manning. «Stanza: A Python Natural Language Processing Toolkit for Many Human Languages». Em: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Website: https://stanfordnlp.github.io/stanza/. 2020. (Ver p. 21).

[76] Steven Loria. *Textblob: Simplified Text Processing*.
URL: https://textblob.readthedocs.io/en/dev/ (acedido em 01/12/2021) (ver p. 21).

[77] *Pesquisa de páginas – sobre.arquivo.pt*. Fundação para a Ciência e Tecnologia.
URL: https://sobre.arquivo.pt/pt/ajuda/pesquisa/ (acedido em 02/01/2022) (ver p. 23).

[78] Entidade Reguladora para a Comunicação Social. «Públicos e Consumos de Média - O consumo de notícias e as plataformas digitais em portugal e em mais dez países». Em: (2014).
URL: www.erc.pt/pt/estudos-e-publicacoes/consum%20os-de-media/estudo-publicos-e-consumos-de-media (ver p. 25).

[79] Diogo Silva da Cunha. «Transformações da presença dos jornais portugueses na web (1996-2016): Correio da Manhã, Diário de Notícias, Expresso e Público. Relatório final de um estudo de caso do projecto "Investiga XXI"». Em: (31 de jul. de 2017). Relatório (121 páginas).
URL: https://sobre.arquivo.pt/pt/publicacoes/relatorios-tecnicos/ (ver pp. 25–27).

[80] Lucas Ou-Yang. *Newspaper3k: Article scraping curation*. 1 de fev. de 2022.
URL: https://newspaper.readthedocs.io/en/latest/ (ver p. 30).

[81] Tianqi Chen e Carlos Guestrin. «XGBoost: A Scalable Tree Boosting System». Em: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785.
URL: http://doi.acm.org/10.1145/2939672.2939785 (ver p. 39).

[82] François Chollet et al. *Keras*. https://keras.io. 2015. (Ver p. 40).

[83] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai e Soumith Chintala. «PyTorch: An Imperative Style, High-Performance Deep Learning Library». Em: *Advances in Neural Information Processing Systems 32*. Ed. por H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox e R. Garnett. Curran Associates, Inc., 2019, pp. 8024–8035.
URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf (ver p. 40).

[84] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest e Alexander M. Rush. «Transformers: State-of-the-Art Natural Language Processing». Em: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, out. de 2020, pp. 38–45. (Ver p. 40).

[85] Tom O'Malley, Elie Bursztein, James Long, François Chollet, Haifeng Jin, Luca Invernizzi et al. *KerasTuner*. https://github.com/keras-team/keras-tuner. 2019. (Ver p. 40).

[86] J. D. Hunter. «Matplotlib: A 2D graphics environment». Em: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55. (Ver p. 44).

[87] Michael L. Waskom. «seaborn: statistical data visualization». Em: *Journal of Open Source Software* 6.60 (2021), p. 3021. DOI: 10.21105/joss.03021.
URL: https://doi.org/10.21105/joss.03021 (ver p. 44).

Apêndices

Apêndice A

Arquitetura geral do projeto

Neste Apêndice, é apresentada a arquitetura geral do projeto desta dissertação (Figura A.1), que descreve, visualmente, as principais etapas realizadas no âmbito do projeto.

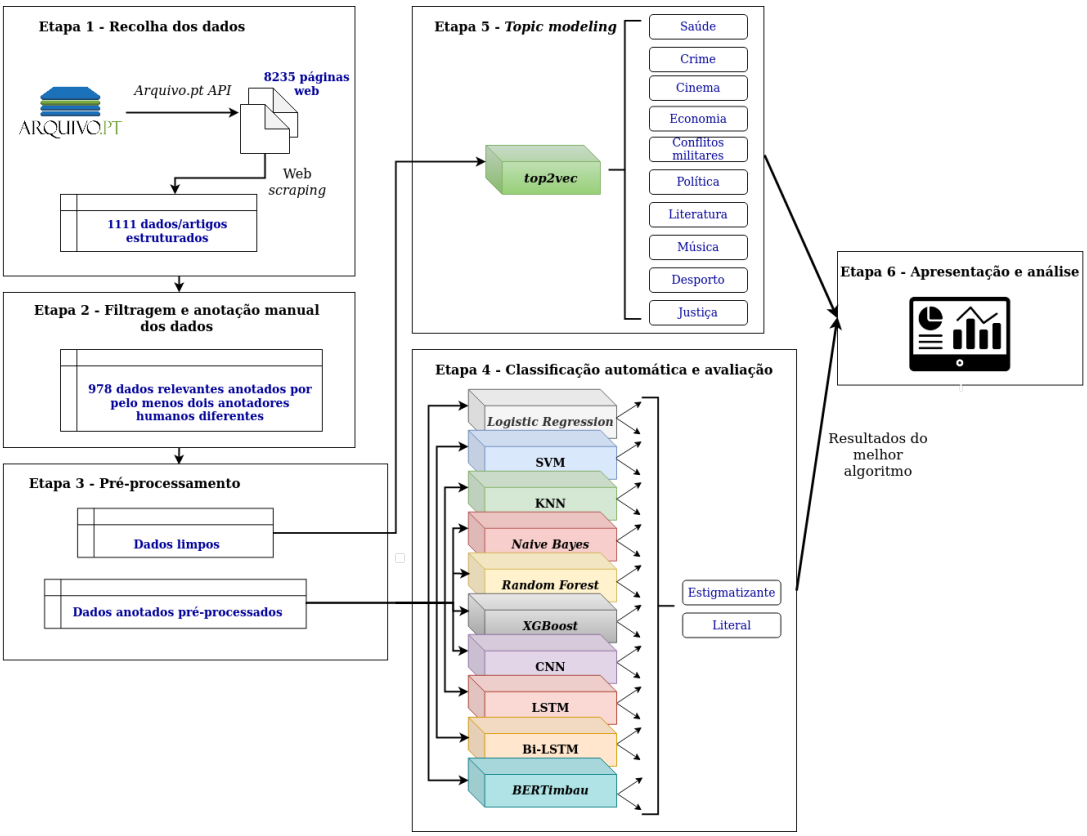


Figura A.1: Arquitetura geral do projeto.

Apêndice B

Descrição da *Arquivo.pt API*

Neste Apêndice, é apresentada a documentação relevante relativa à *Arquivo.pt API*, nomeadamente os parâmetros de pesquisa que podem ser utilizados (e os que foram, no presente projeto, utilizados), os campos de respostas que podem ser retornados e um exemplo de resposta retornado.

B.1 Parâmetros de pesquisa

A API possui um único *endpoint* que é "https://arquivo.pt/textsearch". Na Tabela B.1, podem ser visualizados todos os parâmetros de pesquisa que podem ser utilizados para as pesquisas de texto. Na tabela B.2, podem ser visualizados os valores de todos os parâmetros das *queries* de pesquisa utilizadas no processo de recolha dos dados para este projeto. Os parâmetros {from, to, type, maxItems, fields, prettyPrint} permaneceram constantes em todas as *queries*, enquanto os parâmetros {q, siteSearch} foram atualizando com um novo valor da lista a cada *query*.

Tabela B.1: Parâmetros de pesquisa efetuada com a *Arquivo.pt API*. Adaptado de https://github.com/arquivo/pwa-technologies/wiki/Arquivo.pt-API.

Parâmetro	Descrição	Exemplo
q	Termos de pesquisa. Operadores de pesquisa avançada são: • “ “ : procurar elementos que contêm a expressão exata entre aspas; • - : excluir elementos que contêm termos a seguir ao travessão.	q = James Davis q = "Antonio Costa" q = Albert -Einstein
from	Data (do arquivamento da página) inicial do intervalo de pesquisa. O formato é YYYYMMDDHHMMSS, sendo possível apenas usar o ano YYYY. *Default*: 1996	from = 19960101000000
to	Data final do intervalo de pesquisa. *Default*: Ano atual - 1	to = 20151022163016
type	Conjunto dos formatos dos elementos (p. ex. *pdf, ps, html, xls, ppt, doc, rtf*).	type = pdf
offset	Posição dos índices onde a pesquisa começa. *Default*: 0	offset = 0
siteSearch	Limitar a pesquisa para um dado conjunto de *websites* (separados por vírgulas).	siteSearch = http://www.publico.pt
collection	Limitar a pesquisa para um dado conjunto de coleções.	collection = EAWP13,EAWP21
maxItems	Número máximo de elementos. *Default*: 50. Máximo: 2000	maxItems = 50
dedupValue	Número máximo de elementos por cada *website* referido no parâmetro “siteSearch”.	dedupValue = 5
dedupField	Parâmetro da resposta onde será realizada a remoção de duplicados (p. ex. site, url).	dedupField = site
fields	Conjunto de campos a incluir em cada elemento da resposta. Campos possíveis: title, originalURL, linkToArchive, tstamp, contentLength, digest, mimeType, linkToScreenshot, date, encoding, linkToNoFrame, linkToOriginalFile, collection, snippet, linkToExtractedText.	fields = title, originalURL, linkToArchive, tstamp
callback	Função de *callback*.	callback = hndlr
prettyPrint	Retorna os resultados com indentações e quebras de linha. Quando tem valor "false", pode conduzir a melhores desempenhos. *Default*: true	prettyPrint = true

Tabela B.2: Valores dos parâmetros utilizados nas *queries* para o processo de recolha de dados.

Parâmetro	**Valor/Valores**
q	"esquizofrenia", "esquizofrénico", "esquizofrenico", "esquizofrénica", "esquizofrenica", "esquizofrénicas", "esquizofrenicas", "esquizofrénicos", "esquizofrenicos", "esquizofrenicamente", "esquizofrenizar", "psicose", "psicótica", "psicotica", "psicóticas", "psicoticas", "psicótico", "psicotico", "psicóticos", "psicoticos"
from	1996
to	2021
type	"html"
siteSearch	"publico.pt", "www.publico.pt", "ultimahora.publico.pt", "jornal.publico.pt", "dossiers.publico.pt", "desporto.publico.pt", "www.publico.clix.pt", "digital.publico.pt", "economia.publico.pt", "m.publico.pt", "blogues.publico.pt", "observador.pt", "www.dn.pt", "dn.sapo.pt", "www.dn.sapo.pt", "expresso.pt", "aeiou.expresso.pt", "expresso.sapo.pt", "www.correiomanha.pt", "www.correiodamanha.pt", "www.cmjornal.xl.pt", "www.cmjornal.pt", "www.jn.pt", "jn.pt", "jn.sapo.pt", "www.sabado.xl.pt:80", "www.sabado.xl.pt", "sabado.pt", "www.sabado.pt", "aeiou.visao.pt", "visao.sapo.pt", "abola.pt", "www.abola.pt", "abola.pt:80"
maxItems	2000
fields	"title,tstamp,originalURL,linkToOriginalFile,linkToArchive"
prettyPrint	"false"

B.2 Campos de resposta

Na Tabela B.3, podem ser visualizados todos os campos de resposta que podem ser retornados. Os primeiros seis campos correspondem à parte descritiva da resposta e todos os restantes campos fazem parte do conjunto que descreve cada elemento, que corresponde a uma página web, retornado. Assim, este conjunto repete-se para cada elemento retornado.

Um exemplo de resposta retornada pode ser visualizado na Figura B.1.

Tabela B.3: Campos de resposta retornada pela *Arquivo.pt API*. Adaptado de https://github.com/arquivo/pwa-technologies/wiki/Arquivo.pt-API.

Campo	Descrição	Exemplo
serviceName	Nome do serviço.	"serviceName": "Arquivo.pt-the Portuguese web-archive"
linkToService	URL para o serviço.	"linkToService": "https://arquivo.pt"
request_parameters	Parâmetros de pesquisa.	"request_parameters": { "q": "Costa"}
next_page	URL para os próximos N elementos (se os elementos retornados ultrapassarem o limite). N = offset + limite	
previous_page	URL para os N elementos anteriores. N = offset - limite	
estimated_nr_results	Número estimado de resultados, sem paginação.	"estimated_nr_results": "8654051"
title	Elemento HTML <title> da versão original.	"title": "Antonio Costa"
originalURL	URL original da versão preservada.	
linkToArchive	URL para a versão preservada no Arquivo.pt.	
tstamp	Data de arquivamento. Formato: YYYYMMDDHHMMSS	"tstamp": "19961013191640"
contentLength	Tamanho (em *bytes*) da versão preservada.	"contentLength": "1023"
digest	Hash da versão preservada. Algoritmo: MD5	"digest": "5e8de36a1d6a7d"
mimeType	Tipo MIME do conjunto de caracteres.	"mimeType": "text"
encoding	Codificação do conteúdo. Pode retornar vazio.	"encoding": "windows-1252"
date	Data da realização do *crawling*. Formato: *epoch*	"date": "0845224210"
linkToScreenshot	URL para a versão preservada no formato de imagem.	
linkToNoFrame	URL para a versão preservada sem as partes laterais do Arquivo.pt.	
linkToOriginalFile	URL para o HTML original da versão preservada.	
linkToExtractedText	URL para o texto extraído da versão preservada. Pode retornar vazio.	
linkToMetadata	URL para os metadados do documento.	
snippet	Bloco HTML que contém correspondência com termos de pesquisa.	"snippet": «em>Antonio Costa
collection	Coleção a que pertence a versão preservada. Pode retornar vazio.	"collection": "AWP3"

```
{
  "serviceName" : "Arquivo.pt - the Portuguese web-archive",
  "linkToService" : "https://arquivo.pt",
  "next_page" : "https://arquivo.pt/textsearch?q=esquizofrenia&maxItems=1&type=html&siteSearch=www.publico.pt&prettyPrint=true&offset=1",
  "estimated_nr_results" : 18970,
  "request_parameters" : {
    "offset" : 0,
    "dedupValue" : 2,
    "type" : [ "html" ],
    "dedupField" : "url",
    "q" : "esquizofrenia",
    "maxItems" : 1,
    "siteSearch" : [ "www.publico.pt" ]
  },
  "response_items" : [ {
    "title" : "Esquizofrenia - PÚBLICO",
    "originalURL" : "http://www.publico.pt/culturaipsilon/noticia/esquizofrenia-1690279",
    "linkToArchive" : "https://arquivo.pt/wayback/20150326210311/http://www.publico.pt/culturaipsilon/noticia/esquizofrenia-1690279",
    "tstamp" : "20150326210311",
    "contentLength" : 179709,
    "digest" : "f6fb88c0d5599f9ecc8284f64ee4c0da",
    "mimeType" : "text/html",
    "encoding" : "UTF-8",
    "date" : "1427403791",
    "linkToScreenshot" : "https://arquivo.pt/screenshot?url=https%3A%2F%2Farquivo.pt%2FnoFrame%2Freplay%2F20150326210311%2Fhttp%3A%2F%2Fwww.publico.pt%2Fculturaipsilon%2Fnoticia%2Fesquizofrenia-1690279",
    "linkToNoFrame" : "https://arquivo.pt/noFrame/replay/20150326210311/http://www.publico.pt/culturaipsilon/noticia/esquizofrenia-1690279",
    "linkToExtractedText" : "https://arquivo.pt/textextracted?m=http%3A%2F%2Fwww.publico.pt%2Fculturaipsilon%2Fnoticia%2Fesquizofrenia-1690279%2F20150326210311",
    "linkToMetadata" : "https://arquivo.pt/textsearch?metadata=http%3A%2F%2Fwww.publico.pt%2Fculturaipsilon%2Fnoticia%2Fesquizofrenia-1690279%2F20150326210311",
    "linkToOriginalFile" : "https://arquivo.pt/noFrame/replay/20150326210311id_/http://www.publico.pt/culturaipsilon/noticia/esquizofrenia-1690279",
    "snippet" : "<em>Esquizofrenia</em> - P&Uacute;BLICO Artigos seguintes Artigos anteriores Multim&eacute;dia V&iacute;deo Cinecartaz: Trailer As Vozes T&oacute;picos Cinema <em>Esquizofrenia</em>
Twitter Partilhar no Google+ 0 Cr&iacute;tica <em>Esquizofrenia</em> Por Jorge Mourinha 26/03/2015 - 06:20 (actualizado &agrave;s 06:20 ) Marjane<span class=\"ellipsis\"> ... </span>",
    "fileName" : "IAH-20150326180149-00010-p22.arquivo.pt",
    "collection" : "FAWP2020150326",
    "offset" : 70362692
  } ]
}
```


Figura B.1: Exemplo de resposta JSON, retornada para a pesquisa com os parâmetros “q=esquizofrenia”, “maxItems=1”, “type=html” e “siteSearch=www.publico.pt”.