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

Alina Yanchuk[1], Alina Trifan[2], Olga Fajarda[1] e José Luís Oliveira[1]

[1] Departamento de Electrónica, Telecomunicações e Informática, Universidade de Aveiro, Aveiro, Portugal
{alinayanchuk,olga.oliveira,jlo}@ua.pt
[2] Instituto de Engenharia Electrónica e Telemática de Aveiro, Universidade de Aveiro, Aveiro, Portugal
alina.trifan@ua.pt

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 em 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 da saúde mental e psiquiatria. 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 Arquivo.pt. Foram implementados nove modelos de *machine learning* e *deep learning* para a realização desta tarefa, sendo que a maioria permitiu obter resultados com exatidão e precisão 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.

Keywords: Classificação de texto · Classificação automática · Inteligência Artificial · Artigos *online* · Jornais de notícias · Processamento de Linguagem Natural · *machine learning* · *deep learning* · *topic modeling*

1 Introdução

A presença de estigma na nossa sociedade é ainda uma realidade 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, quando os termos referentes às doenças mentais são utilizados num sentido figurado/metafórico para descrever entidades ou situações fora do contexto da saúde. 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*[3], a maior base de dados *online* de literatura revista por pares, o número de publicações relativas à temática de classificação automática de texto apresenta um grande aumento nos últimos anos, e principalmente a partir do ano de 2015 (Figura 1). Além disso, cada vez processamento de mais têm sido adotadas abordagens computacionais para a realizar, em contraste com os tradicionais métodos manuais. Os métodos manuais caracterizam-se pela anotação, por humanos, dos textos a classificar, enquanto que os métodos computacionais utilizam Inteligência Artificial. Os subcampos da Inteligência Artificial mais relevantes para este processo são a aprendizagem de máquina ou *machine learning*, o Processamento de Linguagem Natural (PLN) e a mineração de texto.

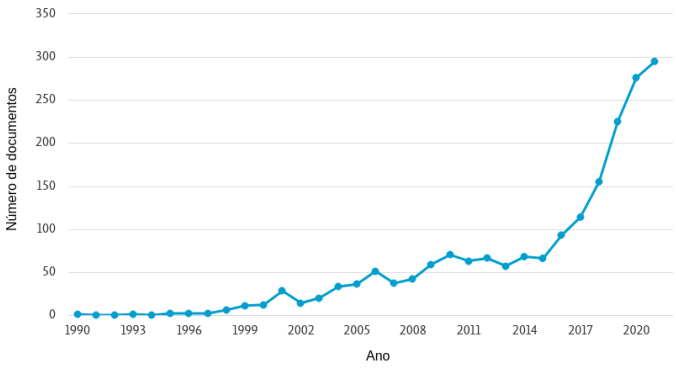

Figura 1. Comparação do número de publicações por ano com os termos "*news*", "*classification*" e "*automatic*", presentes no título, no resumo ou nas palavras-chave. Adaptado do *Scopus*.

Assim, o resultado deste projeto consiste num conjunto de algoritmos de classificação automática de texto, que permitem classificar o sentido dos artigos, presentes em jornais de notícias *online* e detentores de referências aos transtornos

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

mentais da esquizofrenia e psicose, como estigmatizante ou literal. Consequentemente, é também disponibilizado um conjunto de 978 artigos portugueses anotados com as duas classes. Adicionalmente, foi também realizada uma deteção automática de tópicos presentes nos artigos. Todos os artigos foram recolhidos do repositório Arquivo.pt[4], a fonte oficial de dados.

2 Enquadramento

As doenças mentais são condições de saúde diagnosticadas que podem envolver alterações de pensamento, de emoções e de comportamento. As doenças mais graves constituem, em grande parte, a depressão *major*, as perturbações bipolares e a esquizofrenia [1].

Para além dos desafios físicos e psicológicos que estas doenças trazem, os seus portadores sofrem também com o estigma a elas associado. 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 [2]. 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.

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 [3]. 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. 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 [4], 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 [5], lançado pela Sociedade Portuguesa de Psiquiatria e Saúde Mental, que apresenta um manual prático designado por "Guia essencial para jornalistas sobre saúde mental" [5] e cujo objetivo é combater o estigma existente nos meios de comunicação social.

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

[4] https://www.arquivo.pt/
[5] https://www.sppsm.org/informemente

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 entidades de forma negativa, como por exemplo quando a palavra "esquizofrénico" é utilizada para se referir a uma situação ridícula ou contraditória.

Na Europa, um estudo [6] analisou 695 notícias, que apresentavam termos relacionadas 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 [7]. No Reino Unido, esse número constituiu 11% de um total de 600 notícias analisadas [8]. Nos Estados Unidos da América, foram analisados 1740 artigos e a percentagem dos que utilizavam a esquizofrenia como metáfora constituiu 28% [9]. No Brasil, um estudo [10] 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".

Atualmente, apenas dois estudos portugueses neste âmbito foram encontrados. O primeiro [11] 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 [12] 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 consistiram apenas num estudo exploratório, que utilizou a típica abordagem manual de classificação.

Tendo em conta a quantidade massiva de dados existentes em formato digital, o seu processamento e extração de informação manuais exigem 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*. *Machine learning* é um tipo de Inteligência Artificial cujos algoritmos permitem aos computadores extrair conhecimento a partir de um conjunto de dados e aprender a tomar decisões de forma automática e independente. A classificação automática de texto é um problema que pertence à categoria de algoritmos de *machine learning* que usam aprendizagem supervisionada, onde existe um conjunto de dados de treino já classificado, e consiste num processo automático de associar dados textuais a uma dada classe. O processo geral de classificação consiste numa etapa de limpeza e pré-processamento dos textos, numa etapa de extração de atributos relevantes, que consiste na representação de cada texto numa forma numérica, sendo a representação mais comum a vetorial, na etapa de classificação e numa etapa final de avaliação dos modelos perante os resultados obtidos.

No pré-processamento dos textos, são aplicadas técnicas de PLN, sendo as mais comuns a *tokenization*, que é a repartição do texto em *tokens*/termos, a remoção de *stop words*, palavras com alta frequência e sem importância semântica para o texto, e as técnicas de *lemmatization* e *stemming*, que consistem na transformação de palavras derivadas à sua forma raiz, tendo em conta o contexto e ignorando-o, respetivamente.

Quanto à extração de atributos, esta 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, e no modelo *Term Frequency--Inverse Document Frequency* (TF-IDF), que tem também em conta a importância das mesmas no texto. Representações mais complexas passam pelo mapeamento das palavras para certas categorias já estabelecidas. Um exemplo é o modelo de *word embeddings*, que consiste no mapeamento de palavras para vetores densos e de pequena dimensão, permitindo captar melhor a semântica das palavras e fazendo com que palavras similares tenham vetores também similares. Outras abordagens consistem no uso de dicionários, como o *Linguistic Inquiry and Word Count* (LIWC), um dicionário, disponibilizado também na língua portuguesa do Brasil, que mapeia palavras para quatro principais tipos de categorias: processos linguísticos básicos, processos psicológicos, expressões relativas e preocupações pessoais [13]. Onan e Tocoglu [14] 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 típicos de *bag-of-words*.

Quanto à classificação de texto, os algoritmos mais utilizados são as árvores de decisão, *Naive Bayes*, *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN) e *Logistic Regression* [15,16,17,18,19]. SVM com *kernels* lineares e *Naive Bayes* são os algoritmos que têm apresentado melhores resultados neste problema [19,17]. *Deep learning*, um tipo de *machine learning* que utiliza aprendizagem baseada em redes neuronais, também se tem revelado bastante eficaz no processo de classificação textual, nomeadamente na deteção de metáforas. Gao et al. [20] demonstrou 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. No entanto, todos este algoritmos dependem, muitas vezes, de uma grande número de dados de treino. Uma abordagem que tenta ultrapassar essa limitação é aprendizagem por transferência, ou *transfer learning*, onde os modelos são pré-treinados e o conhecimento aprendido num domínio é aplicado noutro domínio similar. Um modelo que faz parte desta abordagem e que tem apresentado resultados estado da arte na área de PLN é o *Bidirectional Encoder Representations from Transformers* (BERT). BERT foi lançado pelos investigadores da Google AI, em 2018, e é apresentado pela primeira vez no artigo [21], 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. BERTimbau [22] é o modelo BERT treinado na língua portuguesa do Brasil.

No âmbito do *topic modeling*, este é um processo de aprendizagem não supervisionada (não necessita de dados já classificados). Os algoritmos fundamentais correspondem, essencialmente, aos algoritmos *Linear Discriminant Analysis* (LDA) e *Probabilistic Latent Semantic Analysis* (PLSA), baseados nas distribuições de palavras. No entanto, estes modelos apresentam a limitação da captura de contexto, dificuldade na captura da semântica das palavras e necessidade de predefinição do número de tópicos a descobrir. Angelov [23] 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 de *topic modeling*.

Apesar dos grandes avanços nas área de PLN e mineração de texto, estas apresentam ainda várias lacunas e limitações, nomeadamente no processamento de textos onde se verifica a presença de ironia, metáforas, expressões idiomáticas e ambiguidade de palavras. Além disso, a própria área de classificação automática de texto encontra-se muito pouco desenvolvida em Portugal. 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 [24] 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 apresenta apenas 0.29% do total. Este projeto, é, assim, pioneiro no ramo de classificação automática de textos portugueses estigmatizantes de doenças mentais.

3 Metodologia

A metodologia adotada é caracterizada por seis principais etapas: (i) recolha dos dados; (ii) filtragem e anotação manual dos dados; (iii) pré-processamento; (iv) classificação automática e avaliação; (v) *topic modeling*; (vi) visualização e análise dos resultados.

Etapa 1: Recolha dos dados

A fonte oficial de dados, que consistem nos artigos de jornais de notícias portugueses disponibilizados na Internet e seus metadados, foi o repositório Arquivo.pt. O Arquivo.pt é um repositório de páginas web portuguesas arquivadas desde 1996 até hoje. Possui armazenadas várias categorias de páginas, que se encontram sob o domínio .PT ou têm interesse para a comunidade portuguesa [25], 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). A Arquivo.pt API[6] é a API

[6] https://github.com/arquivo/pwa-technologies/wiki/Arquivo.pt-API

pública do repositório Arquivo.pt que permite recolher páginas preservadas da Web portuguesa e seus metadados através de pesquisas de texto. Os dados são retornados em formato *JavaScript Object Notation* (JSON) e os pedidos podem ser feitos com diferentes parâmetros. Uma limitação desta funcionalidade é o facto de, para cada pesquisa, apenas ser devolvida uma resposta com no máximo 2000 itens. No entanto, após a realização de uma análise inicial da API, verificou-se que o número de itens retornados, para as pesquisas a efetuar no âmbito deste projeto, 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 dos parâmetros de pesquisa.

Tendo em conta a finalidade do projeto, os dados recolhidos consistiram no conjunto composto pelo título da notícia, data de arquivamento, nome do jornal e *Uniform Resource Locator* (URL) para a versão original e para a versão arquivada, para cada página web retornada. Para isso, foi necessário definir os parâmetros de pesquisa a utilizar.

Começando pelos termos de pesquisa, foi decidido focar-se em artigos 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 (e mais graves) mas, ao mesmo tempo, das perturbações que mais aparecem no nosso vocabulário de termos utilizados fora do seu contexto original. Para 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. Assim, 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 original das mesmas, foram recolhidos, do Arquivo.pt, todos os artigos que possuíssem pelo menos um dos seguintes termos: ["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*.

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 eletrónicos. Os critérios de seleção foram a popularidade do jornal na Internet [26], a sua relevância no âmbito do projeto e a sua longevidade. O endereço de todos os jornais não foi sempre constante ao longo dos anos, podendo estes terem pertencido a outros domínios e apresentarem vários subdomínios. Assim, foi também 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 [27], disponibilizado publicamente na página do Ar-

quivo.pt. No entanto, o relatório apenas possuía 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 do Arquivo.pt, para verificar a veracidade dos dados do relatório e também que diferentes versões do endereço foram arquivadas depois de 2016. É de salientar que da leitura do relatório e das descrições dos projetos premiados no concurso Prémio Arquivo.pt[7] nos anos anteriores, verificou-se que o repositório possui alguns problemas e limitações, sendo que alguns dos endereços podem apresentar problemas de preservação ou nem ser 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[8] de seus domínios e subdomínios entre 1996 e 2019. Os endereços utilizados, referentes aos nove jornais, podem ser visualizados na Tabela 1.

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, definiu-se o intervalo de tempo de pesquisa entre 1996 e 2021. Foi também definido para apenas serem retornadas páginas *HyperText Markup Language* (HTML), um máximo de 2000 resultados, o conjunto de campos {title, tstamp, originalURL, linkToOriginalFile, linkToArchive} a incluir em cada item da resposta e o valor "false" para o parâmetro "prettyPrint". O número total de dados retornados pela API foi de 8235 páginas web.

De seguida, procedeu-se ao processo de web *scraping* do HTML de cada página. O processo de web *scraping* foi realizado recorrendo à biblioteca *newspaper* [28], visto que a mesma conseguiu realizar o processo, de forma rápida e eficaz, para todas as páginas retornadas, mesmo as mais antigas. 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 texto do artigo), e todos os artigos duplicados, sendo a remoção efetuada através da comparação dos conteúdos. O número total de dados estruturados obtidos foi de 1111.

Etapa 2: Filtragem e anotação manual dos dados

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 diferentes anotadores humanos. Cada anotador recebeu um conjunto de dados não anotados e foi pedido para classificar o sentido de um excerto de cada artigo 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;

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

Tabela 1. Os diferentes endereços dos jornais utilizados e intervalos de tempo que possuem suas versões no Arquivo.pt

Jornal	**Endereços**	**Intervalo de tempo**
Público	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	observador.pt	2014 - 2022
Diário de Notícias	www.dn.pt	1998 - 2022
	dn.sapo.pt / www.dn.sapo.pt	2008 - 2012
Expresso	expresso.pt	1998 - 2022
	aeiou.expresso.pt	2008 - 2012
	expresso.sapo.pt	2012 - 2015
Correio da Manhã	www.correiomanha.pt	1996 - 2015
	www.correiodamanha.pt	2001 - 2009
	www.cmjornal.xl.pt	2010 - 2016
	www.cmjornal.pt	2010 - 2022
Jornal de Notícias	www.jn.pt / jn.pt	1998 - 2022
	jn.sapo.pt	2002 - 2011
Sábado	www.sabado.xl.pt:80	2006 - 2007
	www.sabado.xl.pt	2008
	sabado.pt / www.sabado.pt	2009 - 2022
Visão	aeiou.visao.pt	2009 - 2012
	visao.sapo.pt	2012 - 2022
A Bola	abola.pt / www.abola.pt	2000 - 2007
	abola.pt:80	2008 - 2022

– Literal: o excerto do artigo não é estigmatizante, utiliza a doença no seu sentido literal e dentro de um contexto adequado;
– Indefinido: o anotador não consegue decidir a categoria.

Foi disponibilizada uma mesma instrução, a cada anotador, referindo em que circunstâncias, baseadas nas apresentadas em estudo anteriores, cada uma das categorias deve ser atribuída. Um exemplo de excertos de artigos classificados pode ser visualizado na Tabela 2. 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.

Tabela 2. 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
Face à recusa de italiana de aceitar, nos seus portos, o navio Aquarius, que seguia com 690 migrantes resgatados do Mediterrâneo a bordo, e às declarações do primeiro-ministro húngaro, Viktor Orbán, que diz que "a Hungria é contra a mistura" com povos estrangeiros, o Papa acredita que os populistas estão a "criar uma **psicose**" sobre a questão da imigração.	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

No final, foram obtidos 978 artigos manualmente anotados com as classes ['estigmatizante', 'literal']. É também de referir que durante este processo alguns artigos foram descartados por apresentarem problemas estruturais, serem duplicados ou não serem relevantes para o problema.

Etapa 3: Pré-processamento

Durante a fase de pré-processamento, foi realizada uma limpeza dos documentos (N=978) e utilizadas técnicas de PLN, para preparar os textos 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*[9], obtidas do NLTK [29];
- 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 e de todos os pronomes pessoais conectados a verbos.

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.

Etapa 4: Classificação automática e avaliação

A etapa da classificação implicou a ocorrências de duas fases: extração de atributos e implementação dos modelos de classificação. Na extração de atributos, foram utilizados quatro modelos diferentes para gerar representações numéricas dos documentos, sendo eles o modelo de *bag-of-words*, o modelo de TF-IDF, o modelo de *word embeddings*, utilizando vetores de 300 dimensões pré-treinados, para a língua portuguesa, com o algoritmo GloVe [30] e obtidos do repositório NILC-Embeddings[10], e o mapeamento dos termos dos textos para as 464 categorias do dicionário *Brazilian Portuguese LIWC 2007 Dictionary*[11].

O processo de classificação consistiu na implementação dos algoritmos, 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). Foram utilizados cinco algoritmos tradicionais de *machine learning* e quatro algoritmos de *deep learning*. Os algoritmos de *machine learning* foram implementados utilizando a biblioteca scikit-learn[31] e consistiram nos algoritmos:

[9] https://www.nltk.org/howto/portuguese_en.html
[10] http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc
[11] http://143.107.183.175:21380/portlex/index.php/pt/projetos/liwc

- *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;
- 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 "Linear Support Vector Classification"(*kernel* linear);
- *Naive Bayes*: algoritmo probabilístico baseado no Teorema de Bayes e na suposição de independência condicional dos atributos dada uma classe. Foi utilizada a classe "Multinomial Naive Bayes", específica para a classificação de atributos discretos;
- 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;
- *Random Forest*: algoritmo que ajusta um número de classificadores de árvore de decisão em várias subamostras do conjunto de dados de treino e que combina os resultados de cada classificador para determinar a classe final.

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

Os algoritmos de *deep learning* foram implementados utilizando a biblioteca Tensorflow[32], a API Keras[33], a *framework* PyToch[34] e a biblioteca transformers [35], e consistiram nos algoritmos:

- *Convolutional Neural Network* (CNN): tipo de rede neuronal que é tipicamente utilizado no reconhecimento de imagens mas que também tem sido usado em tarefas de PLN. A CNN foi implementada, sequencialmente, com uma camada de *embedding*, que gera vetores de 300 dimensões usando os *word embeddings* do modelo 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"). Foi treinada usando um tamanho de *batch* de 32 e valor de *epochs* de 10.
- *Long Short-Term Memory* (LSTM): tipo de rede neuronal recorrente que mantém apenas as informações necessárias ou úteis para previsão. A rede LSTM foi implementada, sequencialmente, com uma camada de *embedding*, que gera vetores de 300 dimensões usando os *word embeddings* do modelo GloVe, uma camada de LSTM e uma camada densa (com função de ativação "sigmoid"). Foi 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. Foi implementada, sequencialmente, com uma camada de *embedding*, que gera vetores de 300 dimensões usando os *word embeddings* do modelo GloVe, uma camada de LSTM (inserida numa camada bidirecional) e uma camada densa (com função de ativação "sigmoid"). Foi treinada usando um tamanho de *batch* de 32 e valor de *epochs* de 10.
- BERTimbau: 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. Foi treinado usando um tamanho de *batch* de 8 e valor de *epochs* de 4.

A otimização dos restantes hiperparâmetros dos primeiros três modelos referidos foi efetuada recorrendo à biblioteca Keras Tuner [36], com o *tuner* "Hyperband", que utiliza o algoritmo de *random search* e procura acelerá-lo através da alocação adaptativa de recursos e paragem antecipada.

Etapa 5: *Topic modeling*

A deteção automática de tópicos foi realizada usando o algoritmo top2vec, que 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 e pontuações da sua similaridade ao tópico e os documentos semanticamente mais similares a cada tópico.

Etapa 6: Visualização e análise dos resultados

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 [37] e seaborne [38], e, posteriormente, apresentadas num *website* criado usando a biblioteca React. O *website* (Figura 2) foi criado com o objetivo de apresentar o projeto e os principais resultados obtidos de uma forma mais simples, interativa e intuitiva.

4 Resultados e Discussão

De um modo geral, este projeto permitiu obter:

- um conjunto de 978 artigos de jornais portugueses *online*, que fazem referências aos transtornos mentais da esquizofrenia e psicose, manualmente anotados como detentores de um sentido estigmatizante ou literal;
- um conjunto de nove algoritmos de *machine learning* e *deep learning* que realizam a classificação automaticamente;
- um conjunto de tópicos extraídos automaticamente dos artigos.

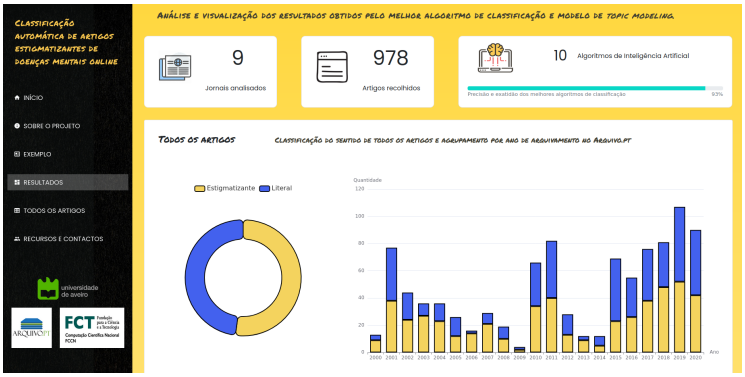

Figura 2. Secção da *dashboard* com os resultados, no *website*.

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. O agrupamento destes resultados por ano de arquivamento no Arquivo.pt pode ser visualizado na Figura 3. 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. Os anos que obtiveram a maior diferença entre número de artigos com sentido estigmatizante e com sentido literal foram os de 2003 e 2018. O agrupamento destes resultados por jornal de notícias pode ser visualizado na Tabela 3. 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.

Quanto aos resultados obtidos pelos algoritmos de classificação automática desenvolvidos, os valores das métricas utilizadas para avaliar o seu desempenho (exatidão, precisão, *recall* e *F1-score*/F1) podem ser visualizados e comparados na Tabela 4. 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 exatidão acima dos 90%, destacando-se no topo os algoritmos de classificação *Naive Bayes* (93.37%) 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 (91.84%) e o Bi-LSTM (91.33%). O modelo de *deep learning* LSTM obteve também o me-

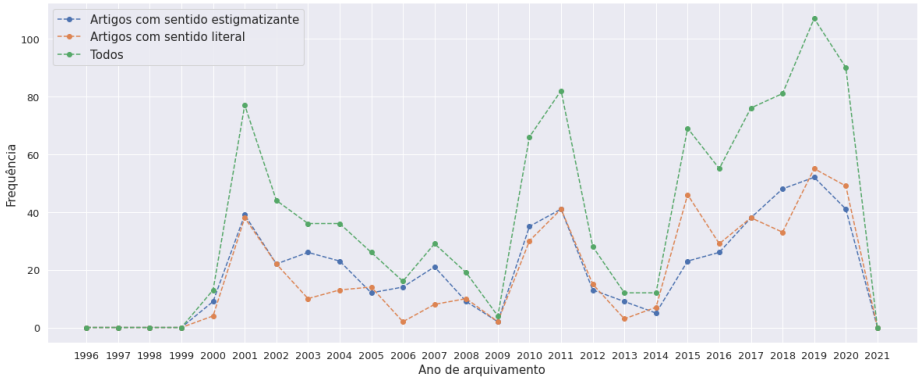

Figura 3. Agrupamento dos artigos, manualmente classificados, por ano de arquivamento no Arquivo.pt.

Tabela 3. Agrupamento dos sentidos 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 4. 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*	92.35	0.92	0.93	0.93
	TF-IDF	**93.37**	0.93	0.94	**0.94**
	LIWC	70.41	0.73	0.69	0.71
SVM	*Bag-of-words*	90.31	0.92	0.90	0.91
	TF-IDF	90.82	0.93	0.90	0.91
	LIWC	80.10	0.81	0.80	0.81
Naive Bayes	*Bag-of-words*	91.33	0.91	0.92	0.92
	TF-IDF	**93.37**	0.91	0.97	**0.94**
	LIWC	52.04	0.52	**1.00**	0.69
KNN	*Bag-of-words*	65.82	0.89	0.40	0.54
	TF-IDF	91.33	0.92	0.91	0.92
	LIWC	70.41	0.72	0.70	0.71
Random Forest	*Bag-of-words*	92.86	0.90	0.97	0.93
	TF-IDF	91.84	0.88	0.97	0.93
	LIWC	79.08	0.77	0.86	0.81
CNN	*Word embeddings*	87.76	0.92	0.83	0.88
LSTM	*Word embeddings*	87.24	**0.96**	0.78	0.87
Bi-LSTM	*Word embeddings*	91.33	0.90	0.94	0.92
BERTimbau	*Tokenizer* do BERTimbau	91.84	0.93	0.91	0.92

lhor valor de precisão (0.96) de entre todos, o que significa que de todos os artigos que o modelo classificou como estigmatizantes, 96% eram realmente estigmatizantes. No entanto, apesar dos bons resultados, os algoritmos de *deep learning* conjugados com *word embeddings*, não superaram, no geral, os resultados dos tradicionais de *machine learning* conjugados com representações de atributos mais simples, o que pode sugerir a necessidade de experimentar com outros algoritmos de *word embeddings* ou gerar novos, através do treino com maior volume de textos portugueses. Por fim, quanto ao *recall*, que calcula quantos dos artigos estigmatizantes foram classificados como tais, o modelo que apresentou melhor resultado (1.00) foi o *Naive Bayes* conjugado com LIWC, apesar da baixa exatidão (52.04%) e precisão (0.52).

Foram automaticamente detetados dez tópicos, cada um definido por um conjunto de 50 termos mais descritivos do mesmo. Na Tabela 5 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%).

5 Conclusão

Neste projeto, foi realizada a recolha e classificação manual de artigos, com referências a doenças mentais, de jornais de notícias portugueses presentes na Internet e arquivados no repositório Arquivo.pt, bem como a exploração de técnicas de Inteligência Artificial para a realização automática das tarefas de classificação e *topic modeling*. Foram propostos nove 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. Além disso, foi 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.

Do nosso conhecimento, este é o primeiro trabalho que explora a classificação de textos portugueses que contêm referências metafóricas através do uso de modelos computacionais, sendo que as grandes conclusões retiradas são que a maioria dos tradicionais algoritmos de *machine learning* permitem obter bons resultados e que o uso de redes neuronais sugere também ser bastante promissor. No entanto, o campo do PLN portuguesa encontra-se ainda muito pouco explorado, o que se revela também na escassa quantidade de modelos treinados para o português de Portugal, existindo também abordagens mais complexas que devem ser, futuramente, consideradas.

Tabela 5. 20 termos mais descritivos de cada tópico, classificação geral atribuída e número de artigos total e estigmatizantes.

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

Referências

1. What Is Mental Illness? American Psychiatric Association. url: https://www.psychiatry.org/patients-families/what-is-mental-illness (acedido em 23/10/2021).
2. ≪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.
3. Ordem dos Psicólogos Portugueses. ≪Desenvolvimento Sustentável e Sustentabilidade dos Cuidados de Saúde Primários≫. Em: Lisboa, Portugal, 2021. isbn: 978-989-53170-2-8.
4. 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).
5. 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).
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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).
12. 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.
13. James Pennebaker, Martha Francis e Roger Booth. ≪Linguistic Inquiry and Word Count (LIWC)≫. Em: (1 de jan. de 1999).

14. Aytug Onan e Mansur Togoclu. ≪Satire Identification in Turkish News Articles Based on Ensemble of Classifiers≫. Em: TURKISH JOURNAL OF ELECTRICAL ENGI- NEERING COMPUTER SCIENCES 28 (28 de mar. de 2020), pp. 1086–1106. doi: 10.3906/elk-1907-11.
15. 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_46.
16. Shitao Zhang. ≪Sentiment Classification of News Text Data Using Intelligent Mo- del≫. Em: Frontiers in Psychology 12 (2021), p. 4398. issn: 1664-1078. doi: 10.3389/fpsyg.2021.758967.
17. 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.
18. 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.
19. Bi-Min Hsu. ≪Comparison of Supervised Classification Models on Textual Data≫. Em: Mathematics 8.5 (2020). issn: 2227-7390.doi: 10.3390/math8050851.
20. 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.
21. Jacob Devlin, Ming-Wei Chang, Kenton Lee e Kristina Toutanova. ≪BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding≫. Em: Pro- ceedings 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.
22. 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. 2020.
23. Dimitar Angelov. ≪Top2Vec: Distributed Representations of Topics≫. Em: ArXivabs/2008.09470 (2020).
24. 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.
25. 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).
26. 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
27. 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/

28. Lucas Ou-Yang. Newspaper3k: Article scraping curation. Em: (1 de fev. de 2022). url: https://newspaper.readthedocs.io/en/latest/
29. Bird, Steven, Edward Loper e Ewan Klein. ≪Natural Language Processing with Python≫. O'Reilly Media Inc. Em: (2009).
30. 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.
31. Pedregosa et al.. ≪Scikit-learn: Machine Learning in Python≫. Em: JMLR 12, pp. 2825-2830 (2011).
32. Martín 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, Rafal Jozefowicz, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Mike Schuster, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, 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 systems≫. Em: (2015). url: tensorflow.org
33. Chollet, Francois e outros. Keras. 2015. url: https://keras.io
34. Paszke, Adam e Gross, Sam e Massa, Francisco e Lerer, Adam e Bradbury, James e Chanan, Gregory e Killeen, Trevor e Lin, Zeming e Gimelshein, Natalia e Antiga, Luca e Desmaison, Alban e Kopf, Andreas e Yang, Edward e DeVito, Zachary e Raison, Martin e Tejani, Alykhan e Chilamkurthy, Sasank e Steiner, Benoit e Fang, Lu e Bai, Junjie e Chintala, Soumith. ≪PyTorch: An Imperative Style, High-Performance Deep Learning Library≫. Em: Advances in Neural Information Processing Systems 32. Curran Associates, Inc..
35. 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, out. de 2020, pp. 38-45.
36. O'Malley, Tom e Bursztein, Elie e Long, James e Chollet, Francois e Jin, Haifeng e Invernizzi, Luca e outros. Keras Tuner. 2019. url: https://github.com/keras-team/keras-tuner
37. J. D. Hunter, ≪Matplotlib: A 2D Graphics Environment≫ Em: Computing in Science Engineering, vol. 9, no. 3, pp. 90-95, 2007.
38. Michael L. Waskom. ≪seaborn: statistical data visualization≫. Em: Journal of Open Source Software, vol. 6, nº. 60, pp. 3021, 2021. doi: 10.21105/joss.03021.