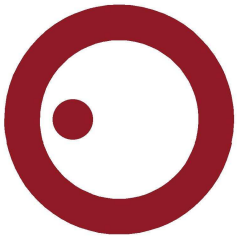


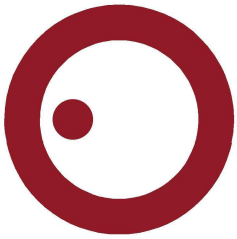
escola superior de tecnologia e gestão
instituto politécnico de leiria

Instituto Politécnico de Leiria
Escola Superior de Tecnologia e Gestão
Departamento de Engenharia Informática
Mestrado em Cibersegurança e Informática Forense

APLICAÇÃO PARA DETEÇÃO AUTOMÁTICA DE DISCURSO DE ÓDIO EM LÍNGUA PORTUGUESA RECORRENDO A APRENDIZAGEM COMPUTACIONAL

ESTUDANTE LUÍS HENRIQUE PEREIRA NEVES

Leiria, setembro de 2022



escola superior de tecnologia e gestão
instituto politécnico de leiria

Instituto Politécnico de Leiria
Escola Superior de Tecnologia e Gestão
Departamento de Engenharia Informática
Mestrado em Cibersegurança e Informática Forense

APLICAÇÃO PARA DETEÇÃO AUTOMÁTICA DE DISCURSO DE ÓDIO EM LÍNGUA PORTUGUESA RECORRENDO A APRENDIZAGEM COMPUTACIONAL

ESTUDANTE LUÍS HENRIQUE PEREIRA NEVES
Número: 2200191

Projeto realizado sob orientação do Professor Doutor Mário João Gonçalves Antunes (mario.antunes@ipleiria.pt).

Leiria, setembro de 2022

Sticks and stones may break my bones, but words will always hurt me.

— *Stephen Fry, Moab Is My Washpot*

AGRADECIMENTOS

Consciente de que o presente projeto advém de um conjunto de esforços que envolvem imprescindíveis e distintas pessoas, venho agradecer:

Ao meu orientador Professor Doutor Mário João Gonçalves Antunes por todas as orientações, toda a disponibilidade prestada e saber transmitido, fundamentais na elaboração deste projeto.

À Claudia Claro minha companheira de vida, uma vez mais termino uma nova fase com o teu apoio incondicional e toda a compreensão possível, muito obrigado, *je t'aime*.

Ao Nuno, o melhor irmão que se pode ter, sempre presente mesmo quando não existe tempo, és o meu orgulho.

Aos meus pais, pelo apoio e incentivo, principalmente pela compreensão da minha ausência.

Por fim, a todos os parceiros de longas noites de estudo, meus caros Sérgio Lavos, Luís Serra, Pedro Realinho, Bruno Jordão e Ricardo Costa o meu muito obrigado!

RESUMO

Assistimos a uma massificação da utilização das tecnologias da informação sem qualquer diferenciação de idade, nem tão pouco requerendo qualquer conhecimento aprofundado de informática, independentemente do tipo de equipamento utilizado (dispositivos móveis, computadores ou até eletrodomésticos equipados com conectividade à Internet). O crescente aumento de utilizadores no ciberespaço tornou-o um local de partilha de informação pessoal e alojamento de documentos pessoais e/ou profissionais, reunindo assim todas as condições para se tornar um meio de diminuição da distância física, através da utilização copiosa das redes sociais. A utilização massificada da Internet e do ciberespaço potenciou o aparecimento de utilizadores mal-intencionados, cuja motivação principal é a prática de cibercrimes. As principais atividades ilícitas exercidas no ciberespaço e que constituem a prática de crime variam, desde o acesso a dados de forma ilegal, apropriação de contas de utilizadores nas mais diversas plataformas, até à invasão de privacidade (física e virtual).

A utilização das redes sociais e a partilha de conteúdos que aí é feita potencia a prática de crimes relacionados com a invasão de privacidade, onde se destacam o *cyberbullying*, a perseguição digital (*cyberstalking*), a exposição pública na Internet de conteúdos digitais íntimos, sem o consentimento da pessoa visada (*revenge porn*) e a troca intencional e abusiva de mensagens de cariz sexual (*sexting*). A proliferação, diversidade e heterogeneidade das redes sociais existentes tem ampliado a prática deste tipo de crimes, com consequências diretas para as vítimas. Por outro lado, o volume de dados a analisar requerem a implementação de aplicações que processem automaticamente as publicações nas redes sociais, com vista à identificação automática e em tempo real de potenciais atividades genericamente associadas à disseminação de discurso de ódio.

O panorama em Portugal está em linha com o que se passa globalmente em todo o mundo, tendo as autoridades nacionais registado um aumento substancial de crimes relacionados com o *cyberbullying* e a utilização indiscriminada do discurso de ódio. A implementação de aplicações para a deteção automática deste tipo de comportamento em publicações nas redes sociais reveste-se assim da maior importância. Em língua inglesa há várias aplicações, mas no contexto da língua

portuguesa, tendo em conta o elevado número de utilizadores, os trabalhos existentes são escassos e direcionados para algumas variantes, como o português do Brasil.

Neste projeto apresenta-se um protótipo de aplicação para a deteção automática de texto que inclua discurso de ódio. Esta aplicação poderá, por exemplo, ser utilizada para analisar, em tempo real, as publicações e correspondentes comentários de um *feed* de uma rede social, detetando discurso de ódio, possibilitando assim uma intervenção célere, como o bloqueio do utilizador ou a remoção da publicação. Foi igualmente desenvolvida uma extensão para o navegador *Google Chrome*, que permite detetar o discurso de ódio numa página web, facilitando assim a sua utilização no contexto da navegação web. No desenvolvimento do protótipo foram utilizados e testados vários métodos de aprendizagem computacional (*machine learning*), designadamente *Naïve Bayes, SVM, kNN, Random Forest* e *Neural Network*.

Ainda no âmbito deste projeto foi criado um dataset anotado e classificado, contendo 354 publicações retiradas da obra de "Para cima de puta"da autora Ferreira, 2021 e de redes sociais como o *Facebook*, o *Twitter* e o *Youtube*, com exemplos reais de texto legítimo e de conteúdos associados a *cyberbullying* e discurso de ódio. Os testes realizados com os métodos *Naïve Bayes, SVM, kNN, Random Forest, Neural Network* e ainda combinações entre os métodos anteriores, permitiram obter um valor de *F-score* de 73% e um precisão de 78%, resultados provenientes da combinação dos algoritmos *kNN* & *Neural Network*.

ABSTRACT

We are witnessing a massification of information technologies without any age differentiation nor requiring in-depth computer knowledge, regardless of the type of equipment used (mobile devices, computers, or even household appliances equipped with Internet connectivity).

The growing increase of users in cyberspace has made it a place for sharing personal information and hosting personal and professional documents, thus meeting all the conditions to reduce physical distance through the copious use of social networks. However, the adherence to and increased use of the Internet and cyberspace has boosted the appearance of malicious users, whose primary motivation is the practice of cybercrime. The main illicit activities carried out in cyberspace and which constitute the practice of crime range from illegal access to data and appropriation of user accounts on the most diverse platforms to invasion of privacy (physical and virtual).

With the use of social networks and the sharing of content that is done there is a potential for the commission of crimes related to invasion of privacy, including cyberbullying, cyberstalking, and public exposure of personal records on the Internet without the consent of the person concerned (preventive porn), and the intentional and abusive exchange of sexually oriented messages (sexual harassment). The proliferation, diversity, and heterogeneity of existing social networks have increased the practice of this type of crime, with direct consequences for the victims. On the other hand, analyzing the volume of data requires applications that automatically process the publications on social networks and identify, in real-time, potential activities generically associated with the dissemination of hate speech.

The panorama in Portugal is in line with what is happening globally, with the national authorities registering a substantial increase in crimes related to cyberbullying and the indiscriminate use of hate speech. Therefore, the implementation of applications for automatically detecting this behavior in publications on social networks is of the utmost importance. In the English language, there are several applications. However, considering the high number of users in the Portuguese language context, existing works are scarce and directed to some variants, such as Brazilian Portuguese.

This project presents a prototype of an application for automatically detecting text that includes hate speech. This application can, for example, be used to analyze, in real-time, the posts and corresponding comments of a social network's feed, detecting hate speech, thus enabling a quick intervention, such as blocking the user or removing the post. In addition, an extension was also developed for the Google Chrome browser, which allows the detection of hate speech on a web page, thus facilitating its use in the context of web browsing.

In the development of the prototype, several machine learning methods were used and tested, namely Naïve Bayes, SVM, kNN, Random Forest, and Neural Network.

Also, in the scope of this project, an annotated and classified dataset was created, containing 354 posts taken from the Ferreira, 2021 "Para cima de puta"work and from social networks such as Facebook, Twitter, and Youtube, with real examples of legitimate text and content associated with cyberbullying and hate speech. The tests performed with the Naïve Bayes, SVM, kNN, Random Forest, Neural Network methods and combinations between the previous methods allowed to obtain an F-score value of 73% and an accuracy of 78%, results from the combination of the kNN & Neuronal Network.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

AdaBoost	Adaptive Boosting.
API	Application Programming Interface.
AUC	Area under the ROC Curve.
BOW	Bag Of Words.
CNN	Convolution neural network.
CP	Código Penal.
CRISP-DM	Cross-industry standard process for data mining.
DCBD	Descoberta de Conhecimento em Bases de Dados.
DT	Decision tree.
FE	Feature Extraction.
FN	False Negative.
FP	False Positive.
HTML	HyperText Markup Language.
HTTP	Hypertext Transfer Protocol.
HTTPS	Hypertext Transfer Protocol Secure.
IDF	Inverse Document Frequency.
IP	Internet Protocol.
KDD	Knowledge Discovery in Databases.
KNN	K-nearest Neighbors.

LR	Logistic Regression.
ML	Machine Learning.
MLP	Multilayer perception.
MNB	Multinomial naive bayes.
mSVM	Multiple view stacked Support Vector Machine.
NB	Naïve Bayes.
NLP	Natural Language Processing.
NN	Neural Networks.
OCR	Optical character recognition.
P	Precision.
R	Recall.
REST	Representational state transfer.
RF	Random Forest.
RNN	Recurrent neural networks.
ROC	Receiver operating characteristic.
SEMMA	Sample, Explore, Modify, Model, and Assess.
SMOTE	Synthetic Minority Oversampling Technique.
SNC	Sistema Nervoso Central.
SSH	Secure Shell.
SVM	Support Vector Machine.
TCP	Transmission Control Protocol.
TIC	Tecnologias da Informação e Comunicação.

TN True Negative.

TP True Positive.

VN Verdadeiro Negativo.

VP Verdadeiro Positivo.

VPS Virtual Private Server.

1

INTRODUÇÃO

As Tecnologias da Informação e Comunicação (TIC) estão em continuo desenvolvimento. Consequentemente, observa-se um crescimento exponencial da utilização da Internet a nível mundial, o que hoje se traduz numa representação do mundo real e da constituição do ciberespaço (Serrão, 2019). Esta crescente e massiva utilização traz benefícios como novas oportunidades de aprendizagem e de socialização, tanto para adultos como para crianças e jovens. Todavia, a utilização das TIC e da Internet por indivíduos com intenções maliciosas e orientadas para a criminalidade, associada ainda a uma utilização descuidada e negligente, pode potenciar a exploração de vulnerabilidades (técnicas e humanas) dos utilizadores em geral.

Globalmente, um terço dos utilizadores da Internet são crianças e jovens até aos dezoito anos (APAV, 2021c). Decorrente deste facto, cada vez mais a Internet é utilizada como meio facilitador para a prática de disseminação de discurso ódio e a difusão ou pretensão de praticar ofensas sexuais contra menores (APAV, 2021a). Neste contexto, destaca-se o *bullying* praticado no ciberespaço (*cyberbullying*), que na sua maioria, atinge as faixas etárias mais baixas.

O *bullying* revela uma violência cobarde, perturbadora e tem repercussões incontestáveis na sociedade. Nos últimos anos, de acordo com Chase e Statham, 2005, este fenómeno ganhou um novo palco: o mundo virtual. Assim, surge o conceito de *cyberbullying* que, devido à liberdade de manifestação de pensamento e partilha nas redes sociais, permite criar uma desvinculação com o ser humano que se encontra do outro lado do ecrã, facultando aos agressores, sob a falsa perceção de liberdade de expressão, a manifestação de pensamentos do mais avassalador que se possa imaginar, efetivando-se através de distintos comportamentos agressivos (físicos, sexuais e/ou verbais), com ou sem contacto ou confrontação direta entre a vítima e o agressor.

Neste contexto, salienta-se o *hate speech* (discurso de ódio), utilizado no *bullying* e *cyberbullying*, e que consiste num discurso em que a dignidade humana é posta em causa, ao ser denegrida e inferiorizada. Assevera-se uma exteriorização ofensiva, tendo como único objetivo inferiorizar minorias, incitar à violência e fazer prevalecer a superioridade do agressor. Pela transversalidade do uso do mesmo, atualmente o

contacto entre um agressor e um menor ocorre de forma instantânea, garantindo simultaneamente o seu anonimato (APAV, 2021b). Atualmente, qualquer utilizador pode ser um alvo de violência na Internet, abrangendo adultos, crianças e jovens. Segundo o relatório da unicef, 2017, é possível identificar três tipos de riscos para as crianças e jovens, nomeadamente: a exposição a conteúdo não desejado e impróprio (pornografia, violência, publicidade ou propaganda racista, discriminatória e o diálogo de ódio), o contacto, formalizado através de conversas que as crianças possam ter com indivíduos mais velhos em busca de encontros sexuais e, por último, a conduta, que envolve os comportamentos dos próprios menores, influenciados pelos atos levados a cabo por outros (desafios que implicam a ingestão de certos alimentos, por exemplo).

Para Antunes e Rodrigues, 2018, qualquer crime que aconteça no ciberespaço é definido como cibercrime. Em Portugal, o Código Penal não considera o *bullying* enquanto um ato criminoso.

Embora não esteja previsto na lei como crime específico, o *bullying* pode ser punível através da ocorrência de diversos e distintos comportamentos, que são considerados crimes na legislação em vigor. Assim, o *bullying* pode ser punível através de um conjunto de crimes, dentro dos quais as agressões (artigo 143º do Código Penal), as ameaças (artigo 153º do Código Penal), as difamações (artigo 180º do Código Penal) e as injúrias (artigo 181º do Código Penal). De igual modo, a prática de *cyberbullying* não possui legislação específica, contudo as ações que a caracterizam estão previstas pelo ordenamento jurídico nacional (Serrão, 2019). Considerando a Constituição da República, o Código Penal Portuguesa, a Lei do Cibercrime, e outros diplomas legislativos nacionais, bem como orientações e determinações europeias, verifica-se um regime sancionatório que visa responder a estes casos e denota vontade de prevenção e supressão deste fenómeno.

1.1 MOTIVAÇÃO

Embora se denote uma preocupação crescente sobre este fenómeno de *cyberbullying*, ainda são limitadas as plataformas e aplicações informáticas para a deteção automática de discurso de ódio, sendo que as existentes são essencialmente focadas na língua inglesa. Do ponto de vista económico, qualquer entidade em que o seu serviço consista na disponibilização de um canal publico de comunicação, deverá considerar a adoção de mecanismos de deteção e bloqueio de discurso de ódio, de forma a garantir a satisfação generalizada dos seus utilizadores e, consequentemente,

aumentar a sua receita. Relativamente à esfera social, detetar discurso de ódio na Internet é de elevada importância para a proteção das minorias, podendo asseverar-se como uma forma de diminuir a existência da radicalização de grupos extremistas (MacAvaney et al., 2019).

A língua portuguesa é oficial em 9 países e falada por cerca de 273 milhões de pessoas. No entanto, a deteção automática de discurso de ódio em língua portuguesa tem pouca expressão no contexto dos trabalhos científicos disponíveis. É, contudo, possível identificar alguns desenvolvimentos em variantes da língua portuguesa, nomeadamente o português do Brasil, embora a terminologia e os significados de determinadas palavras, adjacente à distinção cultural e social, forçosamente diferem.

Nos últimos anos, com relevância para as restrições emergentes da situação pandémica da COVID-19, observou-se um aumento significativo de casos de *cyberbullying*. De facto, as principais organizações mundiais (governamentais e não governamentais) aumentaram a sua preocupação no que toca a discursos de ódio, incentivando e impulsionando diversos estudos científicos neste domínio, nenhum deles em língua portuguesa. Porém, existem diversos estudos, com diferentes abordagens e, na sua maioria, para a língua inglesa. Os estudos analisados reportam a dificuldade da deteção e classificação com elevado nível de assertividade de um discurso de ódio, onde a dificuldade da língua incrementa de forma proporcional o desafio de deteção automática. São vários os aspetos que incrementam de sobremaneira a dificuldade de deteção automática do discurso de ódio, salientando-se a análise automática da semântica de uma mensagem, a utilização do jargão e a existência de vários dialetos numa língua.

O processamento de linguagem natural (NLP - *Natural Language Processing*) tem constituído um recurso valioso no processamento automático de texto e na identificação dos principais termos aí existentes. Além da análise sintática, o uso de NLP tem permitido a análise semântica e, aliado a métodos de aprendizagem computacional (ML - *machine learning*), tem auxiliado na identificação automática de variações no discurso que dificultam a classificação, como o sarcasmo.

A deteção de sarcasmo é um campo ainda pouco explorado em NLP, contrariamente à análise de sentimentos, onde as categorias do sentimento são claramente bem definidas (amor é objetivamente um sentimento positivo, ódio é um sentimento negativo, independentemente de quem pergunte, ou da língua utilizada). Assim, delinear os limites do sarcasmo é uma tarefa mais complexa.

Como previamente mencionado, existem várias abordagens e métodos de deteção de discurso de ódio. Contudo, na maioria dos estudos científicos, os seus autores

identificam como maior desafio a aquisição e a classificação de *datasets*, que servem de treino para os classificadores de ML.

O presente projeto pretende abordar todo o processo de recolha, extração e processamento de mensagens em língua portuguesa, com vista à identificação automática do discurso de ódio. Para tal, foi criado um *corpus* em língua portuguesa (anotado e classificado) e desenvolvido um protótipo de classificação do texto assente em técnicas de ML, designadamente *Naïve Bayes, Support Vector Machines (SVM), Logistic Regression, KNN e Neural Network*. Foi igualmente elaborada uma extensão para a web, designadamente para o navegador *Google Chrome*, que permite testar o modelo de classificação criado, verificando em quase tempo real a existência de discurso de ódio em língua portuguesa, numa página *web*. Desta forma, o utilizador pode navegar através de várias páginas da web e validar se o texto apresentado se refere a discurso de ódio.

1.2 OBJETIVOS E CONTRIBUIÇÕES

Os objetivos delineados para este projeto são os seguintes:

1. Identificar os principais trabalhos científicos relativos à deteção automática de discurso de ódio e *cyberbullying*.
2. Identificar os trabalhos existentes para a deteção deste tipo de mensagens no contexto da língua portuguesa.
3. Construir um *corpus* anotado, em língua portuguesa, constituído por mensagens de discurso de ódio e de texto legítimo, com vista ao *benchmark* de vários métodos de aprendizagem computacional. O *corpus* contempla exemplos de texto de discurso de ódio e de discurso que não é ódio.
4. Desenvolver um protótipo para a deteção automática de discurso de ódio em língua portuguesa. O protótipo inclui o pré-processamento e processamento das mensagens. Na componente de deteção foram utilizados vários métodos de ML, designadamente *Naïve Bayes, Support Vector Machines (SVM), Logistic Regression, K-nn e Neural Network*.
5. Comparar os resultados obtidos com os métodos de ML utilizados e apresentar as principais conclusões.
6. Desenvolver uma extensão para o navegador Google Chrome, que permita a identificação em tempo real de discurso de ódio em páginas *web*.

As principais contribuições obtidas com este projeto são as seguintes:

1. *Corpus* anotado e classificado, em língua portuguesa, com exemplos de discurso de ódio, para ser usado no *benchmark* de aplicações de deteção de *cyberbullying* com recurso a vários métodos de ML e também em aplicações que recorram a *NLP*. O *corpus* desenvolvido encontra-se disponível no *GitHub*, no seguinte *link*: https://github.com/LuisHN/Detetor-Discurso-Odio/tree/main/classificador/dataset
2. Protótipo de uma aplicação para a deteção de discurso de ódio em língua portuguesa, através de uma aplicação desenvolvida em *Python*, que incorpora os modelos obtidos através do *software Orange* [1]. O protótipo desenvolvido encontra-se disponível no *GitHub*, no seguinte *link*: https://github.com/LuisHN/Detetor-Discurso-Odio/tree/main/classificador
3. Extensão para o navegador *Google Chrome*, que permite detetar discurso de ódio em páginas *web*. A extensão encontra-se atualmente em validação pela *Google*, todavia o código está disponível no seguinte *link*: https://github.com/LuisHN/Detetor-Discurso-Odio/blob/main/extension.zip. O Anexo C descreve os passos para a instalação da extensão.

1.3 ORGANIZAÇÃO DO PROJETO

O documento está estruturado da seguinte forma:

1. No capítulo 2 descrevem-se os conceitos considerados essenciais para a correta interpretação deste projeto, bem como uma revisão da literatura quanto aos avanços existentes na área da classificação de texto, mais concretamente quanto à deteção de discurso de ódio.
2. No capítulo 3 caracteriza-se o *cyberbullying* e as suas motivações, a sua disposição legal, os impactos psicossociais, bem como os mecanismos atuais de intervenção.
3. No capítulo 4 são identificadas as várias etapas inerentes ao processo de descoberta de conhecimento em base de dados, nomeadamente aquisição de dados, pré processamento, *data mining*, seleção e teste de métodos e algoritmos e os seus respetivos resultados. É ainda apresentada a prova de conceito, desde a conceção dos objetivos a alcançar até à seleção das tecnologias a utilizar.

1 https://orangedatamining.com/

4. No capitulo 5 são apresentados os resultados obtidos decorrentes do desenvolvimento efetuado.

5. Finalmente, no capítulo 6 sumariza-se a motivação que alavancou o presente projeto, a sua elaboração e dificuldades encontradas, culminando na sugestão de trabalho futuro em prol da evolução da temática em estudo.

2

CONCEITOS FUNDAMENTAIS E ESTADO DA ARTE

Independentemente do nível de intervenção (saúde, laboral, educacional, entre outros), a recolha e análise de dados é um ferramenta útil para o processo de tomada de decisões. Contudo, a quantidade e a complexidade dos dados recolhidos é, muitas vezes, alvo de uma análise deficitária, originando uma aquisição de informação e/ou conhecimento muito aquém do expectável. Este resultado deve-se ao tempo necessário que a complexidade dos dados exige, ou até devido à incapacidade de processamento e identificação de padrões.

Tendo em conta a importância de aquisição de conhecimento nas bases de dados que guardam os dados recolhidos, foram desenvolvidas várias tecnologias e técnicas que tornam a tarefa de análise mais exequível (*Azevedo e Santos, 2008*), destacando-se os três principais:

- *Knowledge Discovery in Databases* (KDD), que poderá traduzir-se para Descoberta de Conhecimento em Bases de Dados (DCBD);
- *Sample, Explore, Modify, Model, and Assess* (SEMMA);
- *Cross-industry standard process for data mining* (CRISP-DM);

Importa ressalvar que, dependendo do contexto em que são aplicados, algumas técnicas poderão apresentar melhores resultados do que outras. No âmbito deste projeto foram aplicadas técnicas de KDD, pela sua aplicabilidade na classificação de texto (Lertvittayakumjorn e Toni, 2022). O presente capítulo apresenta, além dos fundamentos essenciais para a interpretação do trabalho realizado, uma revisão da literatura acerca das várias abordagens utilizadas na classificação de texto.

2.1 CONCEITOS FUNDAMENTAIS

Esta secção descreve os principais conceitos associados às técnicas de classificação de texto, métricas de avaliação dos métodos de classificação e das técnicas de descoberta de conhecimento em bases de dados.

2.1.1 *Técnicas de classificação de texto*

As técnicas de classificação podem ser genericamente divididas em abordagens estatística e de *machine learning*. As técnicas estatísticas visam apenas satisfazer as hipóteses de uma forma manual, não necessitando de grande auxílio de algoritmos, ao contrário das técnicas de ML (Allahyari et al., 2017).

As técnicas de *Machine learning* baseiam-se na utilização em dados (*datasets*) e algoritmos por forma a replicar a maneira como os humanos aprendem, melhorando a sua precisão ao longo do tempo em função da quantidade de informação e dados disponíveis. O seu funcionamento advém da capacidade de reconhecer padrões complexos e tomar decisões, com base no conhecimento adquirido no passado (Han e Kamber, 2012).

No que diz respeito ao funcionamento do sistema de aprendizagem de um algoritmo, Berkeley, 2020 delineou o mesmo em três partes fundamentais.

1. Processo de decisão: Os algoritmos de ML são utilizados na generalidade, com o propósito de previsão ou de classificação. Esta decisão é tomada em função de uma estimativa calculada sobre um padrão identificado nos dados.
2. Função de erro: Esta tem como função validar a predição do modelo, recorrendo à existência de exemplos conhecidos o que possibilita à função de erro efetuar comparações de forma a avaliar a previsão.
3. Processo de otimização do modelo: Por forma a minimizar a discrepância entre o exemplo conhecido e a estimativa do modelo, o algoritmo ajusta a sua árvore de decisão em função do crescimento dos exemplos conhecidos.

Na Figura 1 podemos observar os algoritmos aplicáveis à classificação de texto.

O funcionamento dos algoritmos estão genericamente divididos em três categorias principais:

1. *Machine learning* supervisionado:

   É caracterizado pelo uso de *dataset* anotados para treinar algoritmos que classificam dados ou preveem resultados com precisão. Alguns dos métodos utilizados são Redes Neuronais, *Naïve Bayes, Logistic Regression, Decision Tree, SVM* entre outros.

2. *Machine learning* não supervisionado:

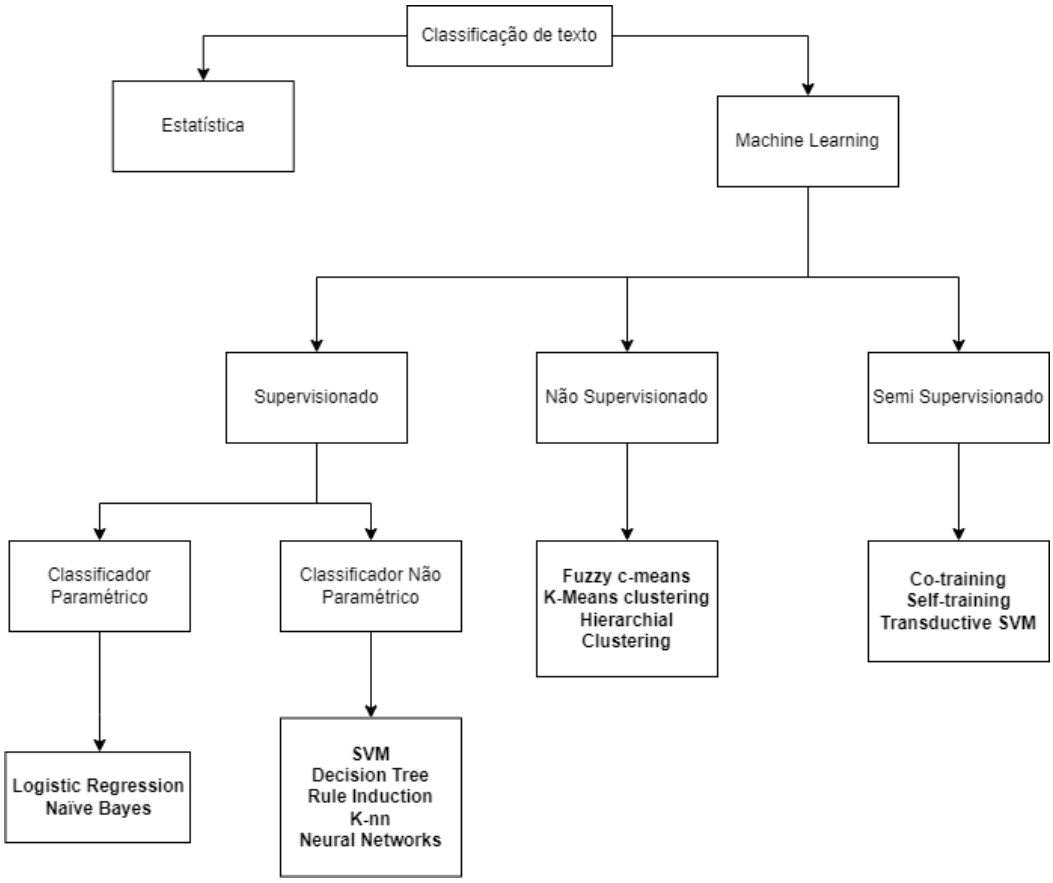


Figura 1: Classificadores de texto

Faz uso de algoritmos de ML para analisar e agrupar *dataset* não anotados. Estes identificam padrões ocultos ou *dataset* sem qualquer intervenção humana. Alguns dos métodos utilizados são *Fuzzy c-means, Hierarchial Clustering*, entre outros.

3. *Machine learning* semi-supervisionado:

   Esta categoria consiste numa simbiose das duas anteriores, sendo que na fase de treino tira partido dos *dataset* anotados. Contudo, o fator diferenciador reside na capacidade de resolução de problemas como a não existência de dados anotados suficientes para o treino de um algoritmo de aprendizagem supervisionado.

Entre os algoritmos de classificação supervisionada, existem duas categorias fundamentais: os algoritmos paramétricos e não paramétricos. Os algoritmos de classificação *Logistic Regression* e *Naïve Bayes* são paramétricos e os mais utilizados (Tsangaratos e Ilia, 2016). Os algoritmos *SVM, Decision Tree, Rule Induction, K-NN e Neural Networks*, correspondem à classificação não paramétrica (Aliwy e Ameer, 2017).

Relativamente à aprendizagem não supervisionada, os algoritmos mais utilizados consistem no *Fuzzy c-means, K-Means clustering* e *Hierarchial Clustering*. Já a classificação semi-supervisionada inclui o *Co-training, Self-training, Transductive SVM* e ainda métodos tendo por base grafos (Calma et al., 2018);

2.1.1.1 *Algoritmos de aprendizagem supervisionada*

Esta secção descreve alguns dos algoritmos de aprendizagem supervisionada que são mais utilizados, nomeadamente *Logistic Regression*, *Naïve Bayes*, *SVM*, *Naïve Bayes* e redes neuronais.

Logistic Regression

É um modelo estatístico utilizado para determinar a probabilidade de um dado evento acontecer, com base num conjunto de variáveis explicativas continuas e/ou binárias. A sua aplicabilidade, dado o seu caráter preditivo, sendo aplicável desde a deteção de fraudes bancárias, até aplicações em *Marketing* (Kleinbaum et al., 2002).

Naïve Bayes

O algoritmo de *Naïve Bayes* consiste em encontrar uma probabilidade num evento, tendo por base a ocorrência de um prévio evento, correspondendo assim à probabilidade condicional (Rish et al., 2001). Pode ser útil na deteção de Spam (dado um e-mail, prever se é spam ou não), Diagnóstico médico (de acordo com os sintomas, prever se o paciente tem determinada doença), entre outras aplicabilidades.

Support Vector Machines (SVM)

Este método faz uso da analise de regressão e classificação procurando reconhecer padrões a partir de um conjunto de dados. Sendo um classificador binário, quando lhe é fornecido um conjunto de dados, este consegue identificar a qual das duas classes esse conjunto pertence (Duan e Keerthi, 2005). O diagrama na figura 2 mostra como os vetores de suporte pertencentes a duas classes diferentes (vermelho versus azul) são separados usando o limite de decisão com base na margem máxima.

Redes Neuronais

As redes neuronais são modelos compostos por neurónios artificiais interligados, que formam um sistema baseado no Sistema Nervoso Central (SNC). De acordo com o autor Haykin, 2009, as redes neuronais, em específico os neurónios, recebem/percecionam entradas e processam-nas produzindo respostas. Têm capacidades de reconhecer padrões e conseguem resolver problemas que os métodos baseados em

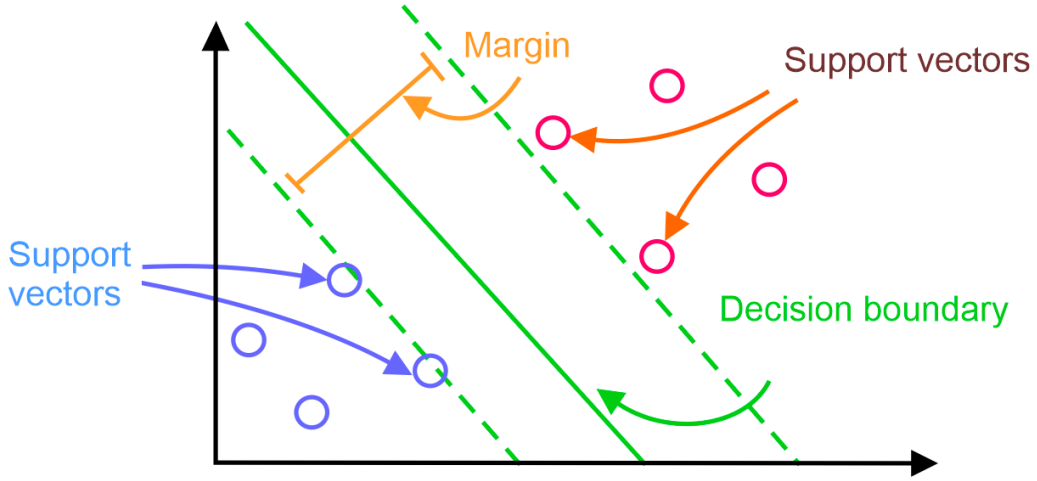


Figura 2: Visão global do método de classificação *SVM*. Fonte packt, 2021

regras não podem resolver. A figura 3 apresenta a comparação entre um neurónio biológico e um neurónio artificial simples (*percetrão*).

2.1.1.2 *Algoritmos de aprendizagem não supervisionada*

A presente secção irá apresentar alguns dos principais e mais usuais algoritmos de aprendizagem não supervisionada, sendo estes *Clustering*, *Fuzzy c-means*, e *Hierarchial clusterin*.

Clustering

Os algoritmos de *clustering* baseiam-se no agrupamento de dados disponibilizados de acordo com similaridade de um determinado critério, formando grupo de dados (*clusters*). É bastante usado sempre que seja necessário encontrar padrões inesperados no *dataset*, (Madhulatha, 2012). A figura 4 apresenta um exemplo do resultado do algoritmo.

Fuzzy c-means

O algoritmo *Fuzzy c-means* (ou *Soft Clustering*), é uma variação do algoritmo *clustering*, que por sua vez permite resolver incertezas. Isto é, dados que possuem características de dois grupos diferentes, podem pertencer a mais de um grupo (*cluster*). A figura 5 apresenta uma comparação entre os algoritmos. O algoritmo *clustering* é identificado como *Hard Clustering*, pelo que é possível analisar que este contém no *cluster 1* apenas o valor *1*, enquanto que o agrupamento efetuado pelo algoritmo *Fuzzy c-means* possui os valores compreendidos entre *0.5 e 1*.

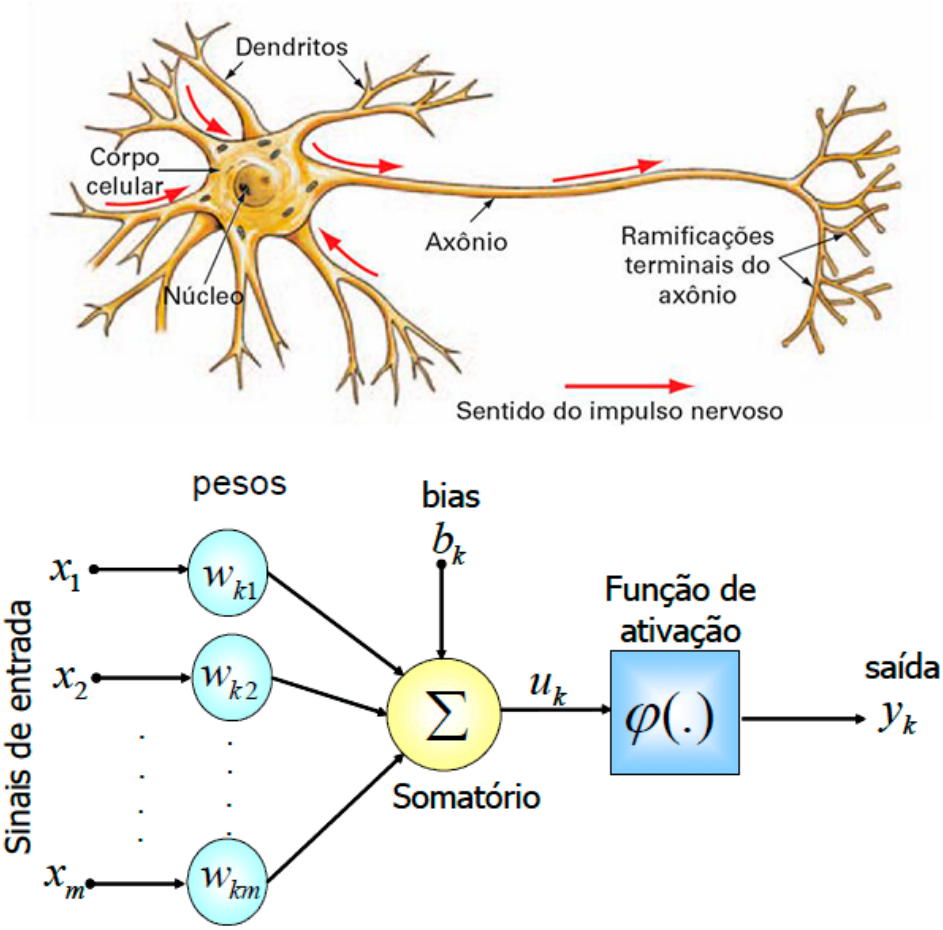


Figura 3: Comparação entre neurónio biológico e neurónio artificial, fonte Haykin, 2001 e Soares e Silva, 2011

Hierarchial clustering

O *Hierarchial clustering* é mais uma variação do algoritmo *clustering*, que tal como o nome sugere, é um algoritmo que constrói uma hierarquia de *clusters* (Azank e Corrêa, 2022). Isto é, uma árvore de *clusters*, conhecida como *dendrogram*. Nesta estrutura, cada *cluster* pode conter outros *clusters*, denominados filhos. Se um *cluster* não tiver nenhum filho, este é denominado de uma folha do *dendrogram*. A figura 6 apresenta um exemplo de *dendrogram*.

2.1.1.3 *Algoritmos de aprendizagem semi-supervisionada*

Por último, serão apresentados os algoritmos mais comuns aplicados na aprendizagem semi-supervisionada.

Co-training

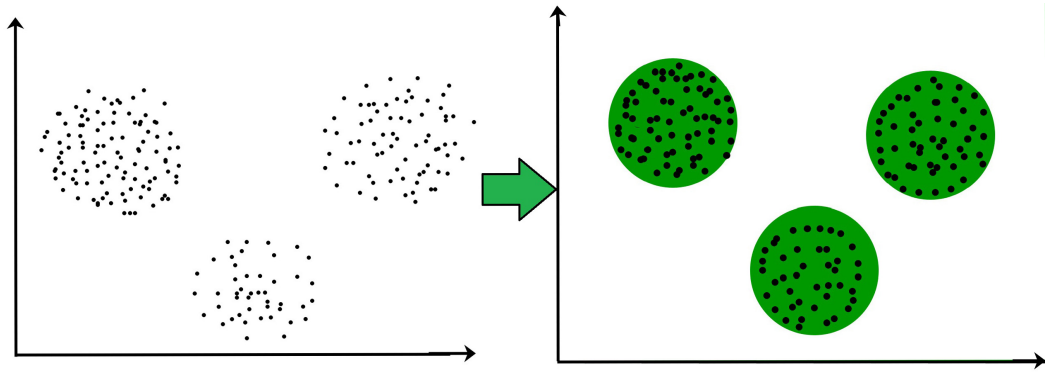


Figura 4: Exemplo resultado algoritmo *clustering*, fonte Priy, 2021

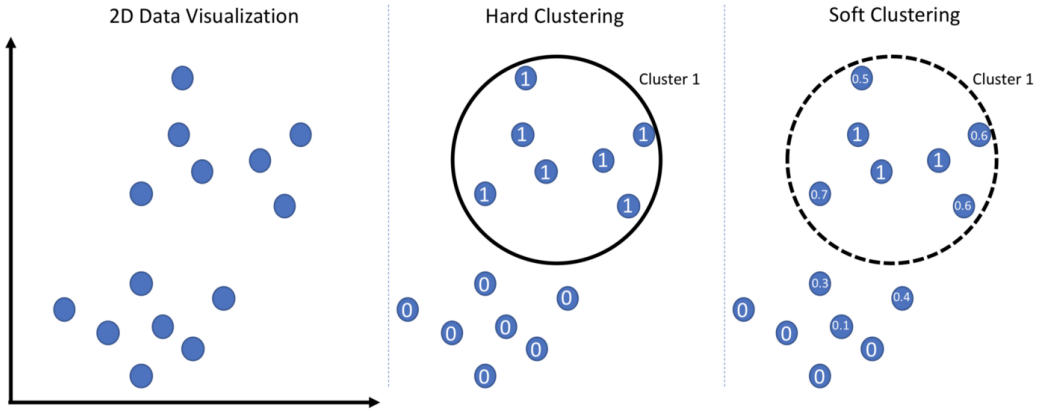


Figura 5: Comparação entre *Clustering* e *Fuzzy c-means*, fonte Yufeng, 2021

O fator diferenciador deste algoritmo, prende-se com o facto de apenas necessitar de uma pequena parte de informação classificada (*labeled*). Na prática, este resolve a escassez de classificação, aplicando dois algoritmos de ML supervisionado, ou o mesmo algoritmo, para induzir duas hipóteses (classificadores). Cada hipótese é induzida através do subconjunto de treino.

Self-training

Consiste num algoritmo treinado inicialmente apenas nos dados classificados, mas que, a partir do output do modelo, é novamente treinado de maneira iterativa, usando dados pré-rotulados e dados que tenham sido pseudo rotulados em iterações anteriores do algoritmo.

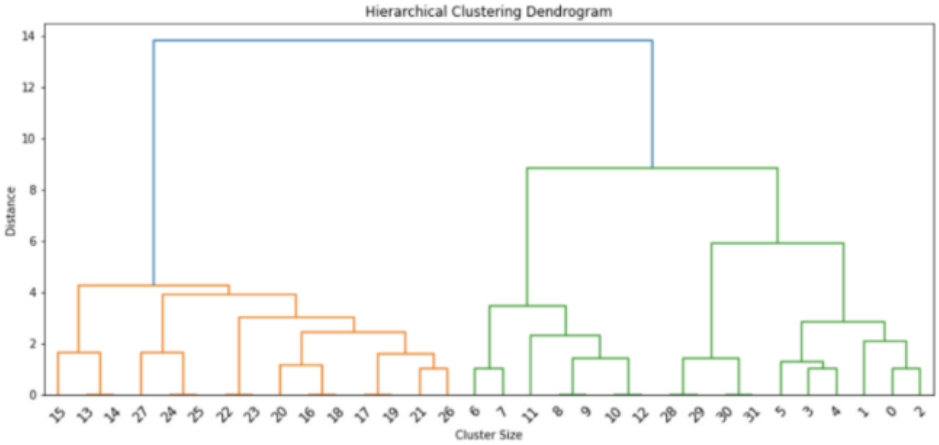


Figura 6: *Hierarchial clustering dendrogram*, fonte Azank e Corrêa, 2022

2.1.2 *Métricas de avaliação de algoritmos*

Existem diversas métricas como a precisão, exatidão, recall, especificidade, AUC e F1, utilizadas para avaliar os resultados de um algoritmo, por forma a perceber se este foi preciso nas suas previsões. O processo de cálculo aritmético de cada métrica utiliza os seguintes valores:

- Verdadeiro Positivo (VP), da designação inglesa True Positive (TP) - corresponde ao número de exemplos da classe positiva corretamente previsto.
- Verdadeiro Negativo (VN), da designação inglesa True Negative (TN) - corresponde ao número de exemplos da classe negativa corretamente previsto.
- Falso Positivo (FP), da designação inglesa False Positive (FP) - corresponde ao número de exemplos da classe positiva incorretamente previsto.
- Falso Negativo (FN), da designação inglesa False Negative (FN) - corresponde ao número de exemplos da classe negativa incorretamente previsto.

A Precisão (P) de um determinado algoritmo corresponde à proporção de valores da classe positiva corretamente prevista e é calculada através da equação (1):

$$P = \frac{VP}{VP + VN} \tag{1}$$

A exatidão (E) considera-se a métrica mais importante, pois denota a precisão do algoritmo, sendo calculada através da equação (2):

		Valor Real	
		Positivo	Negativo
Valor Previsto	Positivo	TP	FP
	Negativo	FN	TN

Tabela 1: Matriz de confusão

$$E = \frac{VP}{VP + FP} \tag{2}$$

A métrica *Recall* (R) corresponde à taxa de sucesso de previsão da classe positiva e pode ser calculada através da equação (3):

$$R = \frac{VP}{VP + FN} \tag{3}$$

A média harmónica entre a Precisão e o *Recall* designa-se por F1 e pode ser calculada através da equação (4):

$$F = \frac{2 * P * R}{P + R} \tag{4}$$

A *Area under the ROC Curve* (AUC) mede a área bidimensional abaixo de toda a curva *Receiver operating characteristic* [1] (ROC). Um modelo cujas previsões sejam 100% erradas terá um AUC de 0, enquanto que um modelo com previsões de 100% corretas terá um AUC de 1, pelo que se pode inferir a precisão do algoritmo em estudo através desta métrica (Google, 2022).

Após obtenção destas métricas é importante analisar a matriz de confusão ou *confusion matrix*, que consiste numa matriz NXN utilizada para avaliar o desempenho de um modelo de ML, onde N corresponde ao número de classes alvo. A matriz compara os valores alvo reais com os previstos pelo modelo de ML. A tabela 1 ilustra a matriz de confusão para uma classificação binária (positivo e negativo).

1 As curvas ROC são uma forma de representar a relação, normalmente antagónica, entre a sensibilidade e a especificidade de um teste diagnóstico quantitativo, ao longo de um contínuo de valores de "*cutoff point*".

2.1.3 *Descoberta de Conhecimento em Bases de Dados*

Formalizado em 1989, com o objetivo principal de extrair conhecimento a partir de grandes base de dados (Usama, 1996). O autor Fayyad et al., 1996 apresenta uma sequência de nove etapas que, segundo o mesmo, constituem um processo interativo e iterativo. Neste contexto, as etapas mencionadas são as seguintes:

1. Definição do domínio de aplicação e objetivos
2. Selecionar a base de dados
3. Pré-processamento e limpeza dos dados
4. Redução e transformação dos dados
5. Seleção do método de *Data Mining*
6. Seleção do(s) algoritmo(s) de *Data Mining*
7. Identificação de padrões (*Data Mining*)
8. Interpretação dos padrões identificados
9. Consolidação do conhecimento obtido

De acordo com Han e Kamber, 2012, a extração de conhecimento de bases de dados afigura-se como uma sequência iterativa constituída por sete etapas, ilustradas na Figura 7, onde podemos observar que estas são na verdade incorporadas nas nove anteriormente identificadas. A maior diferença reside na subdivisão da etapa de *Data Mining* em três, por parte de Fayyad et al., 1996. Em suma, o processo poderá ser definido e descrito pelas seguintes etapas:

1. Seleção - Selecionar informação relevante para o propósito do estudo;
2. Pré-processamento - Aplicação de técnicas capazes de remover erros, omissões ou até solucionar distribuições de dados não uniformes, garantindo os dados alinhados com o propósito do estudo;
3. Transformação - Transformação da informação de uma estrutura ou formato para um formato estruturado capaz de ser processado pelo sistema de uma forma mais eficiente Kusiak, 2001;
4. Data Mining - Seleção dos métodos, algoritmo(s) por forma a extrair padrões dos dados;
5. Interpretação e consolidação - Análise do conhecimento extraído, em forma de técnicas de representação de conhecimento e visualização

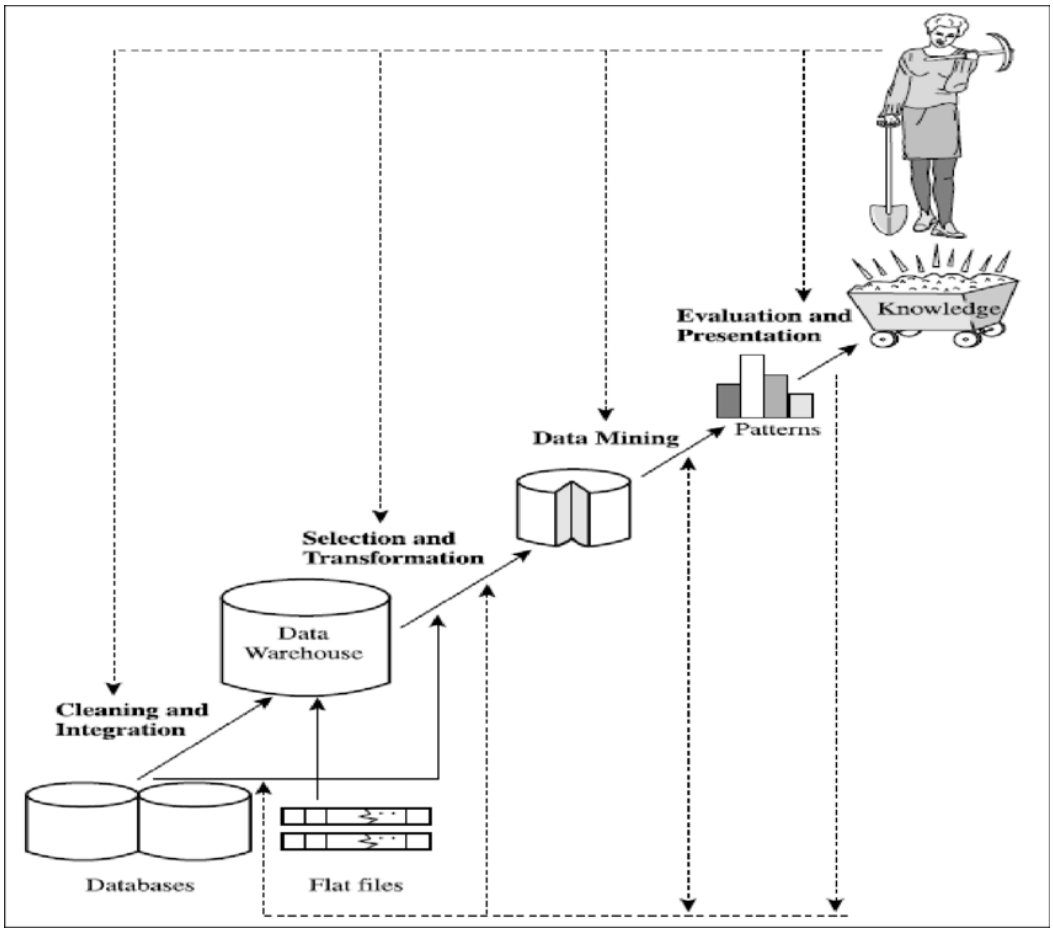


Figura 7: Etapas do processo Descoberta de Conhecimento em Bases de Dados, fonte Han e Kamber, 2012

2.1.3.1 *Pré-processamento e Transformação*

Uma das principais etapas consiste na preparação dos dados para as restantes etapas. De acordo com Kusiak, 2001, o pré-processamento é parte essencial de qualquer análise, uma vez que o resultante desta etapa será passado para todas as etapas posteriores do processo. Para implementar esta tarefa, surgem sub tarefas no âmbito de NLP (*Natual Language Processing*), que permitem a extração e filtragem de entidades e sinónimos, reduzem a dimensão do *dataset*, corrigem palavras com erros ortográficos, entre outras. As sub tarefas, que de uma forma geral detém uma maior utilização, são as seguintes:

1. *Case Folding* - Conversão de todas as palavras para minúsculas ou maiúsculas.
2. Tokenização - Cada frase é decomposta em cada termo que a compõe, sendo geralmente utilizado como delimitador o espaço em branco entre palavras,

quebras de linha, tabulações, e em alguns casos determinados carateres especiais.

3. Remoção de *Stop Words* - Remoção de palavras como artigos, preposições, pronomes, conjunções, etc.
4. Radicalização (*Stemming*) - Cada termo é reduzido ao seu radical.

Ainda dentro do processo, existe uma técnica chamada de *Bag of Words* (*BOW*), que consiste numa matriz atributo/valor que compara a frequência de uma palavra (relevância/peso). Contudo, esta técnica contém alguns problemas conhecidos, como o facto de apenas se limitar a contabilizar o número de palavras no *corpus*, tornando o resultado da técnica *Bag of Words* enorme. E ainda, outro problema que se prende com o facto de não efetuar qualquer correlação com a sintaxe do comentário em análise. Relativamente ao primeiro problema, uma possível solução poderia passar pela integração de *Inverse Document Frequency* (*IDF*). Este método interpreta a importância de uma palavra, aplicando uma pontuação. O *IDF* de uma palavra num *corpus* é calculado pela equação (5):

$$IDF = \log(\frac{N}{nt}) \tag{5}$$

Existe ainda uma derivação desta técnica, designada por *IDF Smoothing*, que garante que as palavras com pontuação 0 na técnica *IDF* não são totalmente suprimidas, sendo o seu cálculo efetuado pela equação (6):

$$IDF = \log(\frac{N}{nt}) + 1 \tag{6}$$

onde N corresponde ao número de itens do *corpus* e nt corresponde à frequência do termo no texto (frase ou comentário).

Relativamente ao facto de não efetuar qualquer correlação com a sintaxe do comentário em análise, uma solução que poderia ser equacionado seria a adoção da técnica de *Word2Vec*. Esta técnica é muito utilizada principalmente em pré-processamento de texto, para realizar tarefas relacionadas com processamento de linguagem natural, tais como análise de sentimentos, tradução de textos, reconhecimento de entidades nomeadas (*NER*), entre outros. Embora seja sem qualquer dúvida muito eficiente, esta carece de muito poder computacional quando comparado com a técnica *Bag*

of Words. Assim, esta deverá ser utilizada de acordo com os objetivos do estudo e quando os resultados da técnica *BOW* forem insatisfatórios.

2.1.3.2 *Data Mining*

Consiste no processo de extração de conhecimento, decorrente da identificação de padrões e/ou relações entre variáveis existentes em *datasets*. Esta representa apenas uma das etapas no processo de descoberta de conhecimento em bases de dados. Todas as etapas anteriores interferem diretamente na qualidade do resultado final obtido pela etapa de *Data mining*, uma vez que uma má execução das etapas anteriores (informações incompletas, informação incoerente com o objetivo em estudo, entre outros), originam uma identificação deficitária e/ou inexistência de padrões e/ou relações entre as variáveis existentes.

Existem diversos métodos de *Data Mining* contudo, devido à sua pertinência, apenas são destacados os mais comuns que são a Previsão e Descrição (Fayyad et al., 1996). A Descrição consiste na descoberta de padrões que permitem a compreensão do conhecimento adquirido. A Previsão advém da descoberta de padrões que permitem prever iterações no futuro Estes métodos encontram-se ainda sub classificados:

- Previsão

  CLASSIFICAÇÃO: é o processo de encontrar um modelo que descreve e distingue classes de dados ou conceitos.

  REGRESSÃO: é o processo de identificar a distribuição das tendências com base nos dados disponíveis.

- Descrição

  CLUSTERIZAÇÃO: é o processo de segmentação de dados que partilham tendências e padrões semelhantes.

  ASSOCIAÇÃO: é o processo capaz de encontrar ações e interligadas com outras ações.

  ANÁLISE DE SEQUENCIAÇÃO: é o processo capaz de identificar uma determinada ação em função de ações anteriores.

  ESTIMAÇÃO: é o processo de efetuar uma pontuação ao invés de efetuar uma classificação binária.

2.1.3.3 *Cross Validation*

O *cross calidation* é uma técnica muito utilizada para a avaliação do desempenho dos modelos de ML. Esta técnica consiste na divisão do *dataset* em partes, onde uma parte é utilizada para treino e a outra parte é utilizada para teste e avaliação do desempenho do modelo. A aplicação desta técnica aumenta a probabilidade de detetar antecipadamente a existência de algum tipo de desajuste na divisão entre treino e teste, designado por *overfitting*. Existem vários métodos de aplicação do *cross validation*, sendo descrito de seguida os métodos *K-fold*, na medida em forma utilizados nesta projeto.

O método *K-fold cross validation* estabelece um divisão do *dataset* de forma aleatória em K [2] subconjuntos (onde K é definido à priori), com a mesma quantidade de amostras em cada subconjunto.

A figura 8, apresenta o exemplo com $K = 5$. Este exemplo, também designado por *5-fold cross validation*, utiliza $K - 1$ subconjuntos para treino e o subconjunto restante é utilizado para teste, gerando métricas de avaliação. Este processo garante que cada subconjunto é utilizado para teste em algum momento da avaliação do modelo.

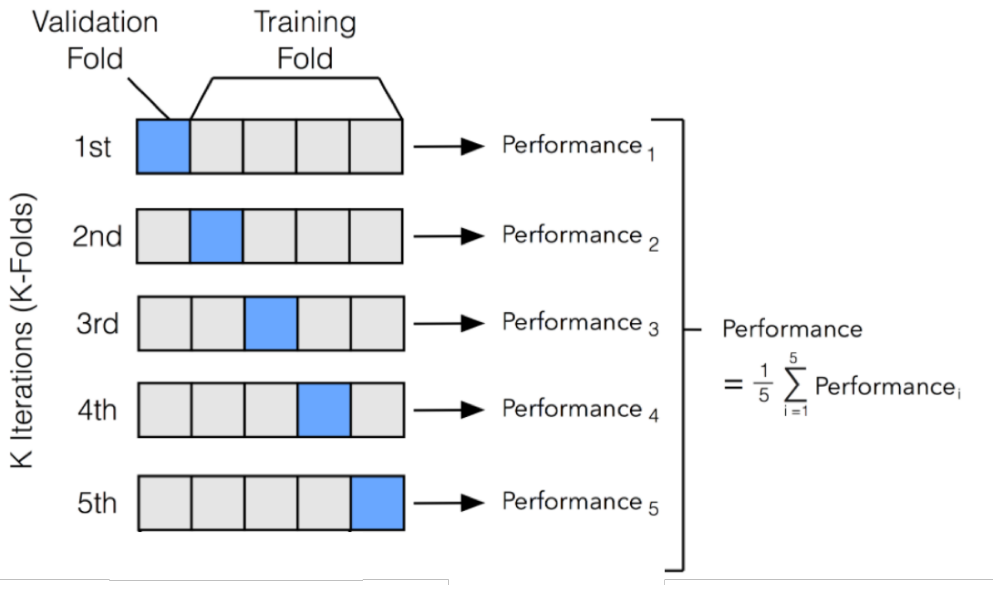


Figura 8: *K-fold Cross Validation*, fonte Andrews, 2020

2 O valor de K deve ser selecionado com atenção afim de garantir que os subconjuntos gerados são suficientemente grandes para representarem estatisticamente o *dataset* original

2.2 ESTADO DA ARTE

Nesta secção são sumarizados seis trabalhos relacionados com o âmbito deste projeto, realçando os principais contributos.

2.2.1 *Deteção de discurso de ódio: Desafios e soluções*

Os autores MacAvaney et al., 2019 identificaram e procederam à análise dos desafios inerentes às abordagens existentes de deteção automática de discurso de ódio em textos. Indicam ainda alguns desses desafios como caraterísticas específicas de cada linguagem, definições divergentes sobre o que constitui discurso de ódio, e limitações de disponibilidade de dados para treino e teste dos sistemas de deteção de discurso de ódio.

Assim, MacAvaney et al., 2019 propõem o modelo de *multi-view SVM* para o classificador de discurso de ódio. O *setup* utilizado partiu de técnicas de pré-processamento (*case-folding*, tokenização e remoção de pontuação),tendo sido utilizados dois *datasets* (*Stormfront* [3] e *TRAC* [4]) no processo de avaliação da precisão dos modelos em teste. No *database Stormfront* o algoritmo *mSVM* atingiu 80% de precisão e um *F-score* 80%. Já no *database TRAC* o algoritmo *mSVM* atingiu 61% de precisão e um *F-score* 53%.

2.2.2 *Uma comparação de algoritmos de classificação para a detecção da fala de ódio*

Os autores Putri et al., 2020 efetuaram uma comparação de algoritmos de forma a detetar de forma automática *tweets* que contenham discurso de ódio na rede social *Twitter*. O estudo comparativo foi efetuado utilizando algoritmos como o *Naïve Bayes, Multi Level Perceptron, AdaBoost Classifier, Decision Tree* e *Support Vector Machine*. Foi igualmente efetuada uma comparação da *performance* dos algoritmos utilizando *SMOTE*, com a finalidade de ultrapassar os dados desequilibrados. Os resultados demonstram que o algoritmo *Multinomial Naive Bayes* apresenta os melhores resultados, tendo obtido um *recall* de 93.2% e uma precisão de 71.2%.

3 https://github.com/Vicomtech/hate-speech-dataset
4 https://sites.google.com/view/trac1/shared-task

2.2.3 *Classificação de um e dois passos para a deteção de linguagem abusiva no Twitter*

Os autores Park e Fung, 2017, apresentaram um estudo comparativo onde pretendiam verificar se existia ou não um incremento de precisão dividindo a classificação de discurso abusivo em duas fases, quando comparado com a classificação em apenas uma fase (classificação como Sexista ou Racista). A abordagem em duas fases classifica inicialmente como linguagem abusiva ou não abusiva e numa segunda fase como Sexista ou Racista. A abordagem de uma fase obteve um resultado de 82.7% utilizando o algoritmo *HybridCNN*, enquanto que a abordagem de duas fases obteve 82.4% utilizado o algoritmo *logistic regression*.

2.2.4 *Deteção automática de discurso de ódio utilizando ML: Um estudo comparativo*

Os autores Abro et al., 2020, apresentam os resultados de um trabalho de investigação, onde comparam oito algoritmos de machine learning (*Logistic Regression, AdaBoost, Naïve Bayes, Random Forest, Support Vector Machines, K Nearest Neighbor, Decision Tree* e *Multilayer Perceptron*). Com objetivo de aferir qual apresentaria melhores resultado na classificação de discurso de ódio, utilizaram o *dataset CrowdFlower* [5] . De forma complementar, utilizaram técnicas de *Feature Engineering* (*TFIDF, Word2vec* e *Doc2vec*) no estudo comparativo. Os resultados obtidos demonstraram como melhor precisão o algoritmo *SVM*, com uma precisão 79% utilizando *TFIDF*.

2.2.5 *Utilização de Redes Neurais Convolucionais para Classificar Discurso de ódio*

Os autores Gambäck e Sikdar, 2017, apresentam um classificador de discurso de ódio fazendo uso de quatro modelos de *Convolutional Neural Network* (*character 4-grams, word vectors based on semantic information built using word2vec, randomly generated word vectors, and word vectors combined with character n-grams*). O classificador atribui cada *tweet* para uma das quatro categorias racismo, sexismo,

5 https://data.world/crowdflower/hate-speech-identification

ambos (racismo e sexismo) e discurso de ódio. O melhor resultado obtido entre os testes executados, foi o *word2vec embeddings* com um *F-score* de 78.3%.

2.2.6 *Deep Learning para deteção de discurso de ódio*

Os autores Badjatiya et al., 2017, apresentam uma abordagem de deteção de discurso de ódio recorrendo a três arquiteturas de redes neuronais. Estas consistiram em *CNN*, *LSTM* e *FastText*. Os resultados obtidos foram consideravelmente bons, onde o método *LSTM* combinado com *Random Embedding* e *GBDT* obteve um precisão de 93% e um *F-score* de 93%.

2.3 SUMÁRIO

Neste capítulo foram apresentados conceitos fundamentais para a compreensão deste projeto, tais como as metodologias de aquisição de conhecimento de base de dados, tendo sido dada relevância à técnica KDD, por ter sido aquela usada. Esta técnica é segmentada em várias etapas, tendo sido cada uma delas analisada e explicada. Foram ainda apresentadas várias técnicas de classificação de texto, o seu modo de funcionamento e áreas de aplicação. De ressalvar, que apenas foram apresentadas as mais usais. O campo da classificação de texto está em constante estudo e melhoria, pelo que as abordagens e técnicas utilizadas estão igualmente em constante alteração.

No processo de revisão da literatura, identificaram-se um conjunto de abordagens muito dispares para o mesmo objetivo, a deteção e classificação de discurso de ódio. Por exemplo, os autores MacAvaney et al., 2019, aplicaram (*multiple view stacked Support Vector Machine*) (mSVM), por outro lado, os autores Putri et al., 2020 efetuaram um estudo comparativo, onde aplicaram *SMOTE* (*synthetic Minority Oversampling Technique*) em conjunto com algoritmos com MLP (multilayer perception), MNB (*multinomial Naïve Bayes*), SVM, e ainda DT (*Decision Tree*). E ainda em estudos mais recentes, a tendência tem sido adotar técnicas de *deep learning*. Os autores de Gambäck e Sikdar, 2017, propuseram o uso do classificador CNN com *word2vec*, por oposição os autores Park e Fung, 2017 sugeriram uma segmentação em duas fases de classificação, combinando assim dois classificadores, o *HybridCNN* e o algoritmo *logistic regression*. Já no estudo Badjatiya et al., 2017 o autor testou três tipos de modelos neuronais diferentes o *CNN*, *LSTM* e *FastText*.

3

CYBERBULLYING

As tecnologias digitais assumem uma grande relevância em termos educativos, informativos, de lazer e sociais. A sua crescente utilização pelas crianças e jovens, em casa, na escola e praticamente em todo o lado, tem-se refletido igualmente em variadas preocupações relacionadas com a privacidade, a segurança, a exposição a conteúdos impróprios, a publicidade ofensiva, entre outros (Seixas et al., 2016).

Neste contexto, surge o *cyberbullying*, entendido como uma agressão intencional, por parte de um indivíduo ou grupo de indivíduos, que recorre repetidamente a aplicações e outras formas de contacto digital, para deliberadamente agredir, perseguir, intimidar, ameaçar, humilhar alguém que não se consegue defender facilmente (Smith et al., 2008; Seixas et al., 2016). O *cyberbullying* requer, assim, a manipulação de tecnologias digitais, com o intuito de perpetuar repetidamente um comportamento hostil, maldoso e agressivo que, intencionalmente, magoa ou prejudica outros utilizadores.

A intensificação da interação com as tecnologias digitais faculta aos utilizadores uma ligação imediata, em qualquer hora e local. A relação interpessoal tem vindo, deste modo, a assumir uma nova configuração, cujo carácter invisível e subtil passa demasiadas vezes despercebido. Devido ao facto de as TIC serem de uso essencialmente individual, aliado à menor competência digital da geração parental, condiciona as possibilidades de controlo e de supervisão no que respeita aos comportamentos online e ao próprio acesso a diferentes conteúdos, sem que exista uma atempada análise do que é adequado a cada faixa etária (Seixas et al., 2016).

Considera-se que a introdução das TIC tem vindo a potenciar novas formas de comunicação e interação, completamente distintas das competências verbais e não verbais (naturalmente desenvolvidas em contextos presenciais). Segundo N. Willard, 2004, os processos de socialização no mundo físico regem-se de acordo com os valores morais e as expectativas sociais (com base nos quais é moldada a conduta humana), o reconhecimento empático de que determinada situação pode suscitar dano a outra pessoa, a desaprovação social de determinada conduta (passível de originar embaraço ou vergonha caso seja praticada), bem como a constatação de que quando algum ato ilícito é cometido, este trará consequências negativas impostas por

figuras de autoridade. Contudo, estes pilares parecem não surtir efeito quando nos referimos aos contextos digitais (*online*). De facto, Seixas et al., 2016, referem que estes contextos favorecem a desinibição dos comportamentos. Ou seja, as pessoas comportam-se de forma mais aberta, mais descontraída e menos constrangida no mundo digital do que no físico, o que pode facilitar o surgimento de condutas ou ações que nunca surgiriam em situações presenciais, como nos casos de uma comunicação mais hostil ou agressiva.

A esta desinibição estão associadas algumas especificidades, nomeadamente:

- O anonimato, uma vez que que a Internet possibilita aos utilizadores permanecerem numa (aparente) obscuridade, possibilitando aos atores das ações escaparem às suas responsabilidades, dado que poderá haver dificuldade em identificá-los;
- A ilusão de invisibilidade, visto que, neste tipo de comunicação, o indivíduo não vê o seu interlocutor (a menos que utilize uma webcam). Assim, pode comportar-se como se estivesse a falar com um ecrã e não com uma pessoa, acreditando que não será identificado. Esta ilusão facilita uma maior divulgação de informação pessoal, comparativamente com situações presenciais (são transmitidas mais informação pessoal em situações em que o emissor é «anónimo visual», ou seja, quando não o conhecemos fisicamente, do que quando este se encontra identificado, visível e em presença);
- A assincronia, uma vez que na maior parte das vezes as pessoas não interagem em tempo real (à exceção dos *chats*), podendo agredir alguém e «desaparecer», por exemplo, deixando um comentário maldoso num sítio e não voltar a visitá-lo, sem saber se a sua ação teve alguma consequência;
- A minimização da autoridade: no mundo online, os indivíduos podem facilmente assumir que a autoridade não existe, que se pode fazer e dizer tudo o que se quer, sem receio de que alguém com autoridade os repreenda ou lhes aplique uma punição.

Além da desinibição, Seixas et al., 2016, os autores referem outras características e propriedades técnicas da comunicação difundidas pelos ecrãs, que se distinguem quando comparadas à comunicação em contextos presenciais:

- A persistência dos conteúdos digitais (textos, imagens, vídeos), considerando que aquilo que é publicado online fica automaticamente gravado e arquivado, independentemente da nossa vontade. "... uma vez na Internet, para sempre na Internet" (Seixas et al., 2016);

- A possibilidade de replicar os conteúdos digitais, ou seja, a informação transmitida online pode ser utilizada por qualquer pessoa (recorrendo ao *copy/paste*) e difundida através da Internet, de variadas formas (mensagens instantâneas, *posts* em redes sociais, entre outros exemplos). Mesmo que um conteúdo seja eliminado depois de ter sido publicado, tal não impede que o mesmo tenha sido copiado ou que uma captura de ecrã tenha sido feita e partilhada;
- A escalabilidade dos conteúdos digitais, que se relaciona com a visibilidade dos conteúdos partilhados como, por exemplo, nos casos de conteúdos que se tornam virais em questão de segundos. A escalabilidade não depende dos utilizadores individualmente, mas do que o coletivo escolhe espalhar;
- A descontextualização, que compreende que tudo o que é publicado online pode ser copiado e publicado de novo, por vezes noutros locais, de modo completamente descontextualizado. Por exemplo, pode utilizar-se uma fotografia retirada de um perfil pessoal/profissional e postá-la num blogue, onde os utilizadores a comentem de modo ofensivo e difamatório;
- As audiências invisíveis, que se expressam através da publicação de conteúdos que, ainda que de forma seletiva, por não se controlar a sua difusão, poderão ser visualizados por desconhecidos.

3.1 TIPOLOGIA DO *CYBERBULLYING*

O *cyberbullying* afirma-se como um fenómeno muito complexo, sendo que a literatura identifica diversas variações de comportamento (Kowalski et al., 2012 ; Montalvão, 2015; Patchin e Hinduja, 2015; N. E. Willard, 2007), pelo que damos a conhecer as distintas formas de expressão do mesmo.

- Mensagens inflamadas (*flaming*): discussão que, podendo começar presencialmente ou online, tende a evoluir para a agressividade através da Internet, incluindo o envio/receção de mensagens inflamadas, rudes, iradas e obscenas, em privado ou em público. Pode originar autênticas «guerras de mensagens» ou comentários, designando-se então de *flame wars* (Seixas et al., 2016);
- Assédio (*harassment*): envio repetido de mensagens de carácter abusivo, visando chatear, ameaçar e alarmar o destinatário (Montalvão, 2015; Seixas et al., 2016);
- Perseguição (*cyberstalking*): perseguição executada através do envio repetido e persistente de ameaças ou mensagens altamente intimidatórias e intrusivas,

com o intuito de causar medo e ameaçar a privacidade da vítima (Seixas et al., 2016);

- Difamação (denigration): publicar declarações falsas ou espalhar através da Internet rumores e boatos sobre outra pessoa com o intuito de causar dano na reputação (Montalvão, 2015; Seixas et al., 2016);
- Personificação (impersonation): fazer-se passar por outra pessoa no ciberespaço ou usando o seu telemóvel, enviando ou publicando mensagens com o propósito de a deixar ficar mal, comprometendo gravemente a sua reputação e amizades (Seixas et al., 2016);
- Exposição (outing): publicar ou enviar mensagens públicas ou privadas com conteúdos relativos a segredos de outra pessoa, de natureza sensível, privada, íntima ou embaraçosa (Montalvão, 2015; Seixas et al., 2016);
- Artimanhas (trickery): utilizar truques com alguém, com o intuito de obter segredos ou informação embaraçosa para depois a divulgar online (Montalvão, 2015; Seixas et al., 2016);
- Exclusão (exclusion): excluir intencional e cruelmente uma pessoa de um grupo online (Montalvão, 2015; Seixas et al., 2016);
- Sexting: troca de mensagens eróticas com ou sem fotos via telemóvel, *chats* ou redes sociais (Kowalski et al., 2012);
- Happy Slapping: agredir fisicamente uma pessoa com o intuito de gravar a agressão e divulgar a mesma online (Kowalski et al., 2012);
- Photoshopping: adulteração de fotos ou vídeos, com a finalidade de denegrir a imagem da vítima (Patchin e Hinduja, 2015);
- Confession Pages: São páginas na Internet ou grupos privados nas redes sociais, onde os utilizadores podem colocar segredos ou informações de forma anónima, despoletando comentários ofensivos (Patchin e Hinduja, 2015);
- Tagging and Untagging: identificar as vítimas sem estas o pretenderem, de forma a relacioná-las a determinadas afirmações, imagens ou vídeos, denegrindo assim a sua imagem (Patchin e Hinduja, 2015);
- Ameaças físicas (physical treats): Quando as ameaças num contexto virtual se concretizam em ameaças à segurança física e ao bem-estar da vítima (Patchin e Hinduja, 2015).

Assim, segundo Seixas et al., 2016, as ações mais usuais, nomeadamente entre crianças e jovens, são as seguintes:

- Espalhar mentiras, ameaças, humilhações ou fotografias de cariz embaraçoso, recorrendo a diferentes formas de difusão;
- Criar páginas de perfil falsas nas redes sociais;
- Usar blogues para difamar outro(s), recorrendo a diferentes tipos de discurso/conteúdos digitais;
- Roubar os *usernames* e as *passwords* para enviar mensagens de provocação e/ou humilhação aos amigos e namorados;
- Partilhar imagens intencionalmente captadas com o intuito de provocar dano, embaraço ou humilhação.

3.2 PERFIL DOS INTERVENIENTES

Em Mason, 2008 reconhece-se a existência de seis tipos de implicados quando falamos no fenómeno de *cyberbullying*:

- Os agressores pró-ativos: aqueles que praticam as suas ações para atingir um determinado fim, para prejudicar terceiros;
- Os agressores reativos: aqueles que executam uma ação como resposta a uma provocação ou ameaça percecionada (seja esta real ou imaginada);
- As vítimas dos agressores pró-ativos;
- As vítimas dos agressores reativos;
- Os observadores que são parte do problema;
- Os observadores que são parte da solução.

Dada a sua pertinência, iremos de seguida apresentar o perfil dos agressores e das vítimas, bem como o papel dos observadores no *cyberbullying*.

3.2.1 *Perfil dos agressores*

Para Seixas et al., 2016, não existe um perfil definido e homogéneo, no que diz respeito aos agressores, uma vez que as suas características são muito diversificadas. Contudo, os autores identificam algumas características mais comuns:

- Impulsividade e baixa tolerância à frustração;

- Extrema necessidade em dominar os outros;
- Dificuldade em aceitar e cumprir normas e regras;
- Maior tendência para a expressão de comportamentos e atitudes agressivas e/ou violentas;
- Reduzida empatia perante as vítimas das agressões, devido ao facto de não verem, em tempo real, o impacto das suas ações.

3.2.2 *Perfil das vítimas*

Também nas vítimas não se consegue apresentar um perfil devidamente definido e homogéneo, podendo estas apresentar perfis totalmente díspares. A título de exemplo, a vítima pode ser alguém bem-sucedido e devidamente integrado a nível profissional, mas que, a dada altura, acabou por ser humilhado ou ameaçado por um agressor. Também a mudança de interações dentro de determinado grupo pode provocar situações de exclusão de determinado elemento. Existem ainda indivíduos que, por não se sentirem adaptados ao contexto onde estão integrados (laboral, educacional, social), se sujeitam a agressões, humilhações, fazendo praticamente tudo o que está ao seu alcance para pertencerem ou, pelo menos, não serem afastados do grupo.

Não obstante tudo isto, e concretamente no caso em que as vítimas são crianças e jovens, as mesmas revelam dificuldade em gerir as suas relações interpessoais, nomeadamente no que respeita à assertividade e defesa dos seus direitos fundamentais. De igual forma, são crianças e jovens que denotam una certa dificuldade em criar e/ou manter amizades, possuindo uma rede de apoio social relativamente frágil ou, por vezes, mesmo inexistente, em contexto escolar e fora dele (Seixas et al., 2016).

3.2.3 *Papel dos observadores*

Quando se aborda a questão do *cyberbullying*, é sobre o papel dos observadores que menos sabemos, nomeadamente porque podem assumir características muito distintas (Seixas et al., 2016). De facto, podem optar por ter comportamentos em que se tornam parte ativa da agressão ou, pelo contrário, protagonizar ações pró-ativas que funcionem como fatores de proteção, evitando perpetuar atos de humilhação, de exposição pública ou até de agressão a outro. Seixas et al., 2016 salientam que não fazer nada quando se é confrontado com uma situação de *cyberbullying* é, na realidade,

uma tomada de decisão, visto que com este tipo de procedimento pode-se contribuir para o fortalecimento e a legitimação dos atos do agressor. Outras vezes, dependendo do seu comportamento, o observador pode também tornar-se um ciberagressor, ao prolongar a extensão do ato agressivo, ao reencaminhar, por exemplo, um e-mail difamatório, aumentando assim o potencial número de espetadores. Assim, aquele que de início era unicamente observador pode com facilidade transformar-se num agressor de segunda linha, ampliando ou perpetuando o ataque inicial (Seixas et al., 2016).

3.3 CONSEQUÊNCIAS DO *CYBERBULLYING*

No *cyberbullying*, o agressor pode ser qualquer pessoa. Tanto pode ser uma pessoa que se encontra bastante perto, como alguém que se encontra a muitos quilómetros da vítima. Pode ser alguém conhecido, de contacto frequente ou até alguém que nunca se tenha visto antes. O fator do desconhecimento de quem é o agressor aumenta consideravelmente os níveis de pressão psicológica, ansiedade e medo nas vítimas (Seixas et al., 2016).

Assim, a inexistência de um lugar verdadeiramente seguro assevera-se um motivo de grande stresse para as vítimas, tendo estas a sensação de estarem constantemente a ser observadas e perseguidas, podendo ser alvo de algum tipo de ações ofensivas por parte do agressor a qualquer momento, em qualquer local. Dado o fácil acesso às TIC, o bullying deixou de «tirar férias» e começou a fazer «horas extraordinárias» (Seixas et al., 2016), ao desenvolver os seus atos no mundo virtual. O *cyberbullying* tem sido associado a uma série de consequências nefastas. Seixas et al., 2016, mencionam o estudo de Elgar et al., 2014 para referir que a frequência de comportamentos de *cyberbullying* está positivamente associada a uma série de problemáticas, nomeadamente sintomas de saúde mental internalizantes (ansiedade, depressão, automutilação, ideação suicida e tentativas de suicídio), problemas externalizantes (agressividade e vandalismo) e consumo de substâncias (consumo de álcool, de tabaco e de drogas).

Esta associação difere na sua configuração, consoante se esteja a considerar cibervítimas ou ciberagressores. Assim, as vítimas de *cyberbullying* são as que tendem a evidenciar níveis superiores de depressão, medo, baixa autoestima, sentimentos de raiva e frustração, impotência, nervosismo, irritabilidade, perturbações do sono e dificuldades de concentração que, em última estância, comprometem o desempenho académico. Seixas et al., 2016, afirmam que a consequência mais grave inerente ao

cyberbullying é o suicídio. Entre as repercussões associadas à ciberagressão salientam-se um menor bem-estar psicológico, os problemas psicossociais e de conduta dos jovens (violência e delinquência), ansiedade, uma considerável ausência de empatia, comportamento agressivo e criminal, maior consumo de álcool e drogas, dependência de tecnologia e absentismo/abandono escolar.

A maioria dos alunos acha que o *cyberbullying* que recorre a imagens e vídeos é mais penoso do que outras formas de *cyberbullying*, devido à larga audiência, acrescendo o facto de serem facilmente identificadas as vítimas. Deste modo, depreende-se que o dano vivenciado no *cyberbullying* assume, essencialmente, uma natureza social e emocional (Seixas et al., 2016).

Neste contexto, foram identificadas as emoções de ambos os lados (vítima e agressor) no estudo efetuado por António et al., 2020, que se revelam inquietantes. As emoções referidas pelos estudantes que foram agressores foi de indiferença (29.4%), culpa (15.7%), raiva (13.7%) e alegria (9.1%). Os alunos que foram vítimas identificaram como emoções principais a insegurança (49.7%), raiva (40.6%), tristeza (39.9%). Estes resultados estão retratados na figura 9.

3.4 MOTIVAÇÃO PARA O *CYBERBULLYING*

A intencionalidade deste fenómeno merece ser refletida. Um estudo efetuado por António et al., 2020 identificou que 41.1% inquiridos revela ter praticado *cyberbullying* "por brincadeira", 23.9% por vingança, seguindo-se 10.2% que recorreram ao *cyberbullying* como um meio para se afirmarem.

Também Law et al., 2012, observou que cerca de 95% dos alunos que referem ter estado envolvidos em comportamentos de *cyberbullying*, consideram as suas ações como sendo inofensivas e destinadas a terem piada, ao passo que apenas 5% afirmaram ter tido a real intenção de magoar a vítima. São números alarmantes e que refletem uma realidade dolorosamente precoce na vida das crianças e jovens. Pode depreender-se que a grande maioria dos alunos que exercem *cyberbullying* não reconhece que os seus atos sejam desagradáveis ou que tenham impacto verdadeiramente negativos nas suas vítimas (Seixas et al., 2016).

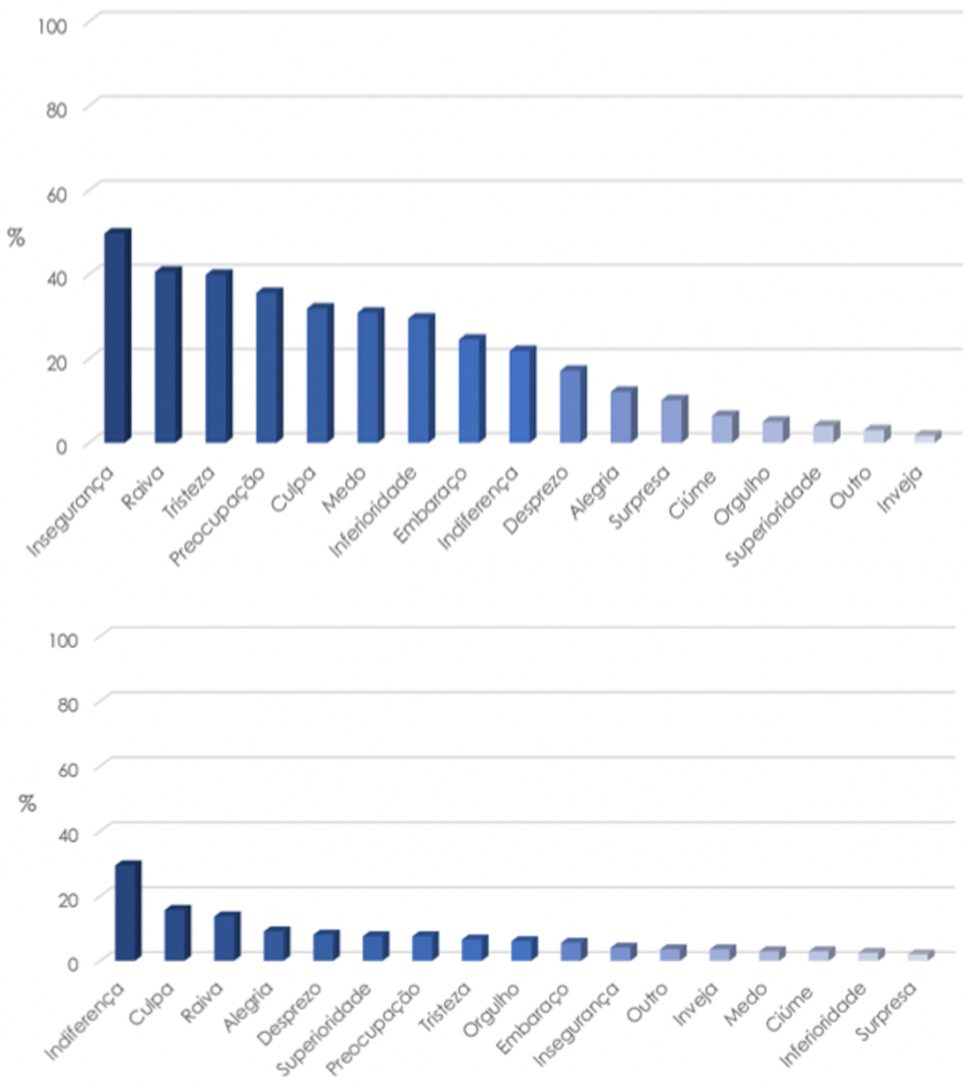


Figura 9: Emoções inquiridos *cyberbullying*, fonte António et al., 2020

3.5 ALGUMAS CONSIDERAÇÕES SOCIODEMOGRÁFICAS

De acordo com Seixas et al., 2016, os resultados das investigações que se têm debruçado sobre diferenças entre os géneros são inconclusivos, às vezes até contraditórios, mesmo quando se considera o mesmo país. Em certos estudos, as raparigas encontram-se mais envolvidas em comportamentos de *cyberbullying*, sendo que o facto de se tratar de um comportamento indireto ajuda a explicar essa maior participação e preferência por esse tipo de agressão. No entanto, noutras investigações, têm sido os rapazes a expressar um maior envolvimento. Considerando os hábitos/comportamentos online, verificam-se, contudo, algumas diferenças: enquanto

as raparigas recorrem ao computador para atividades de socialização, os rapazes utilizam mais frequentemente o computador para pesquisar e para jogar. Dado que as raparigas são mais ativas em redes sociais, salas de chat e blogues, é natural encontram-se em maior risco de serem vitimizadas pelos pares na Internet, em comparação com os rapazes. Em contrapartida, alguns estudos revelam que as raparigas têm maior probabilidade de cometerem *cyberbullying* se tiverem um perfil numa rede social (Seixas et al., 2016).

Relativamente à idade, os estudos têm confirmado uma maior incidência destes comportamentos durante a fase da adolescência. Este facto justifica-se, não somente pela maior autonomia e habilidade para utilizarem as tecnologias digitais, mas também porque, nestas idades, o uso das tecnologias afirma-se como uma oportunidade suplementar de socialização. Seixas et al., 2016, referem que se observa um significativo aumento dos níveis de prevalência até ao secundário (atingindo, aproximadamente, um pico por volta dos 14 anos/ 9° ano de escolaridade), com estabilização ou descida progressiva a partir dos 15 anos. Pese embora não exista muita investigação sobre o assunto, a experiência dos autores indica que, em alguns casos, crianças e jovens pertencentes a minorias (étnicas, raciais, religiosas, orientação sexual, entre outras), com necessidades especiais (sejam elas de cariz físico, sensorial ou mental), bem como pertencentes a grupos já de si vulneráveis (que se encontrem a viver à guarda do Estado ou de organizações não governamentais ou em lares de acolhimento), podem constituir potenciais alvos de *cyberbullying*. O simples facto de ser considerado «diferente» pelos pares pode ser um catalisador para converter uma criança ou jovem em potencial vítima (Seixas et al., 2016).

3.6 *CYBERBULLYING* NA PANDEMIA COVID-19

A pandemia da COVID-19 implicou um longo período de confinamento, e desde cedo que diversos especialistas alertaram para a alta probabilidade do aumento de ocorrências de cibercrimes, entre eles o *cyberbullying*, muito devido ao facto de uma das únicas formas de comunicar ser por via das redes sociais. No ano de 2020, foi efetuado um estudo [1] que visou analisar a frequência de *cyberbullying* por jovens portugueses durante a pandemia.

O estudo contou com uma amostra de 486 estudantes de todos os distritos de Portugal continental e ilhas. Dos 485 inquiridos, 61,4% afirmou ter sido vítima de *cyberbullying* nos últimos meses que antecederam o inquérito. Cerca de 40,8% dos

1 https://ciencia.iscte-iul.pt/publications/files/private/deffa87e217ae129586ff95bed171a6e

inquiridos revelaram estar no papel do agressor/a e 86.6% no papel de observador/a. O estudo revelou ainda diferenças estatisticamente significativas nos/as alunos/as que foram vítimas, de acordo com o seu estatuto socioeconómico e orientação sexual, como se pode analisar na figura 10.

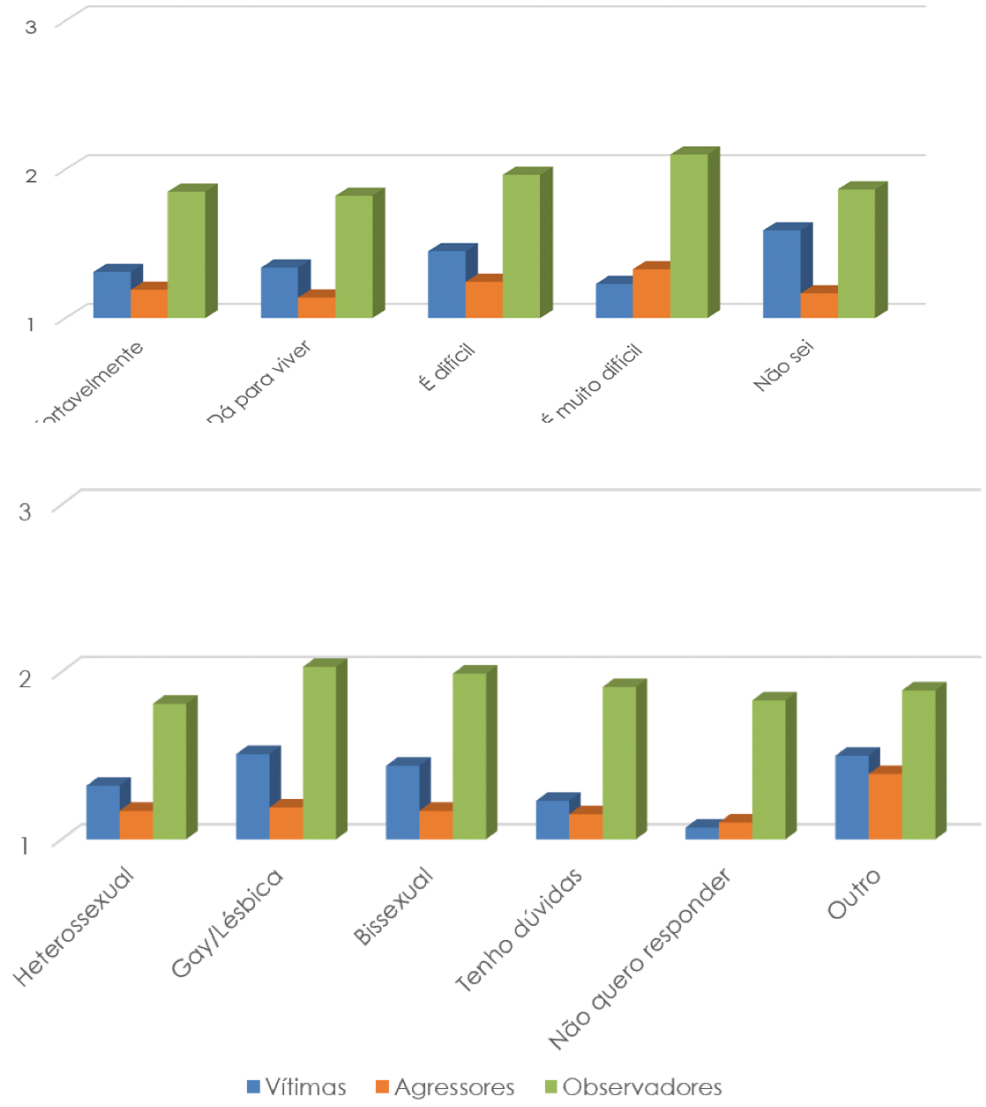


Figura 10: Estatuto socioeconómico e ocorrência de *cyberbullying*, fonte António et al., 2020

3.7 ENQUADRAMENTO LEGAL DO *CYBERBULLYING*

Segundo Seixas et al., 2016, os alunos estão menos cientes das estratégias existentes para solicitar a remoção de conteúdos de alguns *websites* questionáveis, bem como sobre quais as melhores formas de responder de modo útil quando testemunham

um comportamento agressivo *online*. Não obstante a informação existente nas mais diversas aplicações online, em especial nas redes sociais, advertindo sobre a proibição de condutas ilícitas ou ações que vão contra as políticas de utilização definidas, não se verifica que estas metodologias impeçam a prática de agressões virtuais e comportamentos depreciativos entre utilizadores. É de acrescida importância refletir sobre as legislações e organismos que contemplem o *cyberbullying* enquanto flagelo social e cuja atuação tem de ser seriamente considerada, de forma a ser mitigada.

Neste contexto, certos países têm procurado enfrentar o *cyberbullying* através da redação de legislação específica que defina e criminalize os comportamentos inerentes a esta problemática. Outros países, contudo, ainda não preveem legislação específica para o *cyberbullying*, como é o caso de Portugal.

A nível mundial, torna-se pertinente mencionar a Comissão Australiana de Direitos Humanos [2], que encara o *cyberbullying* como uma violação dos direitos humanos em geral, e da criança em particular [3]. Esta perspetiva, que encara o *cyberbullying* como uma violação dos Direitos Humanos e dos da Criança, é igualmente adotada pelo Conselho da Europa [4], cuja disponibilização de diversos recursos permite enquadrar esta abordagem em termos legais e educativos, no quadro da prevenção e combate à violência contra crianças.

Ainda em termos internacionais, os comportamentos e condutas no âmbito do *cyberbullying* podem também ser considerados sob a alçada de tratados internacionais no domínio da criminalidade informática como, por exemplo, a Convenção de Budapeste (Convenção sobre o Cibercrime do Conselho da Europa).

Em termos nacionais, a versão mais recente do Código Penal português (Lei n.º110/2015, de 26/08) contempla uma multiplicidade de crimes que que podem afirmar-se como situações de *cyberbullying*. Ressalva-se ainda que certos comportamentos do *cyberbullying* podem constituir violação de dados pessoais, que poderão cair sob a alçada da legislação nacional.

3.8 APLICAÇÕES DE COMBATE AO *CYBERBULLYING*

Nos anos mais recentes, têm surgido alguns programas e aplicações especializadas, elaboradas no sentido de intervir e prevenir possíveis casos de *cyberbullying*. Limitando-se geralmente a integrar mecanismos de bloqueio e denúncia, a maioria

2 https://humanrights.gov.au

3 https://humanrights.gov.au/our-work/commission-general/what-bullying-violence-harassment-and-bullying-fact-sheet

4 https://www.coe.int/en/web/freedom-expression/guide-to-human-rights-for-internet-users

acaba por se revelar pouco eficaz no que respeita à prevenção. Daí que, recentemente, uma nova onda de programas e aplicações tenha sido desenvolvida, mais direcionadas para as ações de carácter preventivo.

Contudo, este tipo de programas e aplicações ainda não abunda no mercado, os quais incidem sobretudo em aplicações móveis vocacionadas para o uso individual, existindo outras que podem ser adotadas por escolas e agrupamentos escolares. Conseguem encontrar-se algumas delas fazendo pesquisas nas lojas da Google e da Apple. Assim, algumas das referidas aplicações são a Delete Cyberbullying[5] e a KnowBullying[6].

Importa mencionar a existência de linhas de denúncia, como a LinhaAlerta[7] e de ajuda, como a LinhaAjuda[8], que operam telefonicamente, por e/mail e através dos respetivos sítios.

5 https://www.endcyberbullying.net
6 https://healthysafechildren.org/knowbullying-app
7 https://www.internetsegura.pt/lis/sobre-a-lis
8 https://www.internetsegura.pt/linha-ajuda

4

DESENVOLVIMENTO

O fenómeno da propagação de ódio sob as mais diversas formas (mensagens, vídeos, *memes*, entre outros) no ciberespaço é uma realidade em expansão, muito devida à falsa ideia de anonimato que a Internet potencia. A facilidade de contacto com pessoas do outro lado do mundo proporciona uma desconexão da vida real, fazendo com que infringir mal ao próxima seja quase "normal". As consequências para quem é vítima deste tipo de crime são severas e deixam marcas para toda a vida.

A não existência de um *dataset* totalmente em língua portuguesa, demonstra que Portugal ainda se está a adaptar para esta terrível realidade. O presente projeto pretende dar início a algo que poderá vir a ser um grande aliado na prevenção e deteção atempada deste flagelo, através da criação de um *dataset* que poderá ser utilizado para detetar, em tempo real, o discurso de ódio em redes sociais e outras plataformas digitais, como acontece atualmente para outros idiomas.

A recolha de frases e expressões que retratem, tanto discurso de ódio como discurso legitimo, revelou-se uma das tarefas mais complexas deste trabalho, dado que não existem à priori locais onde a língua portuguesa seja o único idioma falado (ou escrito). Assim, após a recolha tornou-se sempre necessário efetuar uma triagem (essencialmente manual), de forma a garantir a consistência de apenas existir conteúdo escrito em língua portuguesa. Esta recolha não foi efetuada apenas num momento, acabou por ser constante durante o desenvolvimento. Foram recolhidas milhares de frases, tendo sido igualmente filtradas e removidas outras tantas, por não corresponderem ao objetivo proposto. Seguiu-se o processo de anotação de todas as frases, construindo assim o *corpus* utilizado na aquisição de conhecimento e consequente classificação do discurso de ódio em língua portuguesa.

O presente capítulo descreve o processo de desenvolvimento prático deste projeto, passando por todas as etapas que deram forma ao classificador, sendo estas:

- A aquisição das frases que compõem o *dataset*, aprofundando todas as etapas inerentes à sua criação:
  - Aquisição de dados
  - Classificação binária (discurso ódio/discurso de legítimo)

  - Pré-processamento
  - Aquisição de conhecimento
  - Apresentação de resultados
  - Desenvolvimento de um classificador de discurso de ódio
- A exploração e apresentação dos algoritmos e técnicas utilizados na etapa *data mining*.
- Apresentação do desenvolvimento da prova de conceito.

4.1 ARQUITETURA

A descoberta de conhecimento em base de dados representa a engrenagem principal e é responsável pela conceção de um classificador de texto com uma precisão o mais próxima possível do resultado expectável, ou seja a deteção de discurso de ódio. O processo de deteção é composto por diversas etapas fundamentais e que devem, tanto quanto possível, ser executadas de forma sequencial, pois em alguns casos existe dependência direta que poderá colocar em causa o desempenho do classificador.

Esta secção descreve a arquitetura aplicada no desenvolvimento da aplicação apresentada neste projeto, que teve em consideração as nove etapas sugeridas por Fayyad et al., 1996. Como identificado anteriormente, a inexistência de um dataset em língua portuguesa, justifica o trabalho de construção e anotação de um para esse fim. Assim é possível analisar na figura 11, que além das nove etapas foram acrescentadas duas que correspondem ao processo de aquisição de dados e correspondente anotação.

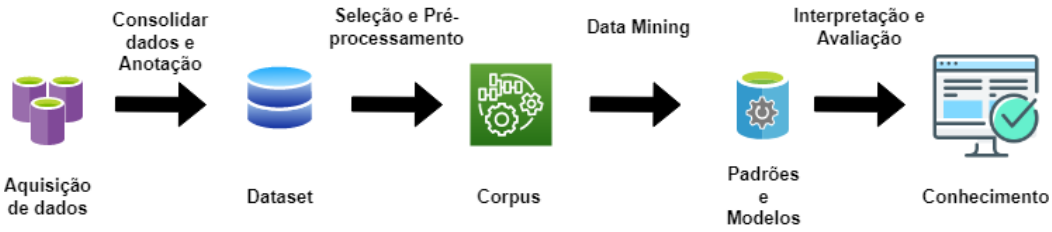


Figura 11: Arquitetura Classificador

4.1.1 *Aquisição de dados*

Uma das tarefas mais complexas da classificação de texto, nomeadamente a sua precisão (independentemente do seu propósito), consiste no *dataset* utilizado. Este deverá ser o mais balanceado possível e devidamente anotado e classificado para uma classificação supervisionada.

Na aquisição de dados para o dataset foram tidas em conta diversas abordagens possíveis, nomeadamente a aquisição de conteúdo proveniente de redes sociais como *Twitter, Facebook, Instagram e YouTube*, por via de *Application Programming Interface (API)* (quando disponível), técnicas de *web scrapping*, passiveis de recolher o conteúdo pretendido. Os programas desenvolvidos para a aquisição dos dados encontra-se disponível no seguinte *URL* `https://github.com/LuisHN/Detetor-Discurso-Odio/tree/main/scraping` e é um dos contributos deste projeto.

Durante o processo de análise dos dados recolhidos, surgiu um problema que viria a tornar-se bloqueante. As frases recolhidas necessitavam de análise individual, pois muitas não estavam em língua portuguesa, outras apenas continham *emojis* e outras não tinham mais do que uma palavra. Após uma análise individual às frases, estas ficaram reduzidas a um subconjunto validado e relevante para a investigação. Por conseguinte, e na tentativa de melhorar o conteúdo dos dados adquiridos até então, foram identificadas entidades nacionais, que poderiam ter uma base de dados sobre esta matéria.

Neste contexto, a organização No Bully Portugal [1], no decorrer de uma reunião, apresentou um livro intitulado "Para cima de puta"da autora Ferreira, 2021, que retrata alguns dos milhares de insultos a si direcionados, através das principais redes sociais. Alguns exemplos desses insultos encontram-se na figura 12. Recorrendo a técnicas de *Optical character recognition* (*OCR*), os insultos no formato de imagem foram convertidos para texto e posteriormente acrescentados ao *dataset*. Os insultos em formato de imagem, bem como o código utilizado, encontra-se disponível no seguinte *URL* `https://github.com/LuisHN/Detetor-Discurso-Odio/tree/main/scraping/book`.

1 https://nobully.pt/

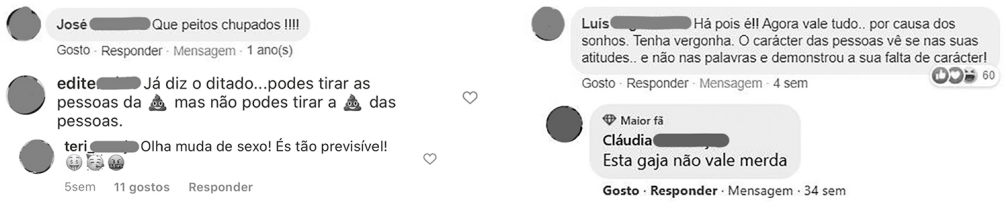


Figura 12: Exemplos insultos, fonte Ferreira, 2021

Fonte de Dados	Discurso de ódio	Discurso legitimo	Total
Redes Sociais	72	151	223
Livro "Para cima de Puta"	131	0	131
Total	203	151	354

Tabela 2: Composição do *Dataset*

4.1.2 *Anotação do dataset*

A anotação do *dataset* é uma tarefa metódica e complexa. Considerando que a dimensão dos dados recolhidos colocaria vários entraves à análise e anotação de cada frase em tempo útil, foi desenvolvido um portal *web*, como retrata a figura 13, com apenas três funcionalidades: eliminar a frase, classificar a frase como discurso de ódio ou classificar a mesma como correspondendo a discurso normal. O portal encontra-se disponível para consulta no *URL* `https://dadol.pt#classificador`, bem como o seu código fonte disponível em, `https://github.com/LuisHN/Detetor-Discurso-Odio/tree/main/frontend`.

O portal foi disponibilizado a um grupo restrito de indivíduos com acesso individualizado, com o intuito de acelerar o processo de anotação, bem como garantir uma classificação assertiva. Cada pessoa, em virtude das suas vivências e aprendizagens, poderá classificar determinada frase como discurso de ódio e outra como discurso legítimo. Assim, a classificação final consistiu no resultado da unanimidade do grupo, e quando tal não se verificasse, a última decisão foi tomada pelo autor.

O resultado obtido no processo de anotação pode ser analisado na tabela 2. O *dataset* contempla um total de 354 publicações, das quais 203 correspondem a discurso de ódio e 151 a discurso legitimo.

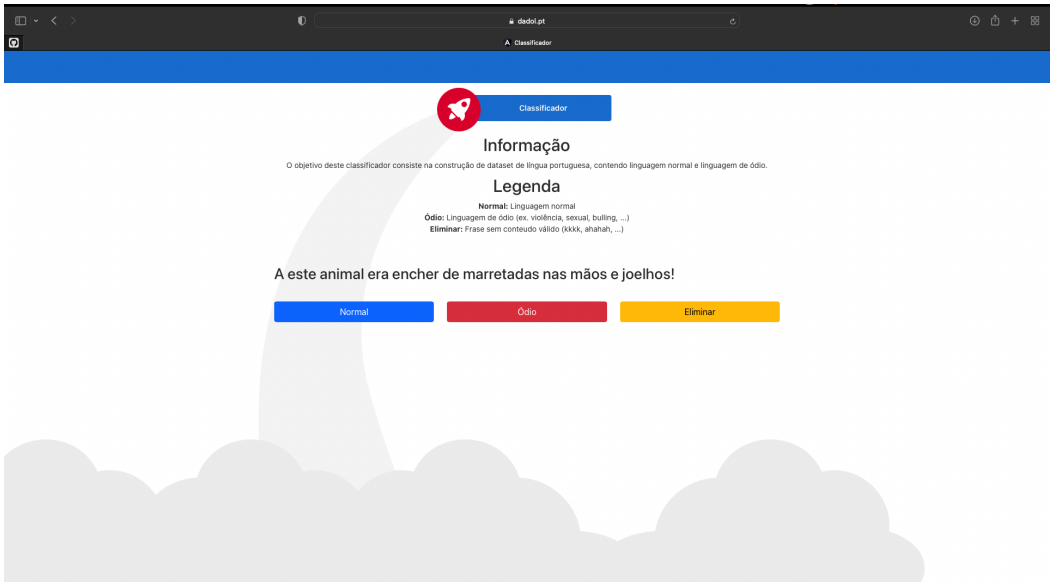


Figura 13: Portal *web*

4.1.3 *Pré-processamento*

A etapa do pré-processamento subdivide-se em duas tarefas complementares, designadamente a limpeza e a transformação dos dados. Como já foi referido, o pré-processamento define a precisão dos algoritmos aplicados na etapa de *data mining*. As tarefas indicadas efetuam uma transformação geral no *dataset*, de forma a constituir um *corpus* completo e consistente, obtendo assim o melhor desempenho e os melhores resultados.

A figura 14 apresenta parte do código *Python* utilizado no pré-processamento, para transformar o *corpus*, nomeadamente a conversão das letras para minúsculas, a remoção de palavras como artigos, preposições, pronomes, conjunções, entre outros e a decomposição da frase em cada termo (a delimitação é efetuada através dos espaços em branco entre palavras, quebras de linha, tabulações, e em alguns casos determinados carateres especiais).

O *dataset* pode conter conteúdo irrelevante ou ausente, isto é, podem existir linhas em branco, *emojis*, carateres especiais, entre outros. Desta forma, a etapa de limpeza de dados normaliza o *dataset*, garante apenas a existência de informações coerentes com o objetivo do estudo. Além desta limpeza inicial, é também essencial transformar o corpus num formato correto para a etapa seguinte.

A titulo de exemplo do resultado obtido no pré-processamento, consideramos o seguinte comentário original: *"Sendo um crime público o ministério público não*

```python
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from nltk import pos_tag
from nltk.corpus import stopwords

def text_preprocessing(text):
    # Step - 1b : Case Folding
    text = text.lower()

    # Step - 1c :Tokenização
    text_words_list = word_tokenize(text)

    # Step - 1d : Remove Stop words, Non-Numeric and perfom Word Stemming/Lemmenting.
    # Declaring Empty List to store the words that follow the rules for this step
    Final_words = []
    word_Lemmatized = WordNetLemmatizer()
    for word, tag in pos_tag(text_words_list):
        if word not in stopwords.words('portuguese') and word.isalpha():
            word_Final = word_Lemmatized.lemmatize(word)
            Final_words.append(word_Final)
    return str(Final_words)

```

Figura 14: Código *Python* Pré Processamento

precisa de apresentação de queixa, bastar-se-a pelas provas, neste caso mais do que evidentes.".

Este comentário é sujeito a uma transformação de todas as palavras para minúsculas, remoção de *URL*, remoção de *stop words* e de números e carateres especiais, e ainda a transformação da palavra para o seu radical. O comentário final, após a transformação, é o seguinte: *"ser crime público ministério público precisar apresentação queixa bastar prova neste caso evidente"*. Por fim é efetuada a tokenização que na prática consiste na criação de um *array* de palavras, conforme se ilustra na Figura 15.

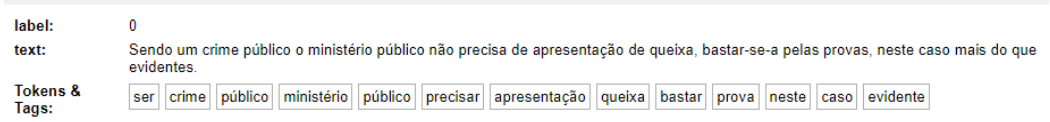


Figura 15: Exemplo pré-processamento

No final do pré-processamento foi aplicada um técnica para transformar o texto num formato numérico. Uma das formas mais simples e populares de o fazer intitula-se *Bag Of Words*, que consiste na representação de cada palavra de acordo com número de ocorrências no *dataset*, fazendo com que as palavras que mais se repetem se destaquem, como é possível visualizar na Figura 16. Tendo em atenção os problemas identificados anteriormente, o *BOW* foi implementado em conjunto

com a variação de *Inverse Document Frequency* (*IDF*), o *Smooth IDF*, considerando o desempenho obtido com esta técnica.

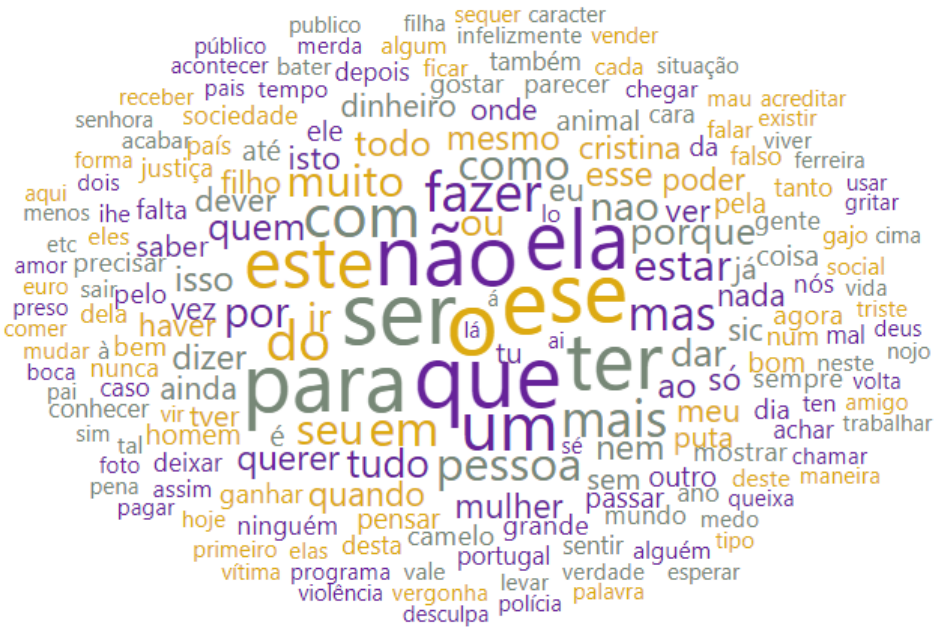


Figura 16: *Word Cloud* gerado após pré-processamento

4.1.4 *Data Mining*

A etapa de *data mining* foi implementada de acordo com o processo apresentado por Fayyad et al., 1996. Considerando o espectro em estudo, e ainda a dificuldade implícita no processo de previsão e classificação de texto, aliada à dificuldade inerente da língua portuguesa, foi necessário explorar diversas técnicas e algoritmos, de maneira a identificar qual seria a que melhor se enquadrava nos objetivos propostos. A implementação desta etapa foi efetuada recorrendo a duas aplicações: o *software Orange* [2] e o recurso a *scripts* recorrendo à linguagem de programação *Python*.

No Orange foram aplicados os algoritmos (*Naïve Bayes, SVM, kNN, Neural Network, Random Forest*), como pode ser observado na figura 17. Paralelamente foram desenvolvidos e executados os scripts que implementam os algoritmos (*Naïve Bayes, SVM e kNN*).

2 https://orangedatamining.com

Ambas as implementações podem ser consultadas no repositório *GitHub*, no seguinte *URL* https://github.com/LuisHN/Detetor-Discurso-Odio/tree/main/classificador.

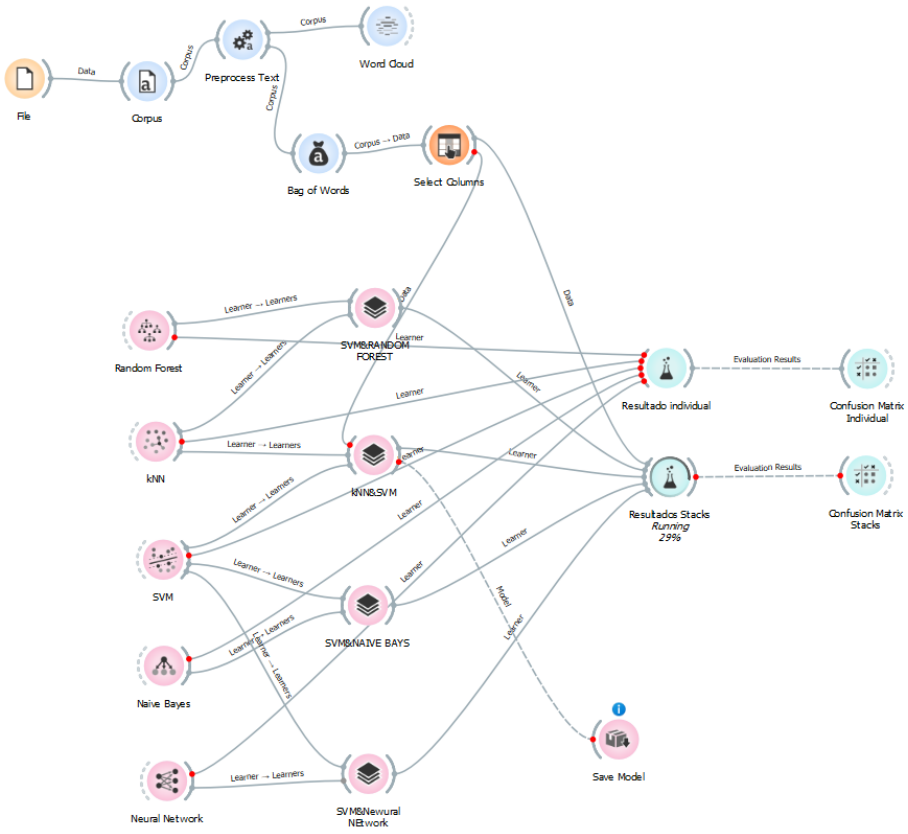


Figura 17: Projeto desenvolvido *Software Orange*

4.1.5 *Algoritmos*

Durante o processo de revisão da literatura e trabalhos relacionados, foram identificados um conjunto de algoritmos e estratégias com resultados promissores, considerando a dificuldade implícita na classificação de texto. Assim, a base de trabalho consistiu na utilização destes mesmos algoritmos e metodologias.

O desenvolvimento e os testes efetuados para a conceção da arquitetura, foram efetuados na aplicação *Orange* pela facilidade visual que oferece. Foi aplicado uma *cross validation k-fold*, com $k = 5$, pelo facto de ser computacionalmente menos exigente e estar em linha com as estratégias de classificação adotadas em vários contextos.

No decorrer dos testes verificaram-se diferenças significativas na utilização ou não da técnica *IDF*. A tabela 3 apresenta os resultados individuais dos classificadores implementados sem o uso da técnica *IDF*. Já a tabela 4 apresenta os resultados obtidos utilizando a técnica *IDF*, onde se observa um incremento generalizado em todos os classificadores. Por sua vez a técnica *Smooth IDF* denota um considerável aumento da performance dos classificadores, como demonstra a tabela 5.

Decorrente dos vários testes efetuados, ficou claro que a utilização da técnica *BOW* e *Smooth IDF* apresenta um desempenho melhor na generalidade dos classificadores. O algoritmo *kNN* usou o número de vizinhos 5, por ser dentro dos testes (kNN 5, kNN 10, kNN 15 e kNN 20) o que obteve melhores resultados. Relativamente ao algoritmo *SVM* este encontra-se configurado com as variáveis $C = 1$, *regression loss epsilon* = 0,1 e *kernel* = *linear*. Já quanto ao algoritmo *Neuronal Network* apresenta a seguinte configuração, *Neuron in hidden layers* = 100, *activation* = *Logistic* e *Solver* = *L-BFGS-B*.

Foi possível identificar que a *AUC* de cada um dos algoritmos é superior a 0,80, o que denota que os modelos efetuam uma classificação binária (0 - discurso legítimo, 1 - discurso de ódio) promissora para o contexto da classificação do texto.

Um dos testes que se revelou bastante interessante, consistiu na aplicação da técnica *Stacking*, onde foram aplicadas várias combinações de algoritmos, pelo que os resultados obtidos, presentes na tabela 7 revelam-se significativamente melhores. A configuração do *Stacking* consistiu na conjugação de modelos de ML (SVM, kNN e Neuronal Network) como *Learner* e o modelo *Logistic Regression* como *Learner Aggregator*.

4.2 PROVA DE CONCEITO

Nesta secção será apresentado todo o processo de elaboração, desde a definição dos objetivos propostos, até ao desenvolvimento técnico da prova de conceito, que visa colocar em prática o classificador de discurso de ódio em língua portuguesa.

Deste modo, foi desenvolvida uma extensão para o navegador Google Chrome, passível de recolher todas as manchas de texto presentes na *tab* ativa do navegador. Esta informação recolhida será enviada para um serviço via *API*, que irá analisar e detetar a presença ou não de discurso de ódio. Tendo em consideração que a extensão recolhe a informação existente no código em formato *HyperText Markup Language* (HTML), foi considerada a possibilidade da extensão também recolher dados sensíveis. Assim, o processo de análise requer uma filtragem e correspondente remoção [3] antes da informação ser armazenada na base de dados. Existe ainda a possibilidade de listar as frases recolhidas e solicitar a sua remoção como é possível analisar na Figura 18.

Considerando o esforço envolvido neste serviço externo, e após uma avaliação do esforço *versus* retorno, tomou-se a decisão de desenvolver uma plataforma integral de deteção de discurso de ódio, capacitada de funcionalidades que lhe conferem a possibilidade de incrementar substancialmente o *dataset* de uma forma simples, precisando apenas que a extensão tenha uma adesão significativa. De facto, num dia normal de utilização, a extensão recolhe uma média de cinquenta mil frases únicas. Claro está que nesta fase é necessária muita intervenção humana, de forma a validar a classificação efetuada pelo classificador e a qualidade das frases. Contudo está aberta a possibilidade para implementar um modelo de ML semi-supervisionado, reduzindo assim a iteração humana no processo de anotação e classificação.

De seguida, serão apresentados os requisitos e funcionalidades da plataforma, a arquitetura da solução implementada, o funcionamento individual dos artefactos que completam a plataforma, bem como os resultados obtidos no decorrer da sua utilização.

3 não estando a cem por cento garantida a totalidade de informação sensível filtrada

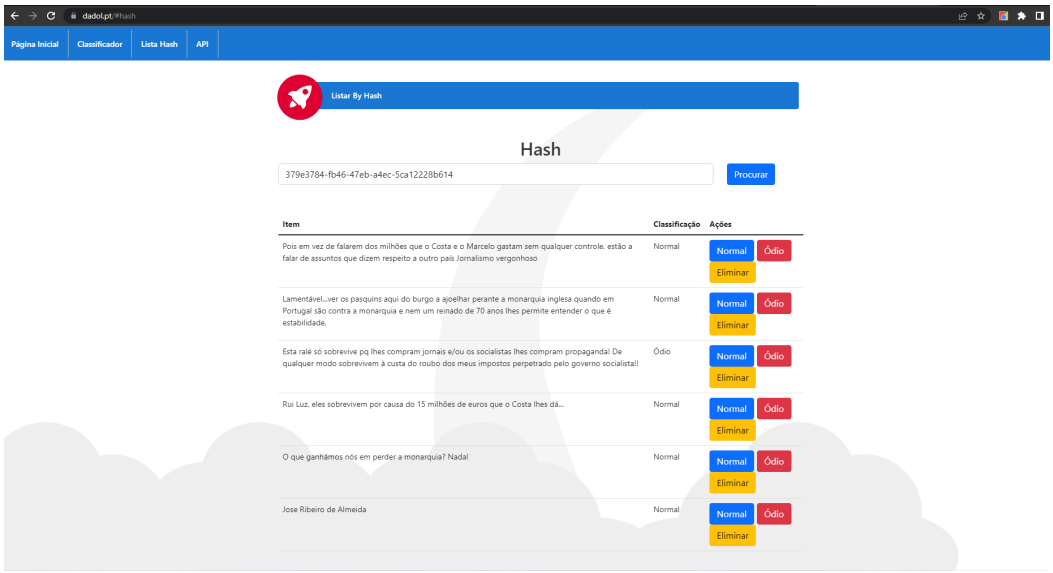


Figura 18: Funcionalidades Portal *Web*

4.2.1 *Objetivos e Funcionalidade*

Esta prova de conceito visa testar a precisão dos modelos de classificação gerados em função do *dataset* recolhido e trabalhado. Contudo, inerente a este objetivo, surgiram novas funcionalidades consideradas bastante interessantes no processo de prevenção do fenómeno em estudo.

Assim, identificam-se as seguintes funcionalidades:

1. Aquisição do conteúdo no navegador Google Chrome através da extensão desenvolvida.
2. Análise e classificação do conteúdo recolhido.
3. Disponibilização dos resultados obtidos.
4. Validação manual das classificações efetuadas, permitindo a sua edição e/ou remoção.
5. Acrescentar novas entradas anotadas no *dataset* e reconstruir os modelos do classificador.
6. Disponibilizar uma *API REST* (*Representational state transfer*) com a capacidade de receber um *input* e devolver em *Near Real Time* a classificação correspondente.

4.2.2 *Arquitetura da Prova de Conceito*

Uma das preocupações durante o desenho da arquitetura desta prova de conceito prendeu-se com a importância de entregar um sistema com respostas rápidas e com facilidade de escalabilidade. Deste modo, a arquitetura implementada segue as tendências atuais no que respeita o desenvolvimento ágil e tira partido da tecnologia *Docker* para um rápido crescimento do sistema.

A Figura 19 representa a solução implementada, sendo a mesma dividida em cinco módulos:

- Extensão de navegador *Google Chrome*.
- Portal *web*.
- *API*.
- Classificador
- *Model Upgrade*.

À exceção da extensão, que está disponível para instalação manual no *URL* `https://github.com/LuisHN/Detetor-Discurso-Odio/blob/main/extension.zip`, e futuramente no *Chrome Web Store*, os restantes módulos encontram-se em funcionamento num (*virtual Private Server*) (*VPS*) alojado na *Oracle Cloud* `https://www.oracle.com/uk/cloud`, estando o seu código disponível no repositório *GitHub* no seguinte *URL* `https://github.com/LuisHN/Detetor-Discurso-Odio`.

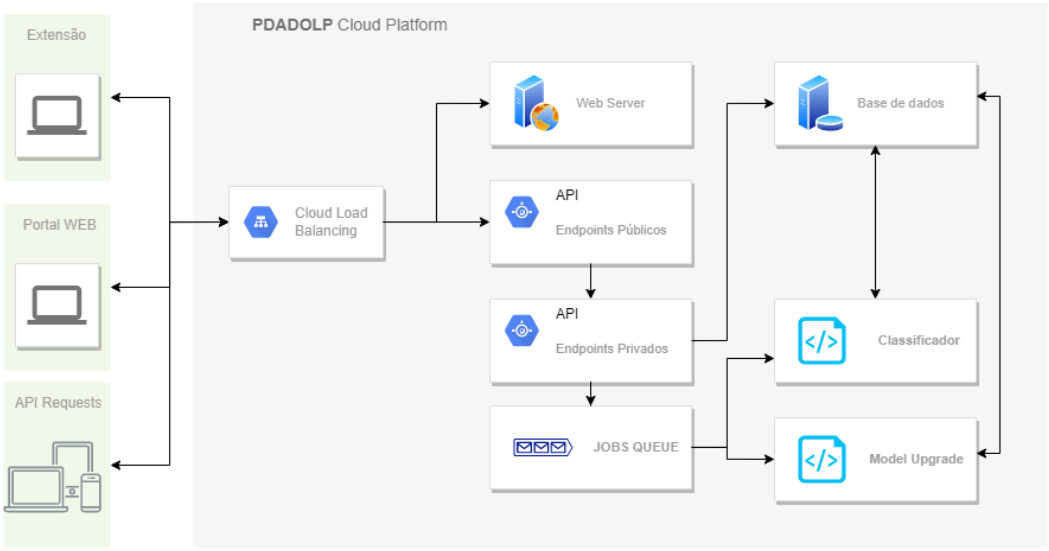


Figura 19: Solução alto nível

4.2.2.1 *Descrição dos Módulos*

Como mencionado anteriormente, existem em funcionamento cinco módulos que, quando combinados, dão origem à plataforma de deteção automática de discurso de ódio em língua portuguesa. Nesta secção, cada um deles será apresentado do ponto de vista técnico e funcional.

A sequência de apresentação dos módulos não têm relação direta com o seu funcionamento interno. assim, o módulo identificado como "Classificador"corresponde a um *script Python*, cuja responsabilidade é a de receber um *input* correspondente a uma frase em língua portuguesa e proceder à sua classificação binária (1 - discurso de ódio e 0 - discurso legítimo).

Esta classificação é guardada na base de dados do sistema desenvolvido, denotando-se o facto de esta classificação não ser definitiva, pois carece de uma validação humana, com recurso ao "Portal *Web*". Tal pode ser observado na Figura 20, antes de estar apta para integrar no *dataset*.

string	clientHash	classification	classification_1	finalClassification	classification_2
És um preto filho da puta	c6401a6a-b091-4faa-a3e3-07c0753e86e6	1	1	1	1
Meu grande paneleiro. devias ter morrido no parto	b2b728f1-9801-4221-bbf3-777873768d01	1	1	1	1
Não quero saber da tua opinião	161a7cc7-d8a9-4bd1-b6e8-0baef474461c	0	<null>	<null>	<null>

Figura 20: Tabela Base de dados

O módulo *API* consiste na integração entre o mundo e as funcionalidades da plataforma, sendo que esta integração expõe apenas o necessário, garantindo assim a integridade dos dados e a estabilidade da plataforma. A *API* foi desenvolvida na linguagem *nodeJS*, utilizando a *framework nestJS*, estando os *endpoints* disponíveis no *URL* `https://api.dadol.pt/api`. Atendendo à figura 21, depreende-se que a mesma se encontra dividida em dois contextos principais:

1. Extensão

   Este contexto é utilizado pelo módulo "Extensão de navegador *Google Chrome*"e apenas são disponibilizados três *endpoints*. Um para devolver um *hash*, identificativo do navegador utilizado no "Portal *Web*", outro para enviar que frases serão classificadas pelo "Classificador"e por fim um que devolve todas as frases já classificadas para um determinado *hash* identificativo.

2. Classificador

   Este contexto é utilizado pelo módulo "Portal *Web*", tendo à sua disposição *endpoints* capacitados para listar frases no estado "pendente de validação", proceder à anotação da frase (1 - discurso de ódio e 0 - discurso legítimo)

e ainda eliminar a frase. Dispõe ainda de um *endpoints* que pode receber uma ou mais frases, para serem classificadas naquele instante.

Figura 21: Swagger API

A funcionalidade de classificação manual de frases possui uma validação interna que garante um mínimo de quatro classificações diferentes para cada frase no estado *"pendente de validação"*. Apenas após este requisito estar garantido, estas frases alteram o seu estado para o estado *"pronto para upgrade"*. Contempla também um mecanismo de filtragem de idioma, fazendo uso da biblioteca *Node Language Detect* [4]. Desta forma reduz-se a probabilidade de serem introduzidas frases ou comentários de idiomas diferentes de língua portuguesa.

O módulo "*Model Upgrade*"tem a responsabilidade de aumentar o *dataset* e consequentemente reconstruir os modelos de ML. Atualmente, este módulo é instanciado de forma manual, considerando a fase embrionária da plataforma. É, no entanto, possível utilizar mecanismos já existentes e criados para serem executados de forma automática pelo sistema. Este módulo procede à seleção de frases no estado *"pronto para upgrade"*, aplicando então todas as etapas de descoberta de conhecimento. No final, os modelos estarão atualizados e operacionais para uso pela plataforma.

O módulo "Extensão de navegador *Google Chrome*"consiste num código que irá correr no navegador do utilizador. A extensão captura todo o texto existente na *tab* em visualização e envia a informação para o *endpoint* disponibilizado para o efeito, como é possível analisar na figura 22. Cada sessão será identificada com uma *hash* devolvida pela *API*. É importante ressalvar que ainda no navegador é efetuada uma filtragem com o propósito de remover alguma informação sensível, sendo esta novamente filtrada na *API* antes de ser armazenada na base de dados.

4 https://www.npmjs.com/package/languagedetect

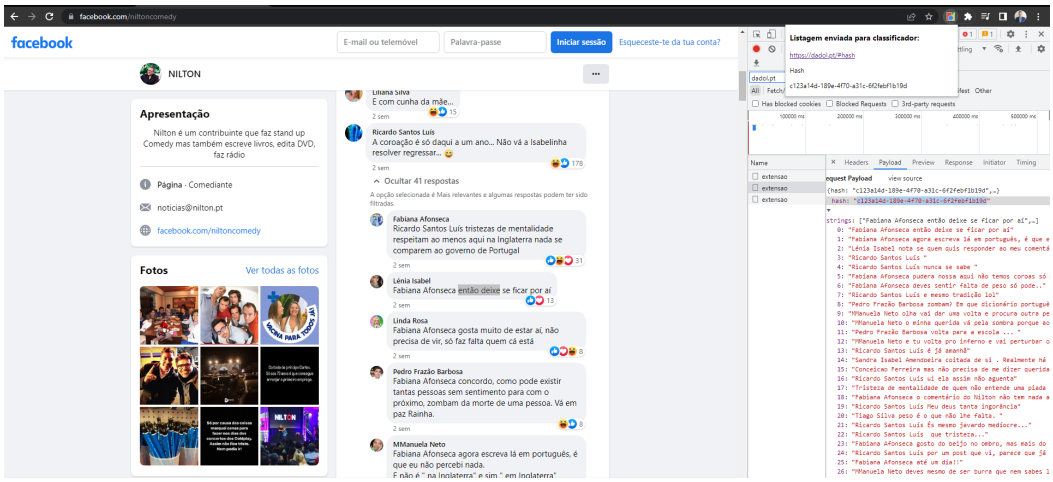


Figura 22: Extensão em funcionamento

Por fim, o "Portal Web"é um portal simples e intuitivo, desenvolvido em *Angular*, que permite uma iteração amigável com os serviços disponibilizados pelo módulo "*API*". O fluxo de funcionamento do portal *web*, tal como é possível verificar na figura 23, realça a existência de duas fontes de dados. A primeira, existente na página principal, permite a inserção manual de novas frases ou comentários para classificação. A segunda fonte de dados possibilita a inserção por parte da extensão do navegador *Google Chrome*, complementada por um classificador manual de frases recolhidas pelas fontes mencionadas anteriormente. Contudo, estas frases ou comentários apenas estarão disponíveis após classificação do utilizador ou automaticamente após seis horas de armazenamento no sistema. A classificação por parte do utilizador é efetuada através da funcionalidade existente no *link* `https://dadol.pt/#hash`. Esta funcionalidade permite classificar as frases que ainda não reúnam as quatro classificações efetuadas.

4.2.2.2 *Infraestrutura*

Para disponibilizar a plataforma online, foi necessário selecionar um fornecedor de *VPS* que respondesse aos requisitos mínimos enquadrados na solução desenhada. Por questões de comodidade, foi selecionado o fornecedor de serviços *Oracle Cloud*, tendo sido criada uma *VPS* enquadrada na opção gratuita. A criação da conta e correspondente configuração encontra-se no anexo A.

Considerando que a *VPs* se encontra apenas com software mínimo para o seu próprio funcionamento, foi necessário instalar e configurar o seguinte software aplicacional (anexo A.2):

1. HAProxy

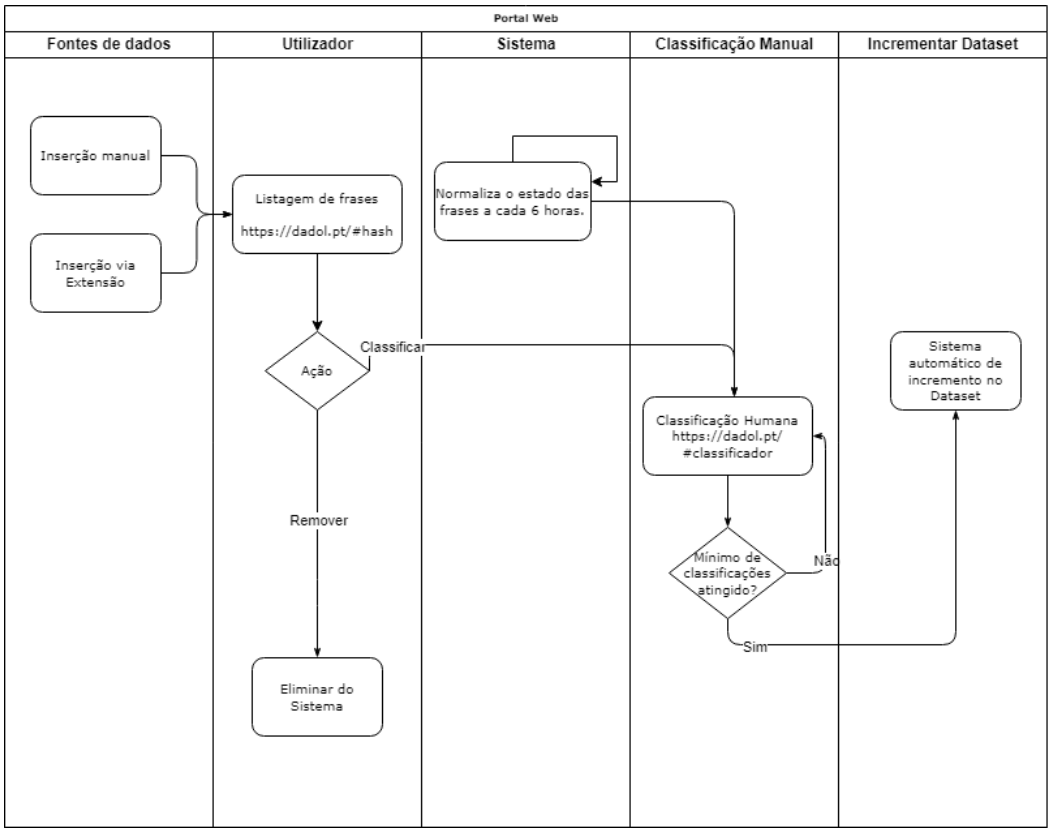


Figura 23: Fluxo funcionamento Portal *Web*

Consiste numa solução *open source* que proporciona alta disponibilidade através de balanceamento de carga de pedidos e *reverse proxy* para aplicações que utilizem *TCP* e *HTTP* [5].

2. Docker

   Trata-se de uma plataforma *open source* pronta para desenvolver, publicar e executar aplicações [6].

3. nodejs

   É uma linguagem de programação *cross-platform* [7].

4. pm2

   Consiste num gestor de processos *daemon* [8].

5. python

   É uma linguagem de programação de *scripting*.

5 https://www.haproxy.com
6 https://www.docker.com
7 https://nodejs.dev/en/
8 https://pm2.keymetrics.io

Para além da *VPS* e devido à necessidade de alterar a *VPS* e o endereço IP, foi criado um domínio para o projeto: `https://dadol.pt`. A configuração está feita com base no endereço *IP* atribuído à *VPS*.

4.2.3 *Dificuldades e melhorias*

O processo de conceção teórica da solução foi, numa primeira fase, moroso e complexo. Tal deve-se ao facto de a implementação de uma plataforma modular com caraterísticas de resposta praticamente em tempo real envolve preocupações de desempenho e escolha da melhor tecnologia a ser utilizada.

Ultrapassada esta fase, iniciou-se o desenvolvimento da extensão do navegador Google Chrome, tendo surgido um problema técnico não identificado na tipificação das funcionalidades. Na prática, a extensão efetua *scraping* da *tab* ativa, num ciclo infinito em busca de alterações. Este facto tem impacto no normal funcionamento do navegador, consumindo muitos recursos computacionais. A solução encontrada consistiu na redução da amostragem, em tempo e quantidade, o que significa uma análise incompleta à página. Ou seja, poderá existir manchas de texto que não são analisadas, não estando assim garantida a total análise da *tab*.

Desde cedo foi uma preocupação garantir que não seriam enviados dados sensíveis (emails, passwords, entre outros) para a base de dados, mas uma vez mais a implementação desta filtragem no navegador iria comprometer o seu desempenho. Desta forma, apenas é efetuada uma pequena triagem no navegador e a restante é efetuada no servidor de destino, garantindo assim que os dados sensíveis não são armazenados na base de dados. Contudo, estes continuam a ser enviados, por *HTTPS*, para o servidor.

As melhorias mais relevantes consistem na otimização do processo de captura efetuado pela extensão, bem como aumentar a capacidade de filtragem por parte do servidor no que diz respeito aos dados sensíveis. Relativamente ao processo de validação das classificações efetuadas, seria interessante integrar um mecanismo de ML não supervisionado, por forma a reduzir a necessidade de atuação humana.

5

ANÁLISE E DISCUSSÃO DE RESULTADOS

Os objetivos do presente projeto são tripartidos: construção de um *dataset* anotado e classificado em língua portuguesa, adequado à deteção automática de discurso de ódio; desenvolvimento de um classificador de discurso de ódio; e implementação de uma extensão para o navegador *Google Chrome* como prova de conceito da aplicabilidade do classificador de discurso de ódio. No presente capítulo procede-se à apreciação detalhada do *dataset* anotado e classificado, à comparação de resultados obtidos no decorrer do desenvolvimento do classificador e à exposição dos resultados alcançados decorrentes da utilização da prova de conceito.

5.1 ANÁLISE DOS DADOS RECOLHIDOS

Os dados utilizados para o treino dos modelos estão presentes num *dataset* que contém 354 comentários e frases anotados e classificados como 0 (discurso legítimo) e 1 (discurso de ódio). O *dataset* não se encontra equilibrado, existindo 203 comentários/frases classificados como discurso de ódio (valor 1) e 151 como discurso legítimo (valor 0). Os dados recolhidos são provenientes de redes sociais (223 comentários/frases) e do livro "Para cima de Puta"(131 comentários/frases). Para a construção do *dataset* foi necessário desenvolver mecanismos de *scraping* aplicáveis às redes sociais e, posteriormente, para a conversão de imagens para texto utilizando técnicas de *OCR*, de forma a recolher o conteúdo de ódio presente no livro "Para cima de Puta".

O *dataset* não tem, ainda, as dimensões que permitam conceber um modelo de *Machine Learning* eficiente e próximo do pretendido, nomeadamente a deteção de todos os tipos de discurso de ódio. Contudo, considerando a não existência de outro dataset, esta pode ser a base de trabalho para alavancar trabalhos de investigação futura nesta matéria.

Método	AUC	F1	Precisão	Recall
Neural Network	0.82	0.70	0.76	0.65
SVM	0.79	0.63	0.69	0.58
kNN	0.54	0.24	0.55	0.15
Naive Bayes	0.87	0.67	0.51	0.98

Tabela 3: Resultados dos algoritmos, sem a utilização de *Bag of Words*

5.2 ANÁLISE DOS RESULTADOS

Com vista a facilitação da interpretação dos resultados obtidos, estes foram agrupados em quatro cenários.

Assim, os resultados obtidos no decorrer do primeiro cenário estão visíveis na tabela 3. Podemos observar que o algoritmo *Neural Network* obtém uma precisão de 76% e um F1 de 70%, enquanto o algoritmo *SVM* apresenta uma precisão de 69% e um F-score de 63%.

Relativamente aos algoritmos *kNN* e *Naïve Bayes*, apresentam uma precisão abaixo de 60%. Para estes dois casos, a curva *ROC* destes algoritmos para a classificação de discurso de ódio pode ser verificada na figura 24. Por sua vez, a curva para o discurso legítimo pode ser observada na figura 25.

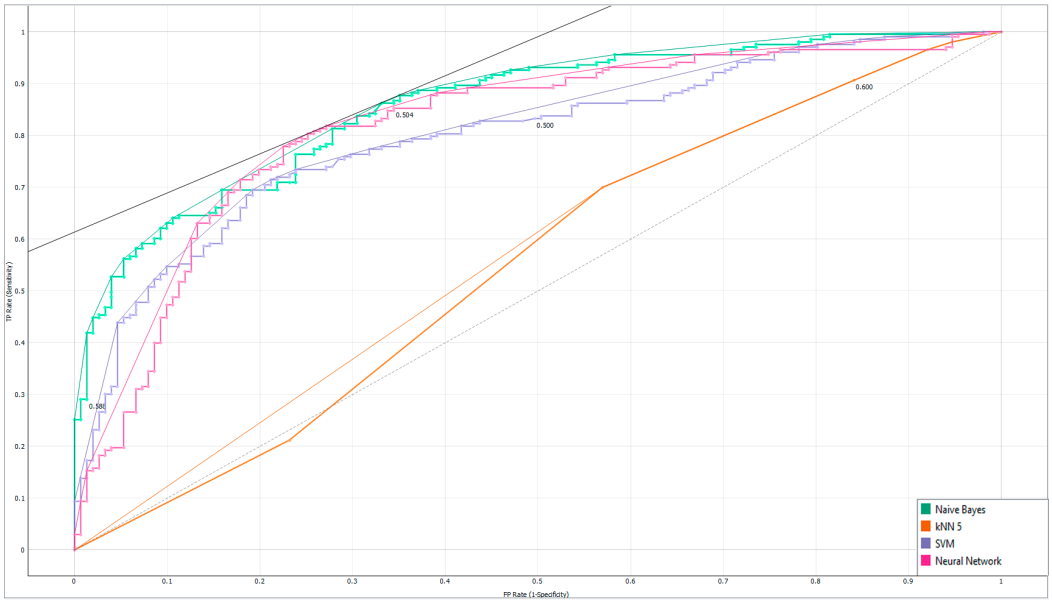


Figura 24: Curva *ROC* de deteção de discurso de ódio, sem *Bag of Words*

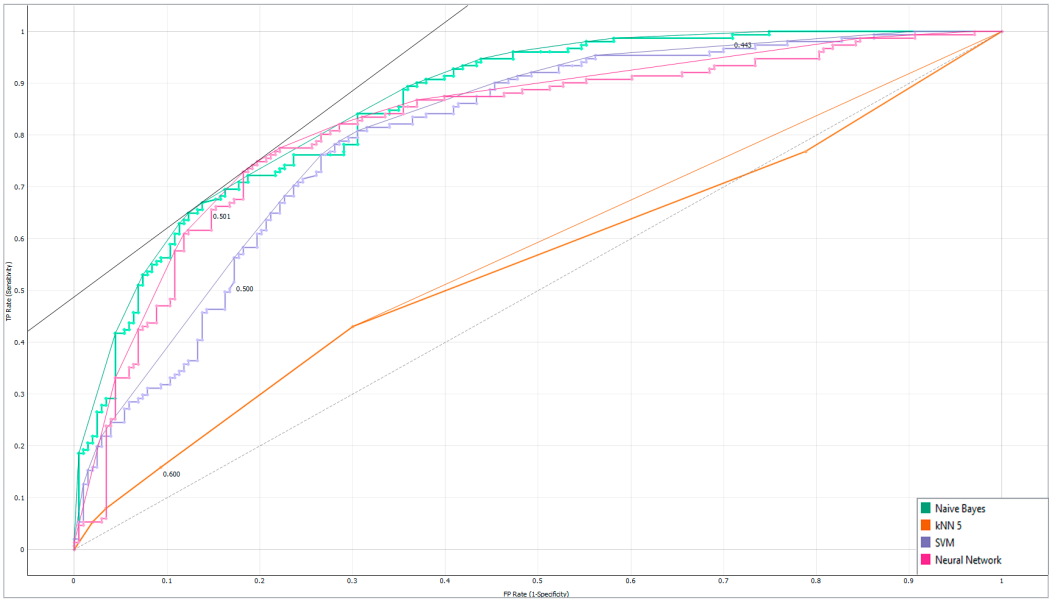


Figura 25: Curva *ROC* de deteção de discurso legítimo, sem *Bag of Words*

Método	AUC	F1	Precisão	Recall
Neural Network	0.84	0.75	0.77	0.76
SVM	0.82	0.76	0.77	0.76
kNN	0.59	0.44	0.75	0.58
Naive Bayes	0.82	0.50	0.72	0.55

Tabela 4: Resultados dos algoritmos, com *Bag of Words*

O segundo cenário diferencia-se do primeiro pelo uso da técnica de *Bag of Words*, sendo possível analisar na tabela 4 um incremento geral do desempenho dos algoritmos. O algoritmo *Neural Network* destaca-se com uma precisão de 77.8% e um *F-score* de 76%. O algoritmo *SVM* apresenta uma precisão de 77% e um *F-score* de 76.3%. Os algoritmos *kNN* e *Naive Bayes* denotam um incremento superior a 41% na precisão e um incremento superior a 80% na métrica *F-score*. Quanto a esta implementação, a curva *ROC* destes algoritmos para a classificação de discurso de ódio encontra-se na figura 26e a do discurso legitimo na figura 27.

O terceiro cenário acrescenta à técnica anterior a técnica *IDF*, que teve implicação no desempenho dos algoritmos, à exceção do algoritmo *Neural Network*. De facto, verificou-se um aumento pouco significativo nos restantes algoritmos, como pode ser comprovado na tabela 5. Relativamente a esta implementação a curva *ROC* destes algoritmos para a classificação de discurso de ódio encontra-se na figura 28 e de discurso legitimo encontra-se na figura 29.

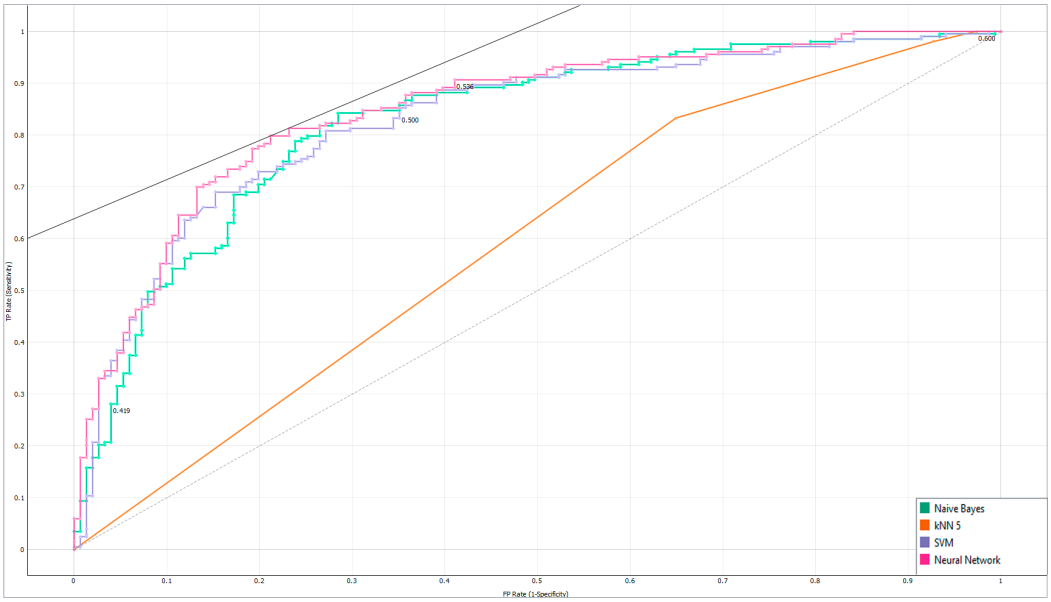


Figura 26: Curva *ROC* de deteção de discurso ódio, com *Bag of Words*

Método	AUC	F1	Precisão	Recall
Neural Network	0.85	0.76	0.77	0.76
SVM	0.83	0.76	0.77	0.76
kNN	0.59	0.44	0.76	0.58
Naive Bayes	0.82	0.50	0.72	0.55

Tabela 5: Resultados dos algoritmos, com *Bag of Words* e *IDF*

O último cenário utiliza *Bag of Words* e *Smooth IDF*. Este cenário revelou ser o mais eficiente, como se pode observar nas curvas de ROC ilustradas na figura 30 e Figura 31, onde é possível denotar uma aproximação do valor 1. Os valores obtidos pelos algoritmos podem ser analisados na tabela 6, verificando-se um incremento significativo no algoritmo *kNN* na generalidade das métricas, enquanto que os restantes melhoraram essencialmente na métrica *AUC* e *Recall*.

Por fim e não descurando os cenários anteriores, foram testadas algumas combinações de *stacking* de algoritmos. De uma forma geral, todas as combinações efetuadas obtiveram valores muito positivos, como se pode analisar na tabela 7. Dá-se especial destaque à combinação *Neural Network* e *kNN*, cujos resultados obtidos ultrapassaram os anteriores, tendo sido alcançado uma precisão de 78% e um *F-score* de 73%. É igualmente visível através das curvas de *ROC* a melhoria no desempenho. A Figura 32 contém as combinações para a classificação de discurso de ódio. O discurso legitimo pode ser observado na Figura 33.

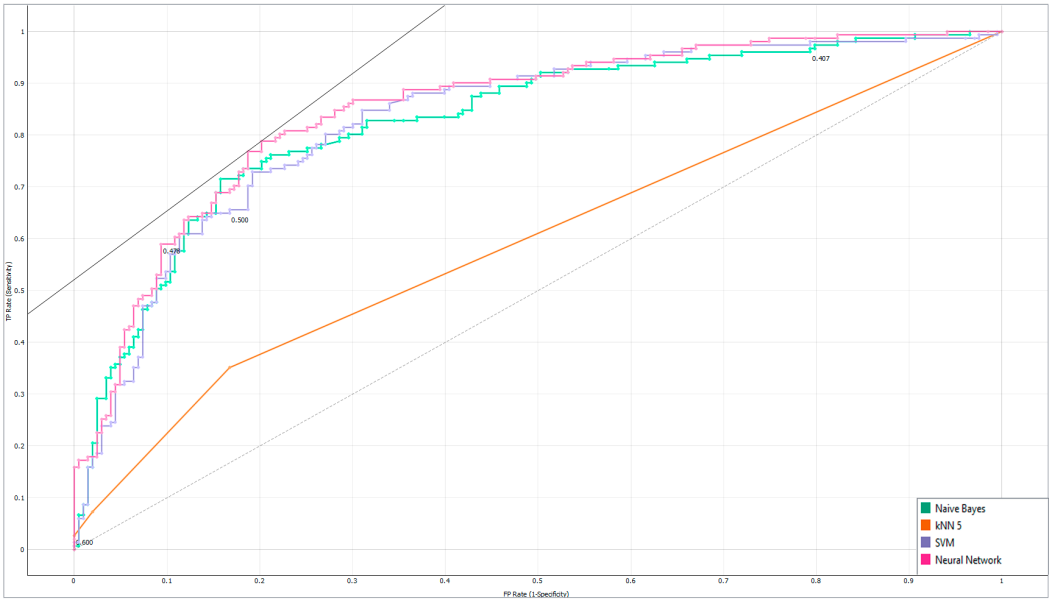


Figura 27: Curva *ROC* de deteção de discurso legítimo, com *Bag of Words*

Método	AUC	F1	Precisão	Recall
Neural Network	0.86	0.76	0.77	0.76
kNN	0.82	0.76	0.76	0.76
SVM	0.83	0.74	0.74	0.74
Naive Bayes	0.84	0.49	0.72	0.55

Tabela 6: Resultados Algoritmos Com *Bag of Words* e *Smooth IDF*

Em suma, o resultado das várias implementações originaram valores de precisão compreendidos entre 51% e 78% e um *F-score* compreendido entre 24% e os 76%. Os algoritmos *kNN & NN* revelaram os melhores resultados com *stacking*. Relativamente aos algoritmos base, o algoritmo Neural Network apresentou os melhores resultados, embora inferiores aos obtidos com *stacking*.

5.3 RESULTADOS DA PROVA DE CONCEITO

Desde o momento da disponibilização, a utilização do portal *web* e da extensão têm sido superiores ao esperado. De facto, a temática da deteção de discurso de ódio é uma preocupação generalizada. Desde o seu lançamento foi possível verificar empiricamente uma maior procura por parte de indivíduos com filhos, que expressaram não ter maneira de "controlar"este comportamento negativo para com a sua criança. Considerando o âmbito de prova de conceito sob uma especulativa hipótese

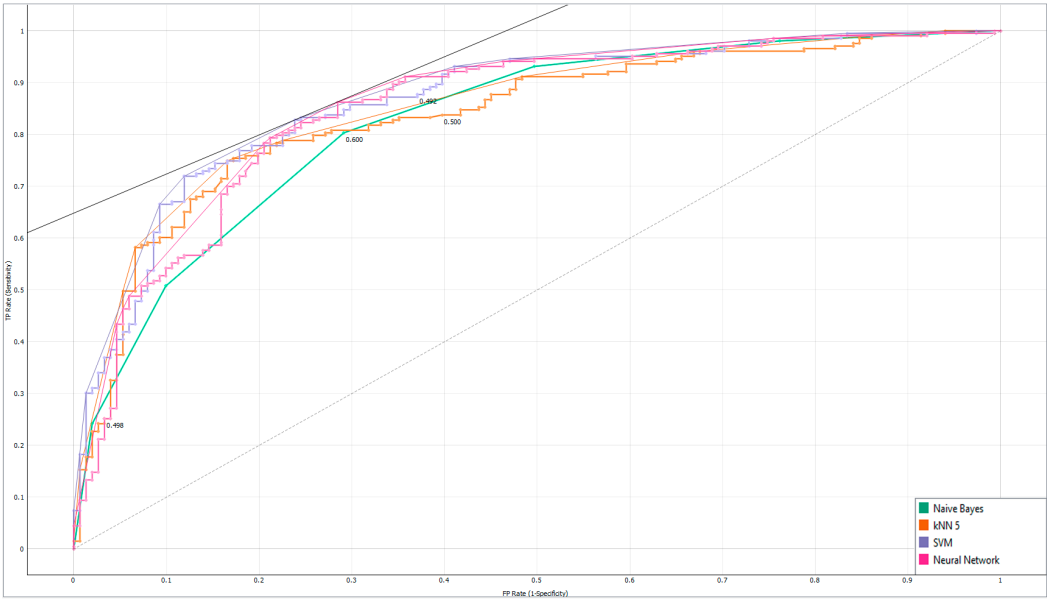


Figura 28: Curva *ROC* de deteção de discurso de ódio com *Bag of Words* e *IDF*

Método	AUC	F1	Precisão	Recall
knn_NN	0.86	0.73	0.78	0.68
knn_NN_SVM	0.85	0.74	0.78	0.70
Knn_SVM	0.86	0.72	0.77	0.67
svm_nn	0.86	0.71	0.76	0.67

Tabela 7: Resultados obtidos com *stacking* dos algoritmos, com *Bag of Words* e *Smooth IDF*

de torná-la num produto operacional, o *feedback* recebido é deveras importante, pelo que importa salientar:

1. O classificador por vezes classifica como ódio frases do tipo *"tudo bem?"*, ou *"o que queres fazer hoje?"*;
2. O portal *web* não está muito intuitivo;
3. A extensão impacta o funcionamento do navegador, tornando-o mais lento;
4. Ajudaria imenso ter uma ferramenta similar à extensão web, mas para dispositivos móveis;
5. É possível utilizar a API integrando-a num projeto pessoal?
6. A API não devia ter autenticação?
7. Os dados recolhidos estão seguros? Como sei que o meu email e/ou dados sensíveis não acabam no classificador?

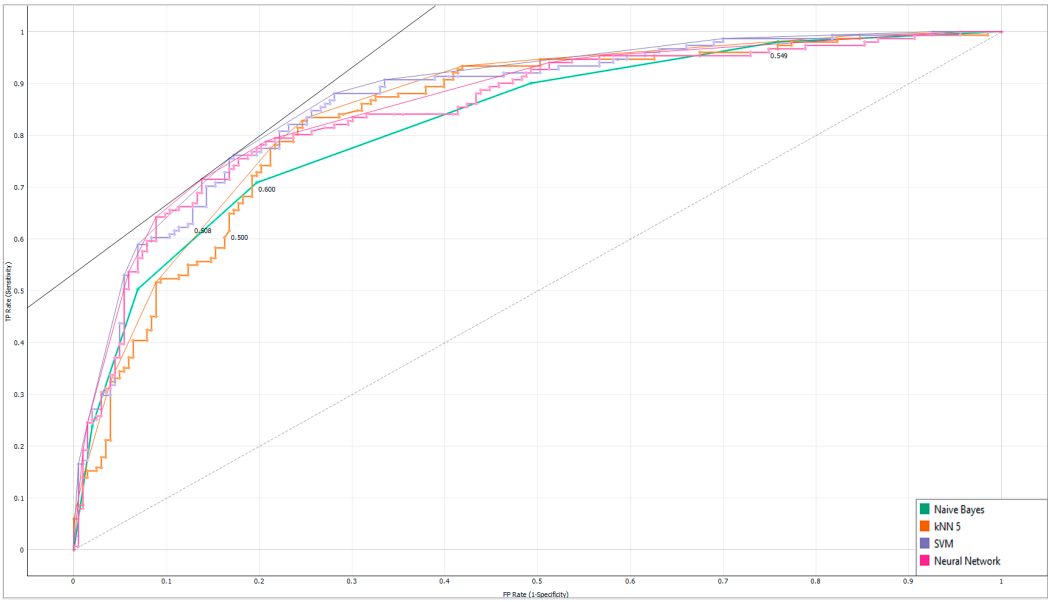


Figura 29: Curva *ROC* de deteção de discurso legítimo com *Bag of Words* e *IDF*

Todo o *feedback* apresentado revela pontos de melhoria desta prova de conceito. Durante o tempo de exposição (cerca de um mês) a extensão já recolheu e validou 480 frases. Destas frases, foram classificada 340 como sendo ódio e 140 como discurso legítimo. Globalmente, 340 frases/comentários foram consideradas bem classificadas. A figura 34 apresenta a relação entre a classificação efetuada pelo classificador e a classificação manual efetuada à posteriori. É possível constatar uma precisão de 70.8% por parte do classificador, o que embora ligeiramente abaixo do valor previsto, é um resulta muito promissor.

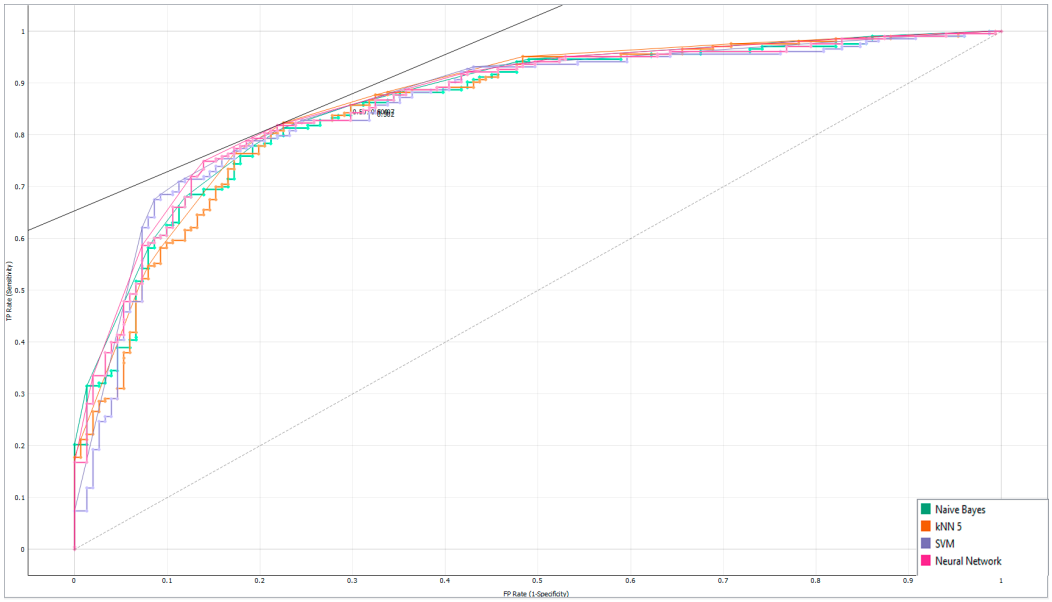


Figura 30: Curva *ROC* deteção discurso ódio Com *Bag of Words* e *Smooth IDF*

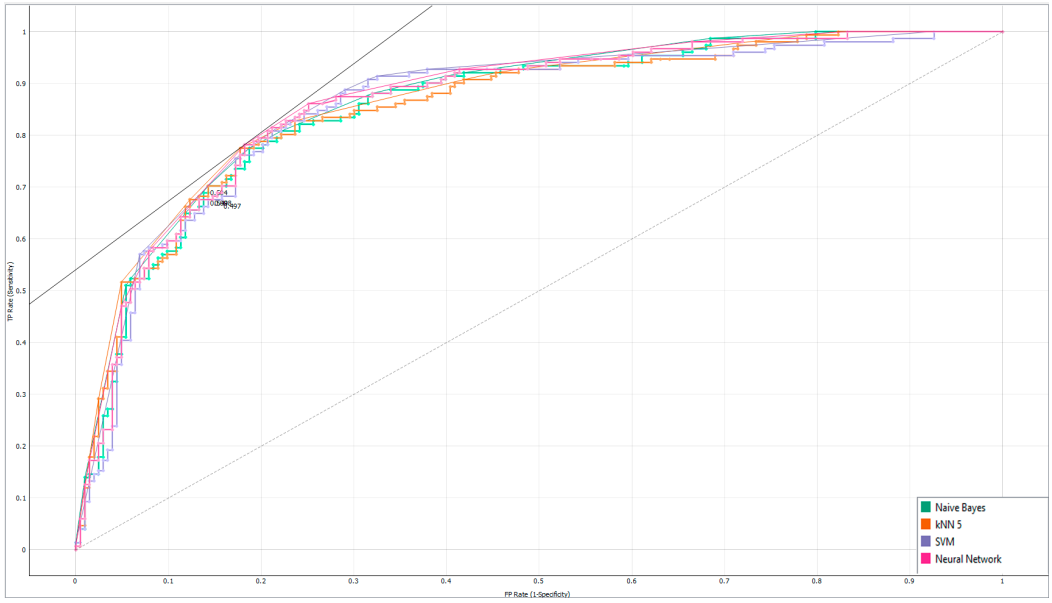


Figura 31: Curva *ROC* deteção discurso legítimo Com *Bag of Words* e *Smooth IDF*

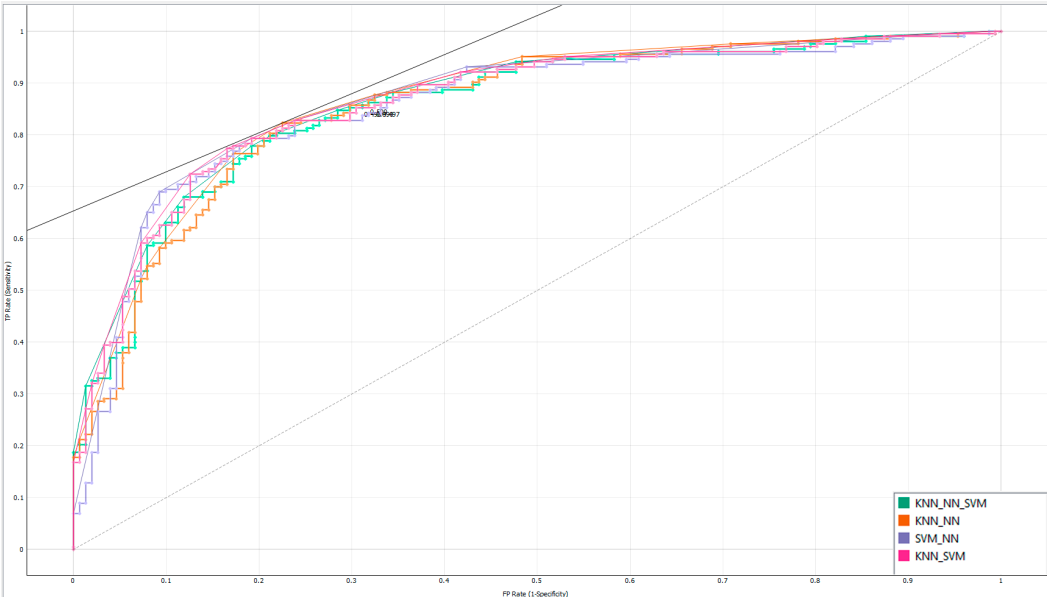


Figura 32: Curva *ROC* de deteção de discurso de ódio, com *stacking*, *Bag of Words* e *Smooth IDF*

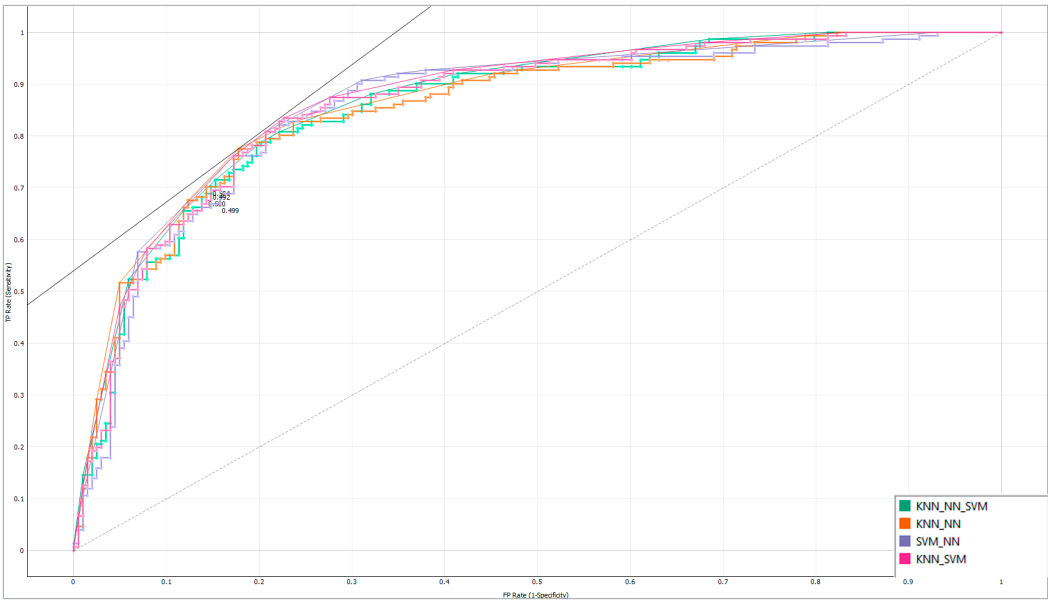


Figura 33: Curva *ROC* de deteção de discurso legítimo, com *stacking*, *Bag of Words* e *Smooth IDF*

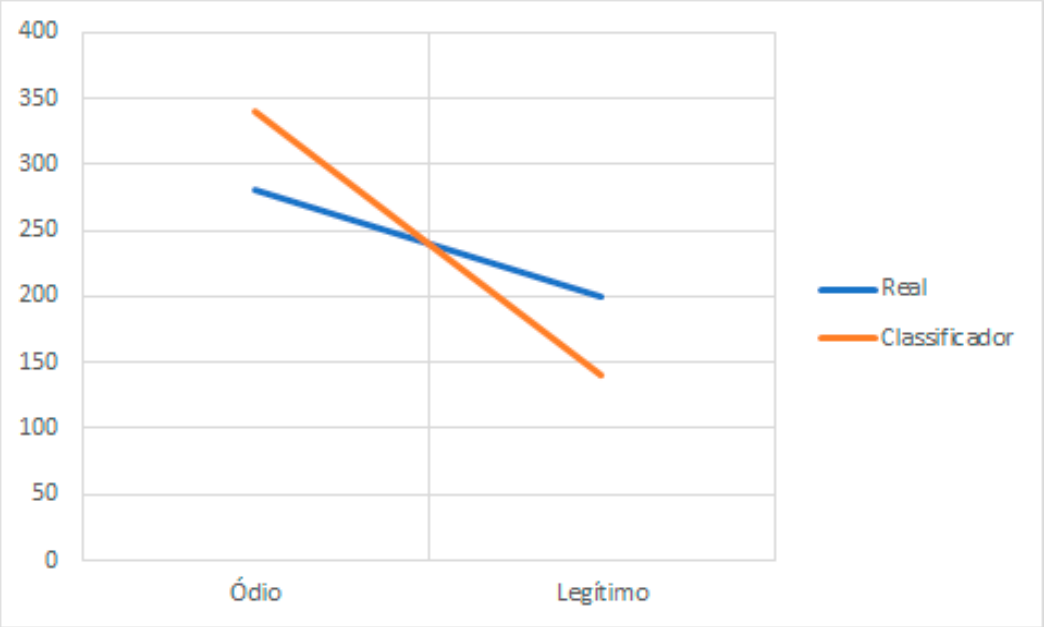


Figura 34: Gráfico resultados Prova de Conceito

6

CONCLUSÕES

O conceito de *cyberbullying* foi utilizado pela primeira vez por Bauman, 2007, assumindo-se como uma variante do tradicional bullying. Genericamente, o *cyberbullying* faz uso das tecnologias da comunicação e informação como meio de denegrir, humilhar e/ou difamar uma ou mais pessoas.

Verifica-se uma preocupação crescente por parte das principais organizações mundiais (governamentais e não governamentais), ao implementar medidas, na tentativa de conter e reduzir este flagelo que tem vindo a vitimar milhões de utilizadores um pouco por todo o mundo. A rápida proliferação deste fenómeno deve-se ao facto de ser praticado *online*, onde não existem limitações espaciais, há um acesso generalizado a tecnologias digitais e existe ainda uma falsa perceção de anonimato. De acordo com Pinheiro, 2009 os resultados do inquérito e entrevistas realizadas, demonstraram que os *cyberbullies*, que são aqueles que praticam *cyberbullying*, sentem que nunca serão identificados. Esta falsa perceção do anonimato que a Internet transparece, agregada à iliteracia digital, a crescente globalização e dinamismo tecnológico do mundo moderno, aumenta a prática deste fenómeno, que merece extrema atenção e investimento na sua prevenção e erradicação Cruz, 2011. Portugal não é alheio a este fenómeno, observando-se um aumento considerável do número de interações consideradas negativas, e/ou alvo de intervenção judicial em Portugal Segurança Interna, 2021.

Esta temática tem estado sob escrutínio mundial o que tem provocado o surgimento de bastantes apoios financeiros para investigações na deteção de *cyberbullying*. Assim, de um modo geral, foi possível constatar através da literatura existente, ainda que focada noutro idioma, que seria exequível desenvolver os objetivos propostos.

Neste projeto abordei o processo de recolha, extração e processamento de mensagens em língua portuguesa, em conjunto com um classificador automático de discurso de ódio (comportamento enquadrado no fenómeno *cyberbullying*) assente em técnicas de ML, designadamente *Naïve Bayes, Support Vector Machines (SVM), Logistic Regression, KNN e Neural Network*. Inicialmente foi efetuada a recolha de informação, que deu origem a um corpus anotado em língua portuguesa. De seguida procedeu-se ao estudo de padrões existentes, por forma a desenvolver um

classificador de discurso de ódio. Este classificador está assente em técnicas de ML, designadamente a combinação de *kNN & Neural Network*, com uma precisão de 78% e um *F-score* de 73%. Simultaneamente, foi elaborada uma plataforma de deteção de discurso de ódio, constituída por uma extensão para o navegador Google Chrome, cuja função básica consiste na recolha de texto existente na *tab* ativa do *browser*. Todo o texto recolhido é enviado via *API REST* para um serviço que implementa o classificador de discurso de ódio. A plataforma encontra-se ainda em testes, embora conte já com dez utilizadores e quatrocentas e oitenta frases únicas classificadas, desde a data de disponibilização online.

Paralelamente, como forma de visualização da classificação atribuída ao texto coletado, foi desenvolvido um portal *web* onde é possível corrigir a classificação efetuada e/ou remover a frase (por se considerar descontextualizado ou por conter informações sensíveis). Complementar a esta funcionalidade, o portal inclui ainda um classificador manual que pretende garantir a correta classificação do *dataset*.

O trabalho desenvolvido encerra algumas limitações, sendo que a principal consiste na dimensão do *dataset*. Embora este seja passível de crescer em função da operacionalização e consequente utilização da plataforma, carecerá sempre de intervenção humana, de modo a garantir a melhoria contínua da qualidade do *dataset*. Uma solução para este problema, poderá passar por integrar técnicas de ML semi-supervisionadas, com vista a aumentar a automatização da classificação. Outra limitação identificada remete para a anotação binária do *corpus*, pelo que será importante alargar o âmbito da classificação, de forma a incluir outros tipos de *cyberbullying* e discurso de ódio, como *sexting* e racismo, entre outros.

Não obstante as limitações referidas previamente, este projeto de investigação representa uma tentativa de prevenção deste fenómeno. Ainda que seja importante ter acesso à Internet, de modo a evitar determinados riscos online, torna-se imperativa a divulgação de um conhecimento geral sobre os riscos que estão associados. A falta de informação e de conhecimento sobre o tema "cyberbullying" é ainda grande, pelo que o investimento na literacia digital entre utilizadores das TIC assume-se como uma necessidade fundamental. Defende-se, assim, uma educação orientada para comportamentos seguros em ambientes online, não somente na ótica da prevenção mas também para esbater o fosso entre jovens e adultos. É previsível que toda a população irá ter presença no mundo virtual e aí desenvolver muitas das suas atividades, num ambiente digital e *online*.

BIBLIOGRAFIA

Abro, Sindhu et al. (2020). «Automatic hate speech detection using machine learning: A comparative study». Em: *International Journal of Advanced Computer Science and Applications* 11.8.

Aliwy, Ahmed H e EH Abdul Ameer (2017). «Comparative study of five text classification algorithms with their improvements». Em: *International Journal of Applied Engineering Research* 12.14, pp. 4309–4319.

Allahyari, Mehdi et al. (2017). «A brief survey of text mining: Classification, clustering and extraction techniques». Em: *arXiv preprint arXiv:1707.02919*.

Andrews, Juan Cristóbal (2020). *K-Nearest Neighbors Classification From Scratch*. URL: https://towardsdatascience.com/k-nearest-neighbors-classification-from-scratch-6b31751bed9b. (accessed: 22.08.2022).

António, Raquel, Rita Guerra e Carla Moleiro (2020). *CYBERBULLYING EM PORTUGAL DURANTE A PANDEMIA DO COVID-19*. URL: https://ciencia.iscte-iul.pt/publications/files/private/deffa87e217ae129586ff95bed171a6e. (accessed: 02.09.2022).

Antunes, M. e B. Rodrigues (2018). *Introdução à Cibersegurança: A Internet, os Aspectos Legais e a Análise Digital Forense*.

APAV (2021a). *Relatório Anual 2019*. https://apav.pt/apav_v3/images/pdf/Estatisticas_APAV-Relatorio_Anual_2019.pdf.

— (2021b). *Relatório Anual 2020*. https://apav.pt/apav_v3/images/pdf/Estatisticas_APAV_Relatorio_Anual_2020.pdf.

— (2021c). *VIOLÊNCIA SEXUAL ONLINE*. https://apav.pt/care/index.php/informacao-para-adult-s/violencia-sexual-online.

Azank, Felipe e Gustavo Corrêa (2022). *Clustering — Conceitos básicos, principais algoritmos e aplicações*. URL: https://medium.com/turing-talks/clustering-conceitos-b%5C%C3%5C%A1sicos-principais-algoritmos-e-aplica%5C%C3%5C%A7%5C%C3%5C%A3o-ace572a062a9. (accessed: 01.09.2022).

Azevedo, Ana e Manuel Filipe Santos (2008). «KDD, SEMMA and CRISP-DM: a parallel overview». Em: *IADS-DM*.

Badjatiya, Pinkesh et al. (2017). «Deep learning for hate speech detection in tweets». Em: *Proceedings of the 26th international conference on World Wide Web companion*, pp. 759–760.

Bauman, Sheri (2007). «Cyberbullying: A virtual menace». Em: *National Coalition Against Bullying National Conference*. Vol. 2. 4.

Berkeley, UC (2020). *What Is Machine Learning (ML)?* URL: https://ischoolonline.berkeley.edu/blog/what-is-machine-learning/. (accessed: 01.09.2022).

Calma, Adrian, Tobias Reitmaier e Bernhard Sick (2018). «Semi-supervised active learning for support vector machines: A novel approach that exploits structure information in data». Em: *Information Sciences* 456, pp. 13–33.

Chase, E. e J. Statham (2005). *Commercial and sexual exploitation of children and young people in the UK—a review. Child abuse review*.

Cruz, Ana Catarina Calixto da (2011). «O cyberbullying no contexto português». Tese de doutoramento. Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa.

Duan, Kai-Bo e S Sathiya Keerthi (2005). «Which is the best multiclass SVM method? An empirical study». Em: *International workshop on multiple classifier systems*. Springer, pp. 278–285.

Elgar, Frank J et al. (2014). «Cyberbullying victimization and mental health in adolescents and the moderating role of family dinners». Em: *JAMA pediatrics* 168.11, pp. 1015–1022.

Fayyad, Usama, Gregory Piatetsky-Shapiro e Padhraic Smyth (1996). «From data mining to knowledge discovery in databases». Em: *AI magazine* 17.3, pp. 37–37.

Ferreira, Cristina (2021). *Pra Cima de Puta de Cristina Ferreira*.

Gambäck, Björn e Utpal Kumar Sikdar (2017). «Using convolutional neural networks to classify hate-speech». Em: *Proceedings of the first workshop on abusive language online*, pp. 85–90.

Google (2022). *Classificação: curva ROC e AUC*. URL: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc. (accessed: 02.09.2022).

Han, Jiawei e Micheline Kamber (2012). *DATA MINING: CONCEPTS AND TECHNIQUES 3RD EDITION*.

Haykin, Simon (2001). *Redes neurais: princípios e prática*. Bookman Editora.

— (2009). *Neural networks and learning machines, 3/E*. Pearson Education India.

Kleinbaum, David G et al. (2002). *Logistic regression*. Springer.

Kowalski, Robin M, Susan P Limber e Patricia W Agatston (2012). *Cyberbullying: Bullying in the digital age*. John Wiley & Sons.

Kusiak, Andrew (2001). «Feature transformation methods in data mining». Em: *IEEE Transactions on Electronics packaging manufacturing* 24.3, pp. 214–221.

Law, Danielle M et al. (2012). «Are cyberbullies really bullies? An investigation of reactive and proactive online aggression». Em: *Computers in Human Behavior* 28.2, pp. 664–672.

Lertvittayakumjorn, Piyawat e Francesca Toni (2022). «Argumentative Explanations for Pattern-Based Text Classifiers». Em: *arXiv preprint arXiv:2205.10932*.

MacAvaney, Sean et al. (2019). «Hate speech detection: Challenges and solutions». Em: *PloS one* 14.8, e0221152.

Madhulatha, T Soni (2012). «An overview on clustering methods». Em: *arXiv preprint arXiv:1205.1117*.

Mason, Kimberly L (2008). «Cyberbullying: A preliminary assessment for school personnel». Em: *Psychology in the Schools* 45.4, pp. 323–348.

Montalvão, Nuno Manuel Martins (2015). «Cyberbullying: caracterização do fenómeno em Portugal». Tese de doutoramento.

packt (2021). *SVM for churn prediction*. URL: https://subscription.packtpub.com/book/all-products/9781789345070/3/ch03lvl1sec24/svm-for-churn-prediction. (accessed: 01.09.2022).

Park, Ji Ho e Pascale Fung (2017). «One-step and two-step classification for abusive language detection on twitter». Em: *arXiv preprint arXiv:1706.01206*.

Patchin, Justin W e Sameer Hinduja (2015). «Measuring cyberbullying: Implications for research». Em: *Aggression and Violent Behavior* 23, pp. 69–74.

Pinheiro, Luzia (2009). «Cyberbullying em Portugal: uma perspectiva sociológica». Tese de doutoramento.

Priy, Surya (2021). *Clustering in Machine Learning*. URL: https://www.geeksforgeeks.org/clustering-in-machine-learning/. (accessed: 01.09.2022).

Putri, TTA et al. (2020). «A comparison of classification algorithms for hate speech detection». Em: *Iop conference series: Materials science and engineering*. Vol. 830. 3. IOP Publishing, p. 032006.

Rish, Irina et al. (2001). «An empirical study of the naive Bayes classifier». Em: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3. 22, pp. 41–46.

Segurança Interna, Sistema de (mai. de 2021). *Relatório anual de segurança interna 2021*. https://www.portugal.gov.pt/download-ficheiros/ficheiro.aspx?v===BQAAAB+LCAAAAAAABAAzNLI0NgcAIUgtZwUAAAA=.

Seixas, Sónia, Luis Fernandes e T de Morais (2016). «CYBERBULLYING: um guia para pais e educadores». Em: *Lisboa, Portugal: Plátano Editora*.

Serrão, Gonçalo Nuno Correia Zambujo (2019). «Cyberbullying». Tese de doutoramento.

Smith, Peter K et al. (2008). «Cyberbullying: Its nature and impact in secondary school pupils». Em: *Journal of child psychology and psychiatry* 49.4, pp. 376–385.

Soares, Pablo Luiz Braga e José Patrocínio da Silva (2011). «Aplicação de redes neurais artificiais em conjunto com o método vetorial da propagação de feixes na análise de um acoplador direcional baseado em fibra ótica». Em: *Revista Brasileira de Computação Aplicada* 3.2, pp. 58–72.

Tsangaratos, Paraskevas e Ioanna Ilia (2016). «Comparison of a logistic regression and Naive Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size». Em: *Catena* 145, pp. 164–179.

unicef (2017). *A Situação Mundial da Infância 2017: As crianças na Era Digital*. URL: https://ciencia.iscte-iul.pt/publications/files/private/deffa87e217ae129586ff95bed171a6e. (accessed: 02.07.2022).

Usama, M (1996). «Advances in knowledge discovery and data mining». Em.

Willard, N (2004). *I can't see you, You can't see me. How the use of information and communication technologies can impact responsible behavior. Retrieved 5, July, 2013*.

Willard, Nancy E (2007). *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress*. Research press.

Yufeng (2021). *Fuzzy C-Means Clustering with Python*. URL: https://towardsdatascience.com/fuzzy-c-means-clustering-with-python-f4908c714081. (accessed: 23.08.2022).

APÊNDICES

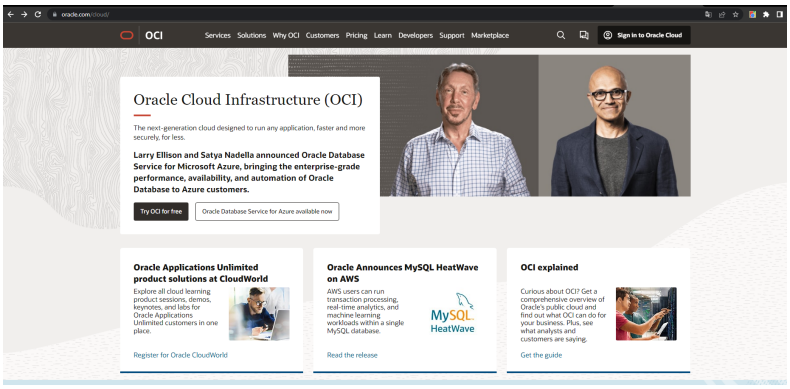
APÊNDICE A

A.1 CRIAÇÃO VPS *ORACLE CLOUD*

De forma a criar uma VPS gratuita no fornecedor de serviços *Oracle Cloud* aceder ao URL https://www.oracle.com/cloud/ e seguir os seguintes passos:

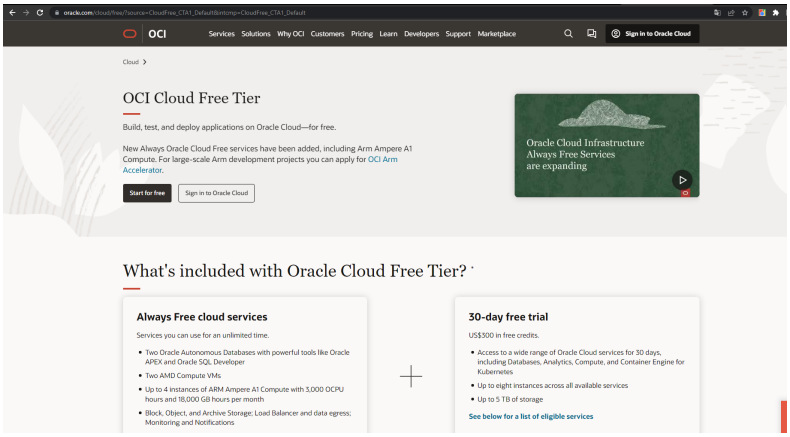
1. Clicar em *Try OCI for free*

Figura 35: Try OCI for free



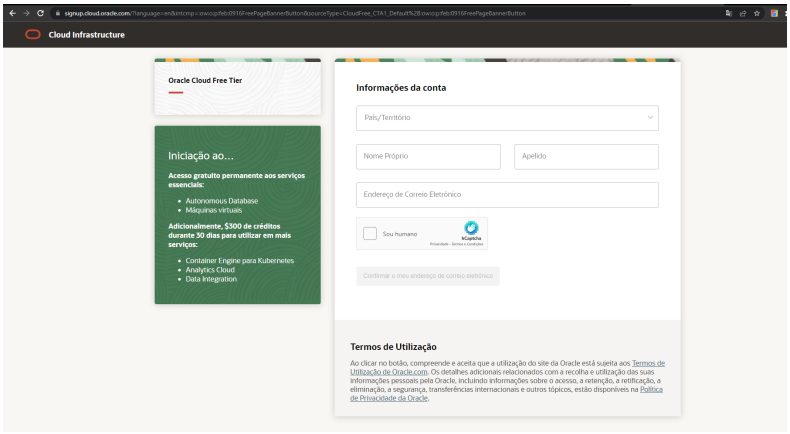
2. Clicar em *Start free*

Figura 36: Start free



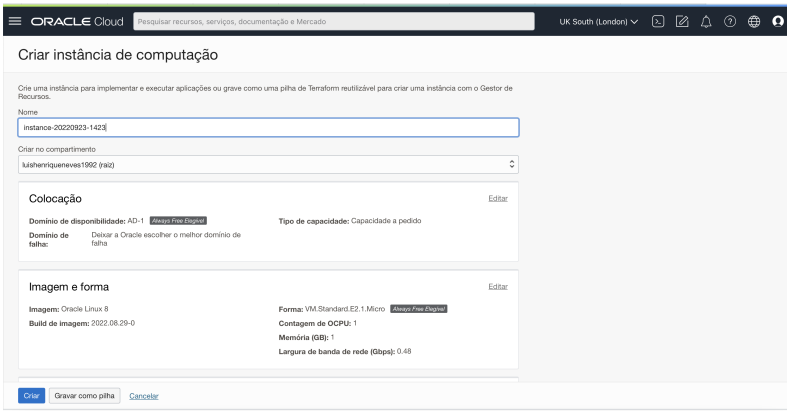
3. Criar conta

Figura 37: Criar Conta



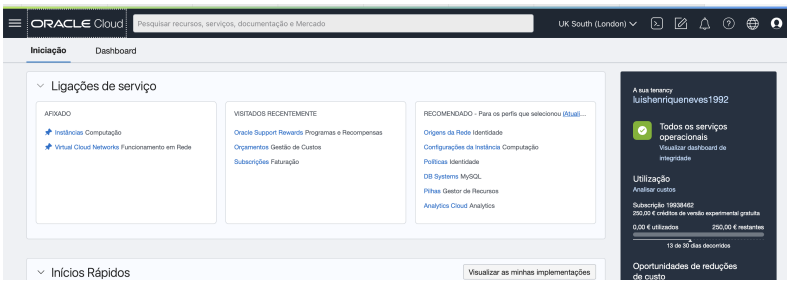
4. Configurar instância de acordo com as especificações pretendidas

Figura 38: Criar Instância



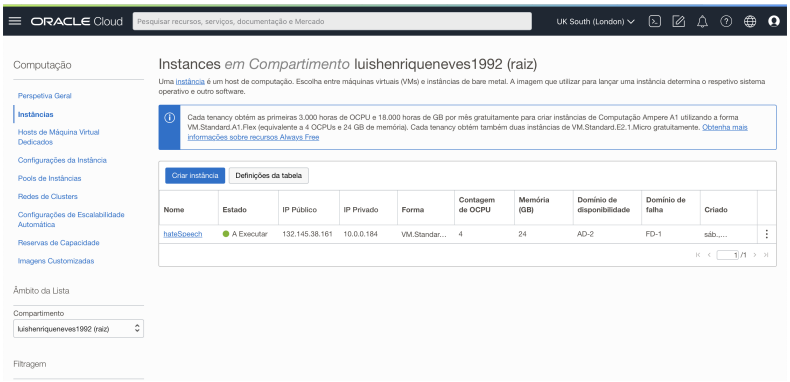
5. Aceder aos detalhes da instância após criação

Figura 39: Aceder aos detalhes da instância após criação



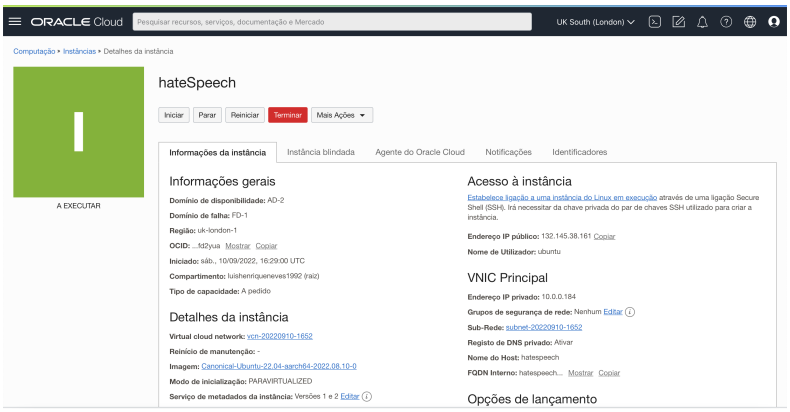
6. Clicar no nome da instância para aceder aos dados desta

Figura 40: Aceder aos dados da instância



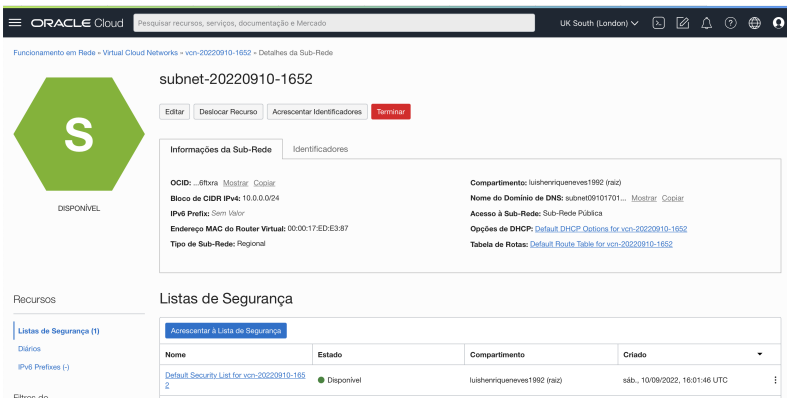
7. Clicar em sub-rede

Figura 41: Aceder à sub-rede



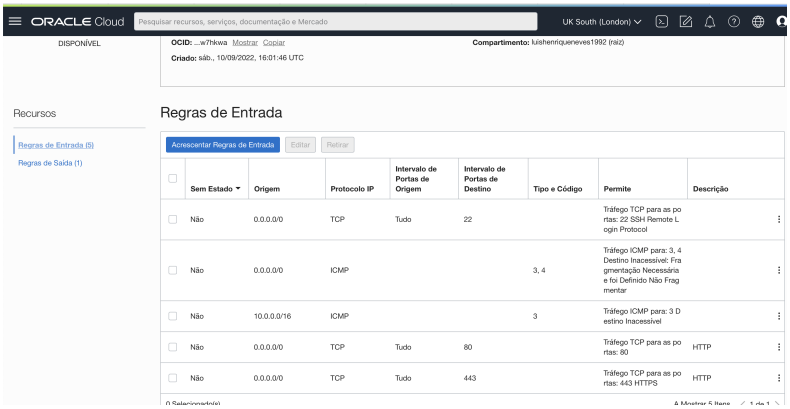
8. Clicar em para adicionar nova regra à *firewall*

Figura 42: Adicionar nova regra à *firewall*



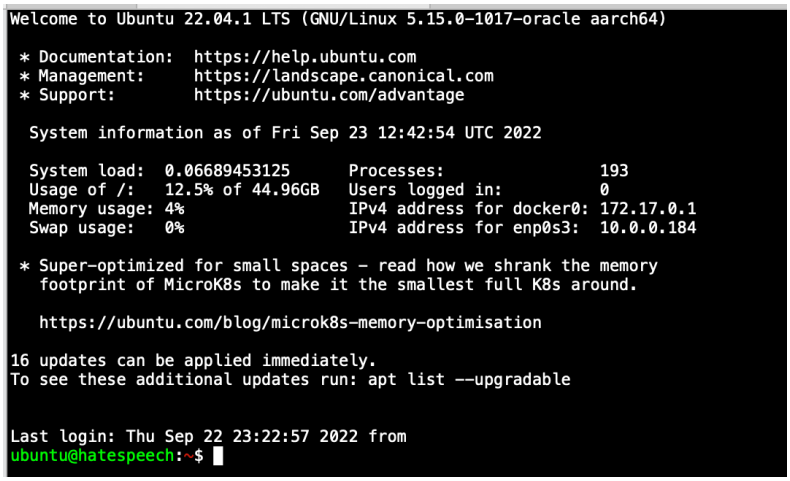
9. Adicionar portos 80 e 443 para acesso *online*

Figura 43: Adicionar portos 80 e 443 para acesso *online*



10. Aceder á instância através *SSH* utilizando os dados definidos no passo 4

Figura 44: Aceder á instância através *SSH*



Está assim instanciada a nossa VPS, de seguida serão apresentados os passos necessários para a preparação necessária por forma a garantir os requisitos do projeto.

A.2 PREPARAÇÃO DA VPS

De forma a facilitar este processo, foi desenvolvido um *script* que irá instalar e configurar todos os *softwares* necessários. Este *script* encontra-se disponível do repositório, no seguinte URL https://github.com/LuisHN/Detetor-Discurso-Odio/blob/main/infrastructure/setup.sh e na figura seguinte.

Figura 45: Fragmento do *Script de setup*

infrastructure > setup.sh

```
#!/bin/bash
sudo su
# Update / Upgrade apt packages
apt update
apt dist-upgrade -y
# Install system dependencies
apt install -y \
  bash-completion \
  curl \
  fasd \
  gnome-tweak-tool \
  htop \
  moreutils \
  shellcheck \
  snapd \
  software-properties-common \
  tilix \
  tldr \
  tree \
  haproxy \
  git \
  nodejs \
  python3 \
  python3-venv \
  libaugeas0
# Install docker
apt-get update && apt-get install -y apt-transport-https ca-certificates curl software-properties-common
curl -fsSL https://download.docker.com/linux/ubuntu/gpg | apt-key add -
apt-key fingerprint 0EBFCD88 | grep docker@docker.com || exit 1
add-apt-repository "deb [arch=amd64] https://download.docker.com/linux/ubuntu $(lsb_release -cs) stable"
apt-get update
apt-get install -y docker-ce
docker run --rm hello-world
# Config docker
groupadd docker
usermod -aG docker $USER
systemctl restart docker
# Config Dev dependencies
npm install -g @angular/cli
npm i -g @nestjs/cli
# Config Firewall
firewall-cmd --permanent --zone=public --add-service=http

# Config and install Certbot
```

Após executar o *Script* anterior é ainda necessário copiar o conteúdo deste ficheiro , para o seguinte */etc/haproxy/haproxy.cfg* alterando o domínio para o que será utilizado, de seguida executar *systemctl restart haproxy*. A VPS encontra-se neste momento operacional e preparada para executar a solução.

B

APÊNDICE B

B.1 CONFIGURAÇÃO PROJETO

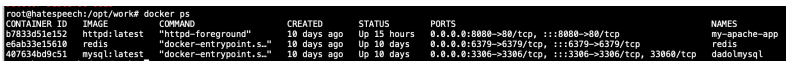
A configuração do projeto tem dependência do anexo A, considerando que todos os passos foram seguidos e concluídos com sucesso, será explicado de seguida os passos necessários para levantar todos os serviços constituintes na plataforma de deteção automática de discurso de ódio em língua portuguesa.

Assim os passos necessários consistem:

1. Criar uma diretoria para o projeto executando o comando na consola SSH "*mkdir -r /opt/work*".
2. Aceder à pasta acabada de criar e executar o comando na consola SSH "*cd /opt/work & git clone https://github.com/LuisHN/Detetor-Discurso-Odio.git*"
3. Aceder à nova pasta criada (Detetor-Discurso-Odio"e conceder permissões de execução ao ficheiro build.sh "*chmod +x build.sh*"
4. executar o script build.sh, este irá efetuar todos os processos necessários para colocar a plataforma *up and running*, "*sh build.sh*"

Se tudo correu como esperado, todos os serviços estão a correr com sucesso. Em suma irá existir o servidor web a correr no porto 80, e uma API gerida pelo *software PM2* no porto 3000. Aparte destes serviços, existe agora também um *container docker* com a base de dados relacional MySQL e um outro com o *Redis*. Executando o comando "*sudo docker ps* o resultado deverá ser o mesmo da imagem seguinte.

Figura 46: *Containers docker*



C

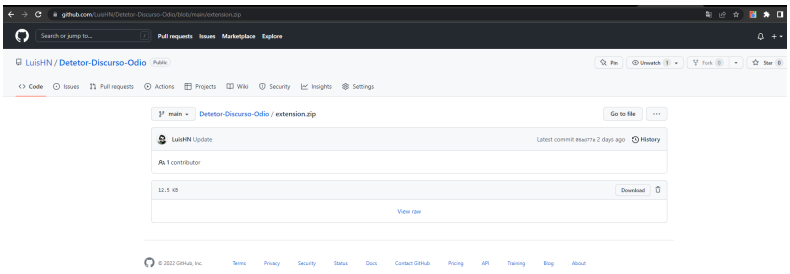
APÊNDICE C

C.1 INSTALAÇÃO MANUAL DA EXTENSÃO

Considerando que à data deste documento a extensão desenvolvida ainda se encontrava em processo de aprovação por parte da Google. E considerando a possibilidade de existir interessa em desenvolver tendo por base a solução já existente. De seguida será apresentada a forma como se pode instalar de forma manual a extensão no navegador Google Chrome.

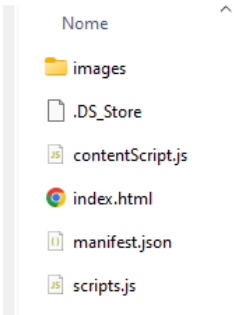
1. Efetuar download do ficheiro extensao.zip utilizando o URL https://github.com/LuisHN/Detetor-Discurso-Odio/blob/main/extension.zip:

Figura 47: *Download* extensao.zip



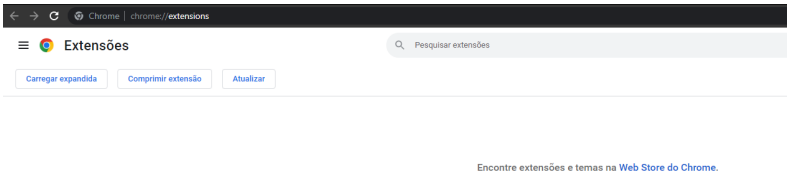
2. De seguida descompactar o ficheiro, obtendo o conjunto de pastas como o da figura seguinte:

Figura 48: Descompressão extensao.zip



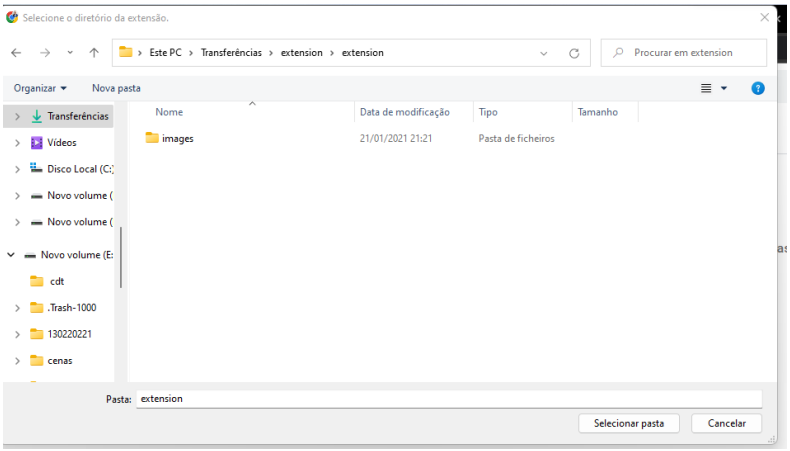
3. Abrir o navegador Google Chrome e aceder ao URL `chrome://extensions/`
4. De seguida clicar em Carregar expandida

Figura 49: Carregar extensão expandida



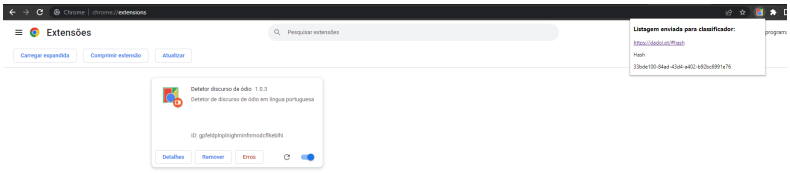
5. Selecionar a pasta descomprimida no passo 2:

Figura 50: Carregar extensão descomprimida



6. Após alguns segundos a extensão estará instalada e já em funcionamento:

Figura 51: Carregar extensão descomprimida



No caso de desenvolvimento, apenas é necessário alterar o código da pasta descomprimida e atualizar a extensão no URL `chrome://extensions/`.

D E C L A R A Ç Ã O

Declaro, sob compromisso de honra, que o trabalho apresentado neste projeto, com o título *"Aplicação para deteção automática de discurso de ódio em língua portuguesa recorrendo a aprendizagem computacional"*, é original e foi realizado por Estudante Luís Henrique Pereira Neves (2200191) sob orientação de Professor Doutor Mário João Gonçalves Antunes (mario.antunes@ipleiria.pt).

Leiria, setembro de 2022

Estudante Luís Henrique Pereira Neves