

**Universidade do Minho**  
Escola de Engenharia

# **Bombedia, uma ferramenta para detecção de Notícias que despoletem Controvérsia**

Laboratórios de Engenharia Informática  
MEI - 4<sup>o</sup> Ano - 2<sup>o</sup> Semestre  
Grupo 95

Constança Elias (PG42820)  
Jorge Brandão Gonçalves (PG42838)  
Maria Araújo Barbosa (PG42844)  
Pedro Pinheiro (PG44421)

20 de agosto de 2021

## Resumo

Este trabalho foi desenvolvido no âmbito da unidade curricular de Laboratórios de Engenharia Informática e tem como principal objetivo contribuir para análise e previsão da capacidade de *Computer-mediated communication* (CMC), em português, despoletarem reações por partes dos leitores e, mais concretamente, controvérsia ou comentários negativos. Trata-se, pois, da criação de uma ferramenta capaz de prever a possível controvérsia de um *post*, desenvolvendo-se, para tal, um modelo de *Machine Learning* com a capacidade de fazer tal previsão. Para o treino deste modelo será usado um *dataset* do *Twitter* e para validação dos resultados obtidos serão utilizados textos de jornais portugueses, *Público* e *Sol*.

**Área de Aplicação:** Análise de Sentimentos, Processamento de Linguagens, *Machine Learning*

## **Agradecimentos**

Este trabalho não seria possível sem a proposta do professor Pedro Rangel Henriques e da professora Cristiana Araújo, a quem agradecemos a ajuda e o entusiasmo com que nos cativaram para o projecto e aceitaram o pedido para serem os nossos orientadores. Gostaríamos também de prestar o nosso agradecimento ao Engenheiro Aragão, do grupo de técnicos do Departamento de Informática, pela disponibilidade imediata para ajudar em todos os problemas técnicos que tivemos com o servidor e pela sempre rápida resposta aos pedidos feitos. É nosso dever agradecer também ao Dr. Ricardo Martins por todas as dicas e conselhos prestados de forma a que o trabalho pudesse tomar o bom rumo que levou.

# Conteúdo

<b>1</b>	<b>Introdução</b>	<b>5</b>
1.1	Contextualização . . . . .	5
1.2	Motivação . . . . .	5
1.3	Objetivos . . . . .	6
1.4	Estrutura do Relatório . . . . .	6
<b>2</b>	<b>Estado de Arte</b>	<b>7</b>
2.1	<i>Computer-Mediated Communication (CMC)</i> . . . . .	7
2.2	Análise de sentimentos e comentários em CMC . . . . .	7
2.3	Análise da controvérsia de CMC em português . . . . .	8
<b>3</b>	<b>Análise e especificação do problema</b>	<b>9</b>
<b>4</b>	<b>Bombedia – Architectura</b>	<b>11</b>
<b>5</b>	<b>Recolha de dados – <i>Dataset</i></b>	<b>13</b>
5.1	Fonte de dados . . . . .	13
5.2	Obtenção dos dados . . . . .	13
5.2.1	<i>Twitter API</i> . . . . .	14
5.2.2	Ferramenta <i>Twint</i> . . . . .	14
5.3	Classificação . . . . .	15
5.3.1	NRC Emotion Lexicon . . . . .	15
5.3.2	<i>Spacy</i> - Lematização e Tokenização . . . . .	15
5.4	Otimização da classificação do <i>dataset</i> . . . . .	17
5.5	Balanceamento do <i>dataset</i> . . . . .	17
<b>6</b>	<b>Análise e preparação dos dados</b>	<b>18</b>
<b>7</b>	<b>Modelos</b>	<b>20</b>
7.1	LSTM . . . . .	20
7.2	GRU . . . . .	20
7.3	CNN . . . . .	21
7.4	<i>Naive Bayes</i> . . . . .	21
7.5	<i>Logistic Regression</i> . . . . .	21
<b>8</b>	<b>Resultados obtidos</b>	<b>23</b>
8.1	Modelos . . . . .	23
8.1.1	Univariáveis . . . . .	23
8.1.2	Multivariáveis . . . . .	25

8.2	Processamento dos jornais . . . . .	25
8.3	Validação dos resultados com o NetAC . . . . .	26
<b>9</b>	<b>Bombastic Media – <i>Web App</i></b>	<b>29</b>
9.1	Requisitos da <i>Web App</i> . . . . .	29
9.2	Público Alvo (Utilizadores) . . . . .	30
9.3	Arquitetura . . . . .	30
9.3.1	<i>App Server</i> . . . . .	31
9.3.2	<i>API Server</i> . . . . .	31
9.3.3	<i>Authentication Server</i> . . . . .	31
9.3.4	<i>TensorFlow Serving</i> . . . . .	31
9.3.5	Servidor <i>Flask</i> . . . . .	32
9.4	Aplicação Final . . . . .	32
9.4.1	Analisador . . . . .	32
9.4.2	Analisador por <i>URL</i> . . . . .	34
9.5	Pesquisa por Palavras Controversas . . . . .	34
9.6	Temas Quentes . . . . .	35
9.6.1	Temas mais procurados . . . . .	35
9.6.2	Análise de Temas ao longo do tempo . . . . .	35
9.7	Manter a classificação atualizada . . . . .	36
<b>10</b>	<b>Alternativas, Decisões e Problemas de Implementação</b>	<b>39</b>
10.1	Decisões . . . . .	39
10.1.1	Dataset . . . . .	39
10.1.2	Otimização da classificação do <i>dataset</i> . . . . .	39
10.1.3	Ferramentas . . . . .	40
10.1.4	Modelos de <i>Machine Learning</i> . . . . .	40
10.2	Problemas de Implementação . . . . .	40
10.2.1	<i>Encoding</i> . . . . .	40
<b>11</b>	<b>Conclusão</b>	<b>41</b>
<b>A</b>	<b><i>Corpus</i> de Palavras Controversas</b>	<b>46</b>
<b>B</b>	<b>Comparação NetAc vs Bombedia</b>	<b>47</b>

# Lista de Figuras

4.1	Arquitetura do Bombedia . . . . .	12
5.1	Exemplos de <i>tweets</i> mal classificados no <i>dataset</i> inicial . . . . .	14
5.2	Pipeline para obtenção do dataset. . . . .	16
5.3	Distribuição da classificação do dataset. . . . .	17
6.1	Palavras mais frequentes nos <i>posts</i> classificados como Negativos . . . . .	19
8.1	Resultados de treino do modelo LSTM ao longo de 30 epochs. . . . .	24
8.2	Resultados de treino do modelo GRU ao longo de 30 epochs. . . . .	24
8.3	Matriz de confusão obtida para o modelo de regressão linear. . . . .	25
8.4	Tabela síntese com os resultados obtidos para o jornal Sol . . . . .	27
8.5	Validação dos resultados obtidos pela ferramenta NetAc. . . . .	28
9.1	Arquitetura da Web App Bombedia. . . . .	31
9.2	<i>Pipeline</i> otimizada da classificação de <i>posts</i> . . . . .	33
9.3	Analisador de <i>posts</i> . . . . .	33
9.4	Exemplo de um resultado fornecido pelo analisador . . . . .	33
9.5	Funcionalidade de procura por palavras controversas no Bombedia . . . . .	34
9.6	Apresentação dos 6 temas mais procurados no <i>Google</i> no dia 16 de junho de 2021 . . . . .	35
9.7	Índice de pesquisas para o termo "Coronavírus" ao longo dos últimos cinco anos . . . . .	36
9.8	<i>Pipeline</i> do processo de atualização do <i>dataset</i> . . . . .	37
9.9	Dataset antes da atualização . . . . .	37
9.10	Dataset após a atualização . . . . .	38

# Capítulo 1

## Introdução

### 1.1 Contextualização

O uso de redes sociais e da comunicação *online* têm vindo a crescer exponencialmente ao longo dos últimos anos. Este facto depoleta a atenção de várias indústrias para as plataformas *online* que proporcionam esta comunicação, como é o caso dos jornais, por exemplo. Consequentemente, a interacção dos utilizadores com notícias *online* tem vindo a adquirir maior importância e vários produtores de meios de comunicação social e conteúdo *online* competem pela atenção dos utilizadores. A presença de notícias sobre temas controversos pode gerar reacções nos utilizadores, levando-os a utilizar a secção de comentários (ou até mesmo as redes sociais) para difundir comportamentos negativos que sejam prejudiciais ao ambiente jornalístico ou outros ambientes de comunicação, onde tais comentários são acessíveis a todos os utilizadores dessas plataformas. [1]. Desta forma, pode surgir um discurso tóxico e ofensivo ou, por outro lado, manifestações a favorecer e a apoiar o que foi dito.

Tendo este contexto por base, foi-nos feita esta proposta de trabalho pelos coordenadores do projeto, no âmbito da unidade curricular de Laboratórios de Engenharia Informática. Surgiu da perceção da importância e do impacto que a criação de uma ferramenta capaz de analisar a controvérsia que *posts* e notícias *online*, poderá ter em áreas não só do processamento de linguagens e da Inteligência Artificial como também em áreas sociais e linguísticas.

Pretende-se então, com este projeto, analisar textos de artigos ou *posts* (concretamente em língua portuguesa) para avaliar a sua potencialidade para despoletar controvérsia, e ainda, procurar estabelecer uma relação entre a existência de discurso pejorativo/grosseiro nos comentários e o sentimento gerado pelo *post* (negativo, positivo ou neutro).

### 1.2 Motivação

A principal motivação que deu origem ao desenvolvimento deste projeto foi a importância da contribuição deste estudo para a análise que existe nesta área em português. Uma vez que esta é bastante escassa quando comparada por exemplo com o trabalho já desenvolvido para a língua inglesa, pretende-se seguir uma abordagem nova, partindo do que já está a ser feito, para a língua portuguesa.

## 1.3 Objetivos

Partindo da motivação acima descrita, pretende-se prever a quantidade de comentários que um conteúdo de *media online* irá gerar ainda antes de ser publicado, ou seja, saber se a notícia será controversa e implicará o movimento das "massas". Para tal avaliação, pretende-se ter por base o tipo de palavras presentes no texto (positivas, negativas ou neutras). Numa fase posterior, pretende-se ainda saber se a reação despoletada será maioritariamente positiva ou negativa e estabelecer uma relação com o tipo de comentários gerados, com o auxílio da ferramenta NetAC (*NetLang Analyser and Classifier*).

## 1.4 Estrutura do Relatório

Este relatório divide-se em onze capítulos.

O Capítulo 1 faz uma breve introdução ao problema e expõe a motivação e objetivos do trabalho a desenvolver.

O Capítulo 2 consiste na análise do estado de arte no que se refere à previsão de comentários e reações que um *post* ou uma notícia *online* poderão despoletar.

O Capítulo 3 analisa em detalhe a contribuição deste projeto para a área bem como os principais requisitos do mesmo.

A arquitetura do **Bombedia** é explicada no Capítulo 4, bem como a *pipeline* de desenvolvimento do projeto.

Os Capítulos 5 e 6 apresentam, respectivamente, como foi obtido o *dataset* para treino dos modelos de *Machine Learning* desenvolvidos e as técnicas de análise e preparação de dados aplicadas aos *posts* selecionados.

Os Capítulos 7 e 8 explicam, respetivamente, os modelos desenvolvidos e os resultados por eles obtidos. Neste último capítulo é ainda efetuada uma comparação dos resultados obtidos pelo **Bombedia** com os resultados obtidos pelo *NetAC*.

A plataforma *web* desenvolvida, **Bombedia**, é detalhadamente apresentada no Capítulo 9.

No Capítulo 10, é feita uma análise resumida das decisões que foram sendo tomadas ao longo do desenvolvimento do projeto bem como dos problemas encontrados.

Por fim, no Capítulo 11, é feita uma breve síntese do trabalho realizado, apresentando as principais conclusões do mesmo.



## Capítulo 2

# Estado de Arte

Ao longo desta secção serão abordados alguns conceitos importantes para a compreensão deste trabalho e ainda vários projetos que têm sido desenvolvidos na área de estudo do presente trabalho.

### 2.1 *Computer-Mediated Communication (CMC)*

De acordo com [2], o termo *Computer-Mediated Communication (CMC)* pode ser definido como “qualquer tipo comunicação humana que envolva um ou mais aparelhos eletrónicos”. Embora tradicionalmente o termo se tenha referido às comunicações que ocorriam e ocorrem através de formatos mediados por computador (por exemplo mensagens instantâneas, correio electrónico, *chat rooms*, *online forums*, redes sociais), este também tem sido aplicado a outras formas de interacção baseada em texto, tais como mensagens de texto (que não envolvem a *internet*) [3].

Vários estudos têm sido feitos para averiguar os diversos campos que envolvem CMC, desde a análise de como os seres humanos utilizam “computadores” (ou meios digitais) para gerir e formar relações interpessoais [4] [5] ao estudo das características paralinguísticas utilizadas neste contexto, tais como *emoticons*, [6] regras pragmáticas, como a tomada de decisões [7] ou a análise sequencial, e organização da conversa [8].

### 2.2 Análise de sentimentos e comentários em CMC

Várias investigações têm sido feitas ao longo dos anos no que se refere à análise de sentimentos e avaliação e previsão das reações que os CMC provocam aos leitores. No entanto, nenhum destes estudos se foca concretamente na língua portuguesa. Recentemente, em 2020, *Branz et al.* [9] realizaram um estudo relacionado com a análise de sentimentos em dados da rede social *Twitter*. Tem-se vindo a estudar como diversas *features* afetam a classificação dos *tweets* [10] [11]. Também foi feito um estudo ficado na previsão de notícias controversas, para italiano, utilizando o *Facebook* [12].

Após a leitura de algum artigo ou notícia *online*, alguns leitores publicam a sua opinião no próprio site onde o artigo foi publicado ou em redes sociais, como o *Twitter*. Estas respostas têm muito conteúdo emocional [13] que muitas vezes vem associado ao conteúdo da notícia ou à forma como a mesma foi escrita. Em 2009, foi feito um estudo com foco na língua holandesa [27], com objetivo de conseguir prever a quantidade de comentários que uma notícia poderá gerar com base em várias *features*,

nomeadamente características semânticas e metadados, como a hora, o autor, entre outros. Outro estudo, de 2019, [13] fez uma previsão das reações emocionais que os utilizadores do *Twitter* expressam depois de ler determinado artigo, utilizando, para isso, uma estratégia de *multitarget*. Para este estudo é utilizado um *Corpus* em língua espanhola. Mais recentemente, em 2020, foi feita uma análise a várias notícias com o objetivo de prever o número total de comentários que uma notícia *online* pode obter, com base na análise de várias *features* (como o tema, o título, os comentários) chegando à conclusão que os primeiros comentários têm muito impacto e definem a existência de reação por parte dos leitores. [14]. Outro estudo semelhante foi feito, também em 2020 para avaliar a toxicidade de notícias, incluindo algumas portuguesas [1].

## 2.3 Análise da controvérsia de CMC em português

A análise da controvérsia e carga emocional associada a CMC em língua portuguesa é muito escassa na literatura. Neste sentido, este projeto pretende trazer alguma novidade na medida em que faz uma análise morfossintática do texto para criar um modelo que seja capaz de prever se um determinado *post online* irá gerar controvérsia. Sendo a língua portuguesa muito rica em termos de vocabulário e expressões próprias, este processo torna-se desafiante e um pouco complexo. No entanto, tirando partido de bibliotecas já existentes e recomendadas noutros estudos, este trabalho pretende obter uma classificação binária em termos do conteúdo da notícia: controversa ou não controversa.

Uma frase é considerada controversa se invoca sentimentos que poderão ser ou não contraditórios e de variados tipos (sentimentos negativos *versus* positivos), prós *versus* contras, argumentos certos ou errados..) [15], podendo dividir o público-alvo em termos de opinião e comentários. [16] sobre um determinado assunto de opinião contraditória [17]. Neste trabalho, a definição de *controvérsia* será reduzida à análise de sentimentos positivos e negativos. Neste contexto, uma frase será então considerada controversa se o sentimento associado ao total das palavras que a compõem for negativo ou positivo (não sendo por isso neutro), indicando que palavras conotadas com um sentimento "forte" têm maior capacidade para despoletar diversidade de opiniões. A classificação como *controversa* pode ser subdividida em duas categorias, *positiva* ou *negativa*, indicando a carga emocional associada à notícia.

Esta abordagem foi inspirada em [18], que concluiu que palavras com conotações negativas ou positivas predominam em textos controversos enquanto que palavras conotação fraca (neutra) aparecem mais em textos não controversos).

Vários trabalhos realizados neste âmbito foram feitos para português do Brasil. [19] [20] [21].

## Capítulo 3

# Análise e especificação do problema

Partindo da avaliação do estado de arte no que se refere à análise de sentimentos e previsão da controvérsia que a comunicação mediada por tecnologias pode despoletar nos leitores, pretendemos desenvolver, neste capítulo, a contribuição que o problema proposto irá trazer para as áreas em estudo.

Como já foi referido, pretende-se estudar a capacidade de um *post* em português gerar controvérsia no público alvo do mesmo. Após uma revisão da literatura, foi constatado que muito trabalho de estudo nesta área tem sido feito para outras línguas, mas no que se refere à língua portuguesa, existe uma lacuna. Encontram-se, no entanto, vários estudos feitos para português do Brasil, como foi referenciado no capítulo anterior.

Da análise do problema, surgiu então a seguinte proposta de trabalho:

- Ter por base o *dataset* para análise de sentimentos do *Twitter* para classificação da capacidade dos *posts* despoletarem controvérsia;
- Utilizar modelos de *Machine Learning* para criar uma ferramenta capaz de identificar, a partir do conteúdo de um *post*, o tipo de reações que este poderá provocar aos leitores;
- Dispor de um *Corpus* de jornais portugueses, fornecidos pelo *NetLang*, para validação dos resultados obtidos;
- Submeter o *Corpus* de jornais à ferramenta *NetAc* (*NetLang Analyzer and Classifier*)<sup>1</sup>, que classifica os comentários de um *post* de acordo com o tipo de discurso de ódio presente, e estabelecer uma relação entre estes resultados e os resultados produzidos pela ferramenta desenvolvida neste projeto;
- Desenvolver uma aplicação *web* que integre as ferramentas desenvolvidas ao longo do projeto, permitindo a diversos utilizadores obter uma classificação de qualquer tipo de CMC.

O foco deste projeto será sobretudo notícias, embora facilmente seja utilizada para a análise de qualquer tipo de texto *online* que envolva a comunicação ou interação entre

---

<sup>1</sup><http://netlang-corpus.ilch.uminho.pt:10100/>

vários utilizadores. Pretende-se, para além disto, permitir aos utilizadores guardarem um histórico dos ficheiros analisados para posterior análise.

Tenciona-se ainda fazer uma comparação dos resultados obtidos neste estudo com os resultados que são fornecidos para ferramenta **NetAC**, que permite classificar os comentários de uma notícia ou outro tipo de CMC, indicando a percentagem de *hate speech* presente nesses comentários, bem como o tipo de discurso de ódio de que se trata (racial, sexual, religioso, etc.).

## Capítulo 4

# Bombedia – Arquitectura

Neste capítulo pretende-se explicar a arquitetura desenvolvida para o **Bombedia** de forma a resolver o problema proposto. As várias fases que compõem a arquitetura do sistema serão analisadas mais em detalhe ao longo dos próximos capítulos.

A arquitetura desta solução encontra-se representada, de forma sucinta, na figura 4.1. Inicialmente, começa-se por extrair dados directamente do *Twitter* que são submetidos a um processo de tratamento (explicado em detalhe no Capítulo 5) com o objetivo de os adaptar de acordo com os requisitos do enunciado. Posteriormente, com o auxílio de um dicionário para palavras em português, *NRC Lexicon*, os *tweets* extraídos são classificados e guardados num ficheiro em formato *csv*. Segue-se, depois, uma fase de processamento (explicada detalhadamente no Capítulo 6), cujo objectivo é preparar os dados para serem submetidos aos modelos de *Machine Learning* desenvolvidos.

Tendo definido vários modelos, sucede-se uma análise dos resultados obtidos com o intuito de seleccionar o modelo com as melhores características. Obtendo o modelo final, procede-se à sua validação através da comparação dos resultados obtidos pelo *NetAC* com os dados obtidos pela plataforma **Bombedia**, definindo, para isso, um classificador que combina os resultados. Finalmente, integra-se o modelo preditivo na interface de consulta desenvolvida, *Bombedia*, na qual é possível o utilizador analisar a controvérsia de *posts* à sua escolha e obter resultados.

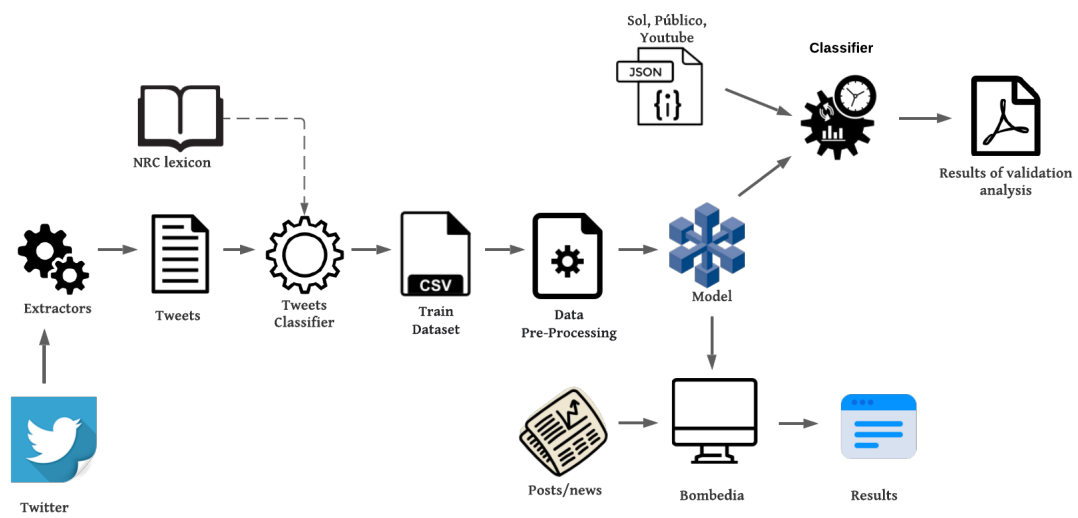


Figura 4.1: Arquitetura do Bombedia

Após se ter apresentado a arquitetura que está na base deste projeto, será explicado cada passo que a caracteriza em particular, ao longo dos próximos capítulos. O capítulo que se segue é referente à etapa de obtenção e classificação do *dataset* utilizado (o *Train Dataset*, que vem referenciado na figura 4.1).

## Capítulo 5

# Recolha de dados – *Dataset*

Nesta secção será explicado em detalhe o processo de obtenção dos dados para a criação de um *dataset* para treino e teste dos modelos de *Machine Learning* que se pretendiam desenvolver.

### 5.1 Fonte de dados

Sabe-se que as redes sociais têm-se tornado uma importante plataforma de comunicação onde os utilizadores expressam opiniões e sentimento em conversas ou mensagens. [22]. Uma plataforma popular nos dias de hoje é o *Twitter* que possui mais de 330 milhões de utilizadores, que partilham cerca de 500 milhões de *tweets* por dia associados a diversos contextos e possuindo diferentes polaridades [23] [24]. Este foi o ponto de partida que levou o grupo a escolher o *Twitter* como fonte de dados. Para além desta razão, vários outros motivos levaram a esta decisão:

- ser uma rede social bastante extensa, com utilizadores de diversas faixas etárias, culturas e condições sociais, contendo, por isso todo o tipo de comentários e *posts*, sobre temas muito variados da atualidade, que permitem obter um *dataset* diversificado [25];
- ser uma fonte para obtenção de *datasets* muito utilizada na literatura, no que se refere a análise de sentimentos e comportamentos dos utilizadores em redes sociais e CMC (como foi analisado no capítulo 2);
- ter sido utilizada anteriormente por um elemento do grupo para o desenvolvimento de um trabalho sobre análise de sentimentos, tendo mostrado resultados bastante positivos;
- ser uma fonte acessível e rápida para recolha de dados de CMC.

### 5.2 Obtenção dos dados

Tendo escolhido a fonte de dados foi necessário proceder à recolha dos mesmos. O *dataset* pretendido deveria ter, no mínimo, duas colunas: *tweets* e *classificação*. Esta última coluna deveria indicar o sentimento associado ao *tweet*, que poderia ser um de três: Positivo, Negativo ou Neutro.

Inicialmente, utilizou-se um *dataset* neste formato que se encontra disponível *on-line*<sup>1</sup>. No entanto, depressa se percebeu que este se encontrava muito mal classificado. Foi detetado, por exemplo, que os *emojis* presentes nos *tweets* tinham um grande peso na classificação dos mesmos induzindo em graves erros. A figura 5.1 apresenta exemplos de alguns *tweets* mal classificados nesse *dataset* (que parecem ser classificados tendo em conta apenas os *emojis* que aparecem no final).

```
Positive => Feliz dias das crianças pra vcs seus lixos :)

Positive => Queridos vizinhos, AGRADEÇO QUE PAREM DE FAZER OBRAS PORQUE NÃO VOS AGUENTO MAIS CARALH
O, atentamente a vizinha do último andar :)

Positive => @requiaopmdb alguns como eu optaram por seguir vc pra ver as aberrações que vc posta ,
as mentiras e promessas falsas , o povo deu a resposta pra vc nas urnas :) por isso não foi reeleito.
Esperta foi a Gleisi Hoffman sabendo que seria estimada de governadora e senadora concorreu a deputad
a

Positive => independente de gostar apenas do que o fp produz, tenha respeito à pessoa que trabalha
nos sons que viralizam no youtube, nas festinhas e nos bailes. hipocrisia master escutar um negro fav
elado e votar num candidato que tem aversão a minorias sociais :-)
```

Figura 5.1: Exemplos de *tweets* mal classificados no *dataset* inicial

Para além disso, este *dataset* estava em brasileiro, o que para o nosso projeto não era benéfico, uma vez que há palavras substancialmente diferentes e/ou com outro significado às palavras equivalentes em língua portuguesa. A adicionar a esse facto, os temas que podem ser considerados controversos no Brasil (nomeadamente, assuntos relacionados com figuras públicas concretas, etc..) podem não o ser em Portugal. Tendo estes dois aspetos em conta e como o foco da análise era a língua portuguesa, o grupo decidiu procurar outro *dataset*. Rapidamente se percebeu que não existia um *dataset* em língua portuguesa já classificado e disponível na *web* pelo que o grupo chegou à conclusão que deveria ser incluída mais uma etapa no projeto: a construção de *dataset* adequado ao objeto de estudo, passando assim pelas fases de extração, normalização e limpeza dos dados, bem como a posterior classificação.

### 5.2.1 *Twitter API*

De forma a proceder à obtenção de novos dados, optou-se por recolher os dados recorrendo diretamente à *API* do *Twitter*. Foi pedida a chave de acesso mas não houve resposta nas duas semanas que se seguiram pelo que foi necessário pensar numa alternativa.

### 5.2.2 Ferramenta *Twint*

Após alguma pesquisa, descobriu-se a existência de uma ferramenta *open source*, *Twint*<sup>2</sup>, que permite, de forma simples, extrair partilhas realizadas no *Twitter*, sem qualquer limitação. Podem ainda ser aplicados filtros de forma a obter um conjunto específico de *tweets* que satisfaçam o pretendido.

Com o objetivo de obter uma distribuição equitativa de dados de todo o país, foram recolhidas mensagens emitidas nas principais cidades portuguesas , até um raio de 30

---

<sup>1</sup><https://www.kaggle.com/leonardoassis/portuguese-tweets-nltk-and-sklearn>

<sup>2</sup><https://github.com/twintproject/twint>



kms de cada cidade. Assim, foi possível garantir que os *tweets* extraídos eram colocados apenas por utilizadores que se encontrassem em território nacional. Optou-se também por fazer a extração de *tweets* escritos a partir de fevereiro de 2021, data em que se iniciou este projeto, de modo a obter comentários mais recentes e atuais. Foi obtido então o primeiro *dataset* com 100000 entradas.

## 5.3 Classificação

Depois de obter o *dataset*, o passo seguinte consistiu em analisar as emoções associadas a cada *tweet* de modo a perceber se o mesmo era controverso ou não.

Para a classificação das várias entradas, desenvolveu-se um pequeno *script* em *Python* que lê cada texto associado ao *tweet* e classifica-o tendo em conta uma lista de palavras classificadas como positivas, negativas ou neutras. Esta lista de palavras corresponde à versão portuguesa do *Corpus* disponibilizado pelo *NRC Word-Emotion Association Lexicon* (ou *EmoLex*).

### 5.3.1 NRC Emotion Lexicon

O NRC Emotion Lexicon consiste num dicionário de palavras em inglês e a sua relação com uma lista de oito emoções. Estas emoções são: (*anger, fear, anticipation, trust, surprise, sadness, joy, and disgust*). Tem também associados dois sentimentos: positivo e negativo. Esta classificação das palavras foi feita por *crowdsourcing*. Para a classificação utilizámos a tradução disponível para português<sup>3</sup>.

### 5.3.2 *Spacy* - Lematização e Tokenização

Para poder aplicar o NRC às várias entradas do *dataset*, foi necessário efetuar um processo de tokenização e de lematização a cada uma dessas entradas. Utilizando o módulo *spacy*<sup>4</sup>, para português, foi possível automatizar o processo de tokenização das frases. Para a lematização dos vários *tokens* obtidos (processo de deflexionar uma palavra para determinar o seu lema) foi definida uma função que extrai o lema a partir dos *tokens* gerados pelo módulo *spacy*. A escolha da utilização deste módulo deveu-se ao facto de o mesmo permitir, de forma simples, efetuar os dois processos referidos para a língua portuguesa.

---

<sup>3</sup>obtida de <https://saifmohammad.com/WebPages/AccessResource.htm>

<sup>4</sup><https://spacy.io/>

Para classificar cada frase do *dataset*, recorreu-se aos módulos previamente indicados e procurou-se verificar se cada um dos lemas associados às palavras das frases se encontravam no dicionário de palavras *NRC*. Para cada *tweet* (a que nos referimos como *frase*) realizou-se uma contagem de palavras positivas e negativas presentes (classificadas utilizando o NRC). Para tal definiu-se um contador que era incrementado por cada palavra que se fosse classificada como positiva no léxico e decrementada se a palavra pertencesse ao grupo das palavras negativas. No final, a frase era considerada positiva se o valor do contador fosse superior a zero e negativa se fosse menor que zero. Se o contador tomasse o valor zero no final, então a frase era classificada como neutra.

O *dataset* final ficou então composto por dois campos: o texto associado ao *tweet* e a sua classificação. O esquema 5.2 resume o *pipeline* seguido para a obtenção do *dataset* classificado.

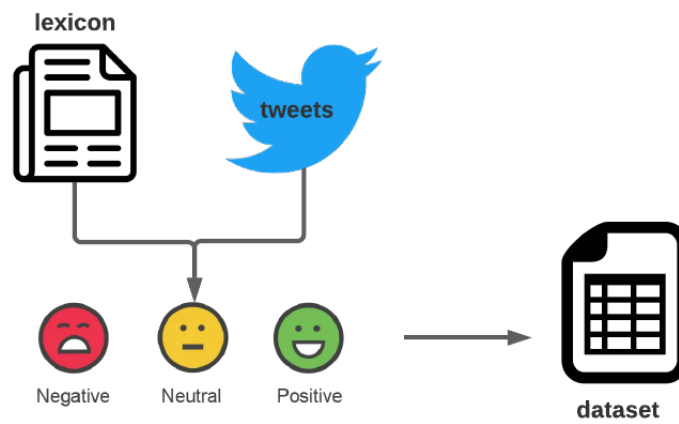


Figura 5.2: Pipeline para obtenção do dataset.

## 5.4 Otimização da classificação do *dataset*

No fim da classificação feita pelo NRC, foi feita uma última fase de processamento em que se classificaram palavras que, embora fossem neutras, pudessem causar controvérsia, sendo estas palavras classificadas como positivas ou negativas. Para tal foi criado um *corpus* onde se encontram palavras que poderão gerar controvérsias em determinados contextos. Exemplos destas palavras são: racismo, transgênero, entre outras. O *corpus* completo pode ser consultado no apêndice A. Esta fase contribuiu para uma melhor classificação dos *tweets* do *dataset* em relação à capacidade de despoletar controvérsia e não só uma simples análise de sentimentos.

## 5.5 Balanceamento do *dataset*

Com a consulta de algumas informações relativas ao *dataset* e a visualização de alguns gráficos, foi possível verificar que o *dataset* se encontrava bastante desbalanceado, apresentando mais de 55 % de *tweets* classificados como neutros, 32% como positivos e apenas 13 % como negativos. Assim sendo, o passo seguinte passou por balancear o *dataset*. Para isso foram extraídos *tweets* associados a temas polémicos aquando da realização do trabalho e que se sabia que iriam gerar controvérsia, isto é, comentários negativos. Estes foram os temas utilizados para o efeito: "Cristina Ferreira", "bbtvi", "Sócrates", "Pinto da costa", "Desconfinamento".

Com esta extração foi possível expandir o *dataset*, aumentando o número de *tweets* associados a uma reação negativa. Aplicando o *script* de classificação definido e explicado anteriormente, foram adicionados os novos *tweets* classificados ao *dataset* tornando-o equilibrado. O gráfico apresentado na figura 5.3 permite visualizar esse equilíbrio, mostrando que o *dataset* apresenta cerca de 60 mil entradas para cada um dos três sentimentos, ficando com um total de 186000 entradas.

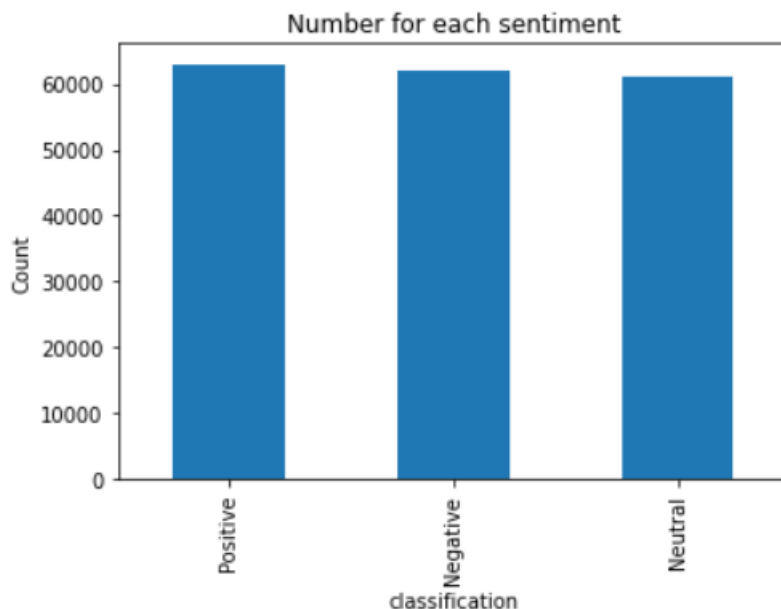


Figura 5.3: Distribuição da classificação do *dataset*.

## Capítulo 6

# Análise e preparação dos dados

Após concluído o processo de tratamento do *dataset*, verificando que não existiam valores em falta e que este se encontrava balanceado, procedeu-se à análise dos componentes dos textos dos *tweets* que seriam realmente importantes para o treino dos modelos definidos numa fase posterior. Ao longo deste capítulo será explicado em detalhe o processo de preparação dos dados. As decisões tomadas tiveram em vista a obtenção de uma melhor *performance* de classificação do modelo. Assim, nesta etapa destaca-se a realização dos seguintes métodos:

- Conversão de todas as palavras para minúsculas;
- Remoção das *stopWords* (palavras que não adicionam qualquer significado à frase);
- Remoção dos sinais de pontuação, caracteres especiais, *links* e *hashtags*;
- Conversão das palavras para um formato *standard* (transformar as abreviaturas nas suas palavras originais).

Todos estes passos foram feitos de modo a minimizar as palavras que eram classificadas numa frase, visto que não é necessário classificar certas palavras que não adicionam significado à frase (*stopWords*). Foram retirados também os caracteres especiais desnecessários à classificação, através de expressões regulares e foram removidos sinais de pontuação e *emojis* recorrendo ao módulo `gensim`<sup>1</sup> do *Python*. Por fim, trocaram-se abreviaturas pelas palavras corretas originais (como por exemplo "tb" para também), de modo a poderem ser corretamente classificadas.

A figura 6.1 apresenta as palavras mais frequentes que foram encontradas nos *tweets* classificados como negativos. Os temas que aparecem em grande estão relacionados com as datas em que foram recolhidos os *tweets*.

---

<sup>1</sup><https://pypi.org/project/gensim/>



## Capítulo 7

# Modelos

Concluída a fase de processamento de dados, procedeu-se ao desenvolvimento do modelo preditivo. Como foi referido inicialmente, decidiu-se recorrer a Inteligência Artificial para esta etapa.

Neste capítulo, será apresentado o trabalho desenvolvido para a obtenção de um modelo de *Machine Learning* (ML) capaz de prever a capacidade que uma notícia tem de gerar controvérsia. Para isto, o modelo criado vai ter em conta a classificação feita anteriormente e o mesmo vai prever a classificação de uma frase. Sabendo a classificação de uma frase, caso esta seja positiva ou negativa conclui-se que se tratará de uma frase capaz de gerar controvérsia. Se a classificação for neutra, a frase considera-se que a frase não irá gerar controvérsia (ver definição na secção 2.3).

Com o objectivo de explorar várias abordagens no que se refere ao desenvolvimento de modelos de ML e perceber aquela que se mostrava mais vantajosa para o estudo em questão, desenvolveram-se vários modelos que serão explicados de seguida.

### 7.1 LSTM

O modelo LSTM (*Long Short Term Memory*) é muito usado em ML, no campo de *Deep Learning*. É do tipo de RNN (*Recurrent Neural Networks*) e possui uma grande capacidade de aprender dependências a longo prazo [30]. Após a análise do estado de arte, foi fácil perceber que este é um dos modelos mais utilizado no âmbito de análise e processamento de linguagens e por isso decidiu-se testar a sua implementação.

A LSTM definida para este trabalho é composta por 3 camadas: uma primeira camada *embedding* seguida da camada LSTM bidirecional e por fim uma *Dense layer* com a função de ativação *softmax*. Este modelo utiliza a entropia categórica cruzada como função de *loss*, a acurácia como métrica e o otimizador *rmsprop*, por ser considerado o otimizador mais rápido em ML [31].

### 7.2 GRU

A GRU (*Gated recurrent unit*) é também um tipo de RNN semelhante à LSTM, muito usada para desenvolver modelos envolvendo processamento de frases e trabalho com linguagens. Este modelo é mais simples do que a LSTM, o que leva a melhores resultados em alguns casos, exigindo na maioria dos casos um menor esforço computacional [32].

Por estas razões decidiu-se experimentar este tipo de redes neuronais de modo a verificar se se obteria melhores resultados (ou semelhantes à LSTM) numa tentativa de encontrar o melhor modelo. A arquitetura da GRU desenvolvida tem exactamente a mesma estrutura, métricas, função de *loss* e otimizador que a LSTM, alterando-se apenas a camada bidireccional.

## 7.3 CNN

CNNs (*Convolution Neural Networks*) são reudes neuronais *feed-forward* (redes neuronais em que a conexão entre os nodos não forma um ciclo) muitas vezes definidas com 20 ou 30 camadas. O poder deste tipos de redes vêm da sua camada chamada *convolutional layer*. Estas camadas consiste num filtro que é aplicado a um determinado *input* com o objetivo de extrair alguma característica desse determinado *input*. Com três ou quatro camadas convolucionais é possível reconhecer dígitos escritos à mão e com 25 camadas é possível distinguir faces humanas [33]. Estas redes são usadas essencialmente para lidar com reconhecimento de imagem e vídeo, classificação de imagens, análise de imagens médicas e até processamento de linguagem natural [34], sendo que neste último campo têm apresentado sucesso na classificação de textos [33].

A ideia inicial de criar um modelo de previsão usando uma CNN partiu dos conhecimentos adquiridos nas Unidades Curriculares do perfil de Sistemas Inteligentes. No entanto, após alguma pesquisa, percebeu-se que devido ao formato do *dataset* criado, esta poderia não ser a melhor opção. Como após a implementação dos modelos mais recomendados no estado de arte, LSTM e GRU, se obteve resultados bastantes positivos nos dados de validação do *NetLang* (jornais do *Público e Sol*), optou-se por não focar no desenvolvimento nem afinação desta rede, acabando por se desistir da sua utilização.

## 7.4 Naive Bayes

Anjaria e Guddeti (2014) [35] [36] realizaram um projeto que culminou no desenvolvimento de um sistema de predição de resultados de eleições com o uso de dados do *Twitter* aliados à mineração de dados e à implementação de um algoritmo de *Naive Bayes* simples, apresentando resultados muito satisfatórios. Este é um modelo probabilístico supervisionado simples com base na regra de Bayes. Quando aplicado neste contexto, o seu objetivo é estimar a probabilidade de um texto ser positivo ou negativo, tendo em conta o seu conteúdo.

Partindo dos bons resultados demonstrado pelo algoritmo *Naive Bayes* em trabalhos prévios, procedeu-se a criação deste modelo e fez-se uso da função *MultinomialNB()* presente no ambiente de programação *sklearn*<sup>1</sup> do *Python*.

## 7.5 Logistic Regression

Um classificador de Regressão Logística, também conhecido como Máxima Entropia, baseia-se num modelo discriminativo que calcula a probabilidade de y ocorrer sabendo que x ocorre.

---

<sup>1</sup><https://scikit-learn.org/stable/>

Este modelo foi também definido à custa do ambiente de programação *sklearn* do *Python*, fazendo-se uso da função *LogisticRegression* a qual foi aplicado o *solver lbfgs* na otimização.

Após se ter selecionado e definido os modelos mais adequados ao problema que se pretendia resolver, passou-se ao processo de obtenção de resultados para cada um dos modelos. Estes resultados serão analisados em detalhe no próximo capítulo.



## Capítulo 8

# Resultados obtidos

Neste capítulo serão apresentados e discutidos os resultados obtidos tanto no treino como na validação dos modelos definidos. Serão também avaliados e validados os resultados obtidos com os jornais escolhidos para validação das previsões geradas pelos modelos.

### 8.1 Modelos

Para treino dos modelos, foram adotadas duas estratégias: passando apenas o conteúdo do *post* e passando outros parâmetros para além deste (como o número de comentários, e de *likes*). Apresentaremos os resultados para as duas abordagens.

#### 8.1.1 Univariáveis

A primeira abordagem consistiu em treinar modelos univariáveis, em que a única variável passada para treino do modelo era o conteúdo do *tweet*. A tabela 8.1 apresenta uma síntese dos resultados obtidos para os diferentes modelos. Facilmente se verifica que o modelo GRU obteve o valor de acurácia mais elevado, atingindo os 92.6 %. Por outro lado, o modelo de *Naive Bayes* apresentou o valor mais baixo, situando-se nos 78%.

Modelo	accuracy
GRU	92.6%
LSTM	91%
Naive Bayes Model	78%
Logistic Regression Model	91%

Tabela 8.1: Síntese dos resultados obtidos na validação dos modelos.

Analisando os gráficos 8.1 e 8.2, referentes à variação de *loss* e acurácia durante o treino dos modelos LSTM e GRU, percebe-se a semelhança dos resultados obtidos pelos dois modelos. No início do treino o valor de *loss* encontrava-se elevado, perto dos 0.7 e após 30 épocas passou a situar-se no intervalo [0.2 , 0.3]. Seguindo o mesmo padrão, o valor de acurácia começou a estabilizar para valores superiores a 90%, a partir da época 5.

A figura 8.3 apresenta a matriz de confusão obtida com o modelo de regressão linear e permite confirmar não só que a acurácia atingida pelo modelo foi de 91%

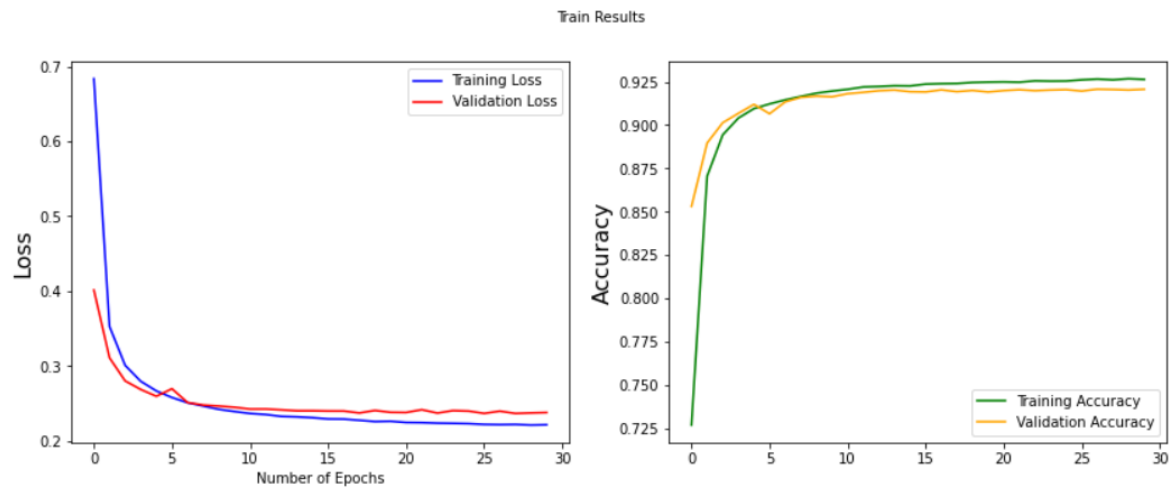


Figura 8.1: Resultados de treino do modelo LSTM ao longo de 30 epochs.

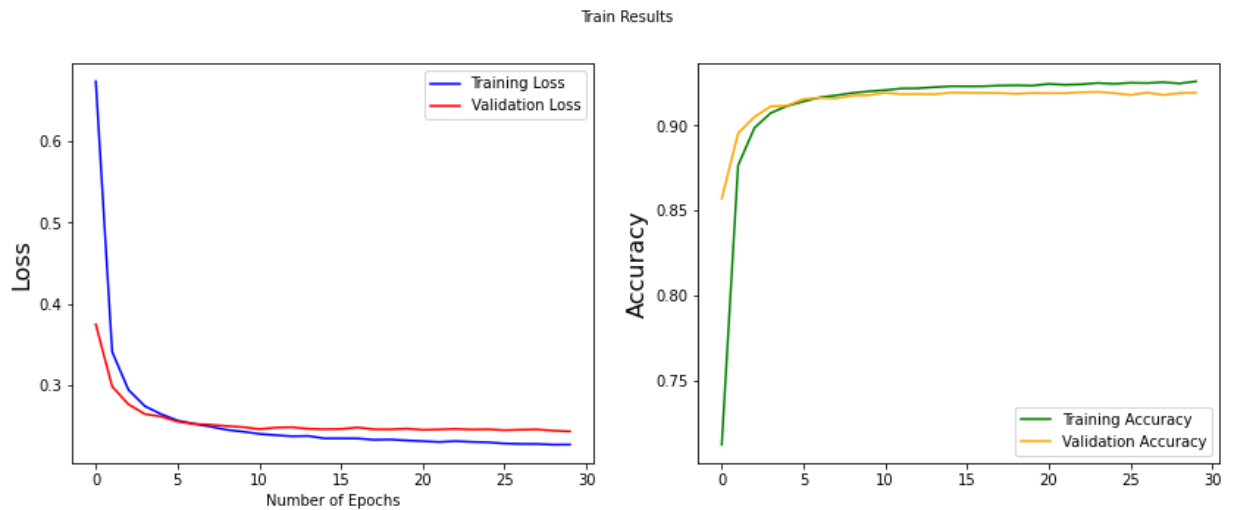


Figura 8.2: Resultados de treino do modelo GRU ao longo de 30 epochs.

mas também que os valores de precisão e *recall* são bastante positivos. Destaca-se os 97% de precisão obtidos quando a classificação é negativa, o que indica que em praticamente todas as previsões do modelo a avaliação de uma *tweet* com sentimento negativo corresponde efetivamente a um *tweet* classificado no *dataset* como negativo. Por outro lado, apesar do modelo ter identificado 95% de frases Neutras do *dataset*, apenas 84% das classificações *Neutral* estão corretas. Ou seja 16% dos *tweets* foram erradamente classificados como Neutros.

	precision	recall	f1-score	support
Negative	0.97	0.89	0.93	12441
Neutral	0.84	0.95	0.89	12226
Positive	0.93	0.89	0.91	12533
accuracy			0.91	37200
macro avg	0.91	0.91	0.91	37200
weighted avg	0.91	0.91	0.91	37200

Figura 8.3: Matriz de confusão obtida para o modelo de regressão linear.

Apesar de todos os modelos criados terem resultados muito positivos, tal como foi visto anteriormente, a GRU destaca-se por ter obtido valores ligeiramente superiores e por isso este foi utilizado como modelo final a integrar na aplicação *BomBedia*.

### 8.1.2 Multivariáveis

A segunda abordagem pensada para o treino dos modelos consistia na utilização de mais variáveis, numa tentativa de obtenção de melhores resultados. Para além do conteúdo do *tweet*, pretendia-se também passar como argumento o número de *likes*, data de publicação do *tweet*, número de *replies* e número de *retweets*.

No entanto, após se analisar o *dataset*, concluiu-se que grande parte dos valores nestes campos eram nulos, pelo que, ao invés de acrescentar informação útil ao resultado da classificação, só iriam tornar o treino destes modelos mais lento e possivelmente com piores resultados dos que os obtidos pelo modelo univariável (treinado apenas com o conteúdo dos *tweets*).

A adicionar a este facto, a correlação existente entre estes campos e a classificação era muito baixa, pelo que desistimos do treino dos modelos multivariáveis.

## 8.2 Processamento dos jornais

Com o objetivo de validar os resultados, procurou-se estabelecer uma conexão entre os resultados obtidos pelas ferramenta NetAC e BomBedia. Para isso, utilizou-se como ponto de partida notícias do jornal *Sol* e *Público*, previamente extraídas pela coordenadora deste projeto, professora Cristina Araújo. Idealmente, o grupo gostaria de ter utilizado mais jornais para teste, nomeadamente jornais que geram mais controvérsia como o *Correio da Manhã*. Não foi possível obter extrações deste jornal em concreto porque os comentários não estão acessíveis, não possibilitando a sua extração e, consequentemente, a comparação de resultados que se pretendia fazer.

A ferramenta NetAC <sup>1</sup> está construída assumindo um formato JSON específico para receber a notícia e por isso foi necessário aplicar conversores criados pelos alunos do professor Pedro Rangel Henriques, coordenador do projeto, no âmbito da Unidade Curricular de *Processamento de Linguagem e Compiladores*, para obter as notícias no formato pretendido.

Para poder submeter os ficheiros em formato JSON à ferramenta que continham os metadados e os comentários dos jornais foi apenas necessário aplicar o extrator fornecido para extrair os comentários e depois fazer uma junção aos metadados já obtidos previamente.

### 8.3 Validação dos resultados com o NetAC

Depois de realizar toda a preparação indicada no tópico anterior, foi possível fazer a análise das notícias na ferramenta NetAC e posteriormente desenvolver um *script* que permitisse, de forma eficiente, submeter todas as notícias ao modelo, extraíndo os resultados. Tendo em conta que em alguns casos o corpo da notícia era muito extenso, optou-se por se avaliar o texto associado ao título concatenado com o subtítulo da notícia, com o objetivo de melhorar a performance da execução do *script*.

A tabela 8.4 apresenta parte do documento síntese obtido, em formato PDF, gerado para mostrar, para cada notícia do Sol, o seu título, o nº de comentários que possui, a classificação da potencialidade para gerar controvérsia (Sim ou Não), a percentagem de *hate speech* calculada pelo NetAC e a classificação do Bombedia. Importa referir que o título da notícia, o número de comentários e a percentagem de *hate speech* foram extraídos dos resultados obtidos da submissão das notícias à ferramenta NetAC. Por outro lado, a potencialidade para gerar controvérsia (a existência de sentimentos positivos ou negativos associado ao conteúdo da notícia, explicado na secção 2.3) é obtida através da submissão ao modelo de ML desenvolvido pelo grupo (e explicado no Capítulo 7), que atribui **sim** no caso do resultado ser **Controverso** e **não** caso contrário. A classificação do Bombedia (coluna *Resultado Bombedia* da tabela apresentada na figura 8.4, não é nada mais do que um complemento de informação da métrica *potencialidade para gerar controvérsia* (obtida através do modelo), uma vez que amplifica o conhecimento sobre o texto, indicando se o mesmo é controverso no sentido positivo ou negativo.

Realizou-se uma análise semelhante para o jornal *Público* e *Youtube*. PEDRO: As linhas destacadas a cor verde correspondem àquelas que, sob o ponto de vista do grupo, foram correctamente classificadas como Controversas (por possuírem mais de 30 comentários ou uma percentagem de discurso de ódio superior a 1) ou não Controversas (por possuírem menos de 30 comentários ou menos de 1% de hate speech). Para se conseguir entender mais facilmente como definimos a Controvérsia, desenvolvemos as seguintes equações :

$$\text{Nº comentários} > 30 \ \&\& \ \% \text{ Hate Speech} > 1 \Rightarrow \text{Controverso}$$

$$\text{Nº comentários} < 30 \ \&\& \ \% \text{ Hate Speech} < 1 \Rightarrow \text{Não Controverso}$$

No anexo B é possível consultar toda a análise efectuada sobre as notícias extraídas para o Sol, Público e Youtube.

<sup>1</sup><http://netlang-corpus.ilch.uminho.pt:10100/>

É de notar, no entanto, que embora as métricas definidas estejam a ser aplicadas de igual forma às três fontes, é normal que o jornal Sol, por exemplo, contenha muito menos comentários associados ao *post* que o *Youtube*. Isto deve-se ao facto de este jornal ter sido criado para apresentar notícias de forma mais objetiva e não tão controversa, como é o caso de jornais como o Correio da Manhã. Já no *Youtube*, a movimentação das massas é muito maior e consequentemente, o número de comentários associados a cada vídeo é mais elevado. Assim sendo, é justificável a maior quantidade de notícias mal classificadas pelo modelo para o jornal Sol, que se pode verificar no gráfico 8.5.

Table 1: Síntese dos resultados por ficheiro

Título	N de comentários	Controverso	Hate Speech(%)	Resultado Bombedia
Esta solteira? Multimilionário britânico procura mulher para construir família e passar férias de luxo	12	nao	1,4433	Neutral
Manuel Luis Goucha escreve carta aberta ao pai	8	nao	2,1672	Neutral
Marido de Goucha deixa provocação a Joacine Katar Moreira: 'Sera que e xenofobia?'	24	nao	1,5421	Neutral
Ele nao esta na TVI e nao estara mais. Marido de Goucha expulso da estacao de Queluz de Baixo	22	nao	1,6746	Neutral
Marido de Goucha indignado com alojamento de migrantes em Lisboa: 'Portugueses sem-abrigo vaguem ao frio'	40	sim	0,5739	Positive
Manuel Luis Goucha defende Judite Sousa: 'Tenho vergonha de seres humanos assim'	20	sim	0,0	Negative
Goucha comenta piada do marido: 'Nem eu resisto a um arroz de pato ou a uma canja de pombo'	10	nao	0,0	Neutral
Manuel Luis Goucha brinca com telefonema de Marcelo a Cristina Ferreira	3	nao	0,0	Neutral
Claudio Ramos responde a marido de Goucha: 'Tenho 46 anos, uma filha para criar e uma carreira maior que a tua, Rui'	28	nao	0,7989	Neutral
Manuel Luis Goucha passa fim de semana na Holanda para assistir a provas de equitação	16	sim	1,9608	Negative

Figura 8.4: Tabela síntese com os resultados obtidos para o jornal Sol

Na figura 8.5 verifica-se que:

- Para o jornal *Sol* foram avaliadas cento e noventa e duas notícias. Do total das notícias avaliadas para este jornal, o modelo realizou uma previsão correta em 145 e falhou em 45.
- No jornal *Público* avaliaram-se 122 *posts*. O modelo errou a previsão em apenas 18 textos.
- Por fim, para os noventa e nove títulos de vídeos do *Youtube* analisados, o modelo classificou correctamente 81 títulos.

Os resultados permitem concluir que o modelo apresenta uma boa capacidade de deteção da presença ou não de controvérsia associado a um texto, possuindo maior dificuldade ao decidir se a controvérsia gerada é positiva ou negativa, o que leva aos erros obtidos.

Estes erros encontram-se essencialmente associados a textos cuja a controvérsia é gerada por Personalidades. A título de exemplo, se numa pesquisa no nosso modelo forem incluídas as palavra *Cristina Ferreira*, independentemente do conteúdo do texto, ele será classificado como *Controverso*. No entanto, a frase poderá não estar a fazer referencia à apresentadora Cristina Ferreira e sim referir-se a outra pessoa que com ela partilhe o nome. Consideramos que esta falha pode estar relacionada com o facto

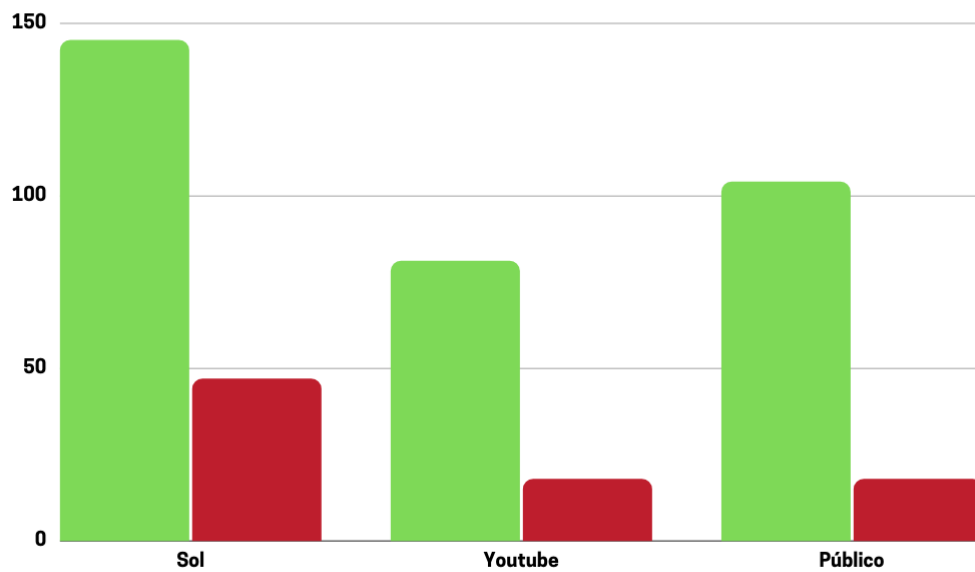


Figura 8.5: Validação dos resultados obtidos pela ferramenta NetAc.

de a apresentadora se destacar por ser uma das Personalidade Portuguesas com mais seguidores nas suas redes sociais e a quem está associado um elevado fluxo de comentários/*posts* controversos. Acresce ainda o facto de estar constantemente no topo dos assuntos do momento no *Twitter* o que lhe garante muitos *posts* com o seu nome no *dataset* usado para treino do modelo.

## Capítulo 9

# Bombastic Media – *Web App*

Concluídas as fases de desenvolvimento e avaliação dos modelos preditivos, passamos agora a explicar, em detalhe, a aplicação *web* que foi desenvolvida para integrar a ferramenta descrita anteriormente. Ao longo deste capítulo, começaremos por enumerar os requisitos estabelecidos pelo grupo para a aplicação, passando à explicação da arquitetura desenvolvida para a criação da *web app* e, por fim, a apresentar a aplicação em si.

### 9.1 Requisitos da *Web App*

De forma a pensar na arquitetura que iria ser desenvolvida e os passos que seriam precisos seguir para obter a aplicação pretendida, começou-se por elaborar uma lista de requisitos inicial para a aplicação. O grupo estabeleceu as seguintes metas para a *Web App*:

- Possibilidade de autenticação para que os utilizadores tenham a oportunidade de guardar um histórico de *posts* analisados na ferramenta de classificação;
- Integração da ferramenta principal, o analisador, que seja capaz de obter a classificação de *post*, recorrendo ao modelo final desenvolvido;

Após inicializar o processo de desenvolvimento da aplicação, foram definidos mais requisitos, por se ter considerado serem importantes para uma melhor utilização da *app*.

- Uma funcionalidade que permita, dada uma palavra, obter palavras positivas ou negativas associadas a essa palavra.
- Integração de dados fornecidos pelo *Google Trends* de modo a que os utilizadores possam saber quais os temas ”mais quentes” actualmente.

## 9.2 Público Alvo (Utilizadores)

A aplicação foi desenhada tendo em conta um público-alvo específico. Pretende-se que os utilizadores principais da aplicação sejam os que se apresentam listados a seguir, embora não estejam restritos a estes:

- Pessoas na área de estudo de Ciências Sociais relacionadas com CMC;
- Imprensa nacional que pretenda avaliar a capacidade de uma notícia gerar ou não controvérsia antes de ser publicada;
- Os próprios criadores dos *posts*.

Tendo em conta o público alvo definido, optou-se por utilizar um *design* gráfico simples e intuitivo, de modo a facilitar a interação com a plataforma.

## 9.3 Arquitetura

A arquitetura desenvolvida para a aplicação *Web* é orientada aos microserviços, possuindo 5 servidores. A figura 9.1 apresenta a arquitetura desenvolvida que será explicada em detalhe de seguida.

- Servidor de autenticação
- Servidor da API
- Servidor da APP
- Servidor de Flask
- *Tensorflow Serving*

A arquitetura foi pensada desta forma por dois grandes motivos:

- O facto de o grupo já ter desenvolvido uma aplicação com esta estrutura no passado, pelo que foi fácil adaptar para os requisitos deste projeto.
- A necessidade de incorporação de código Python e do *TensorFlow* não deveria diminuir a performance do tempo de resposta da aplicação pelo que seria uma vantagem manter cada um num microserviço.

Os pontos listados anteriormente motivaram, consequentemente, a escolha das tecnologias utilizadas.



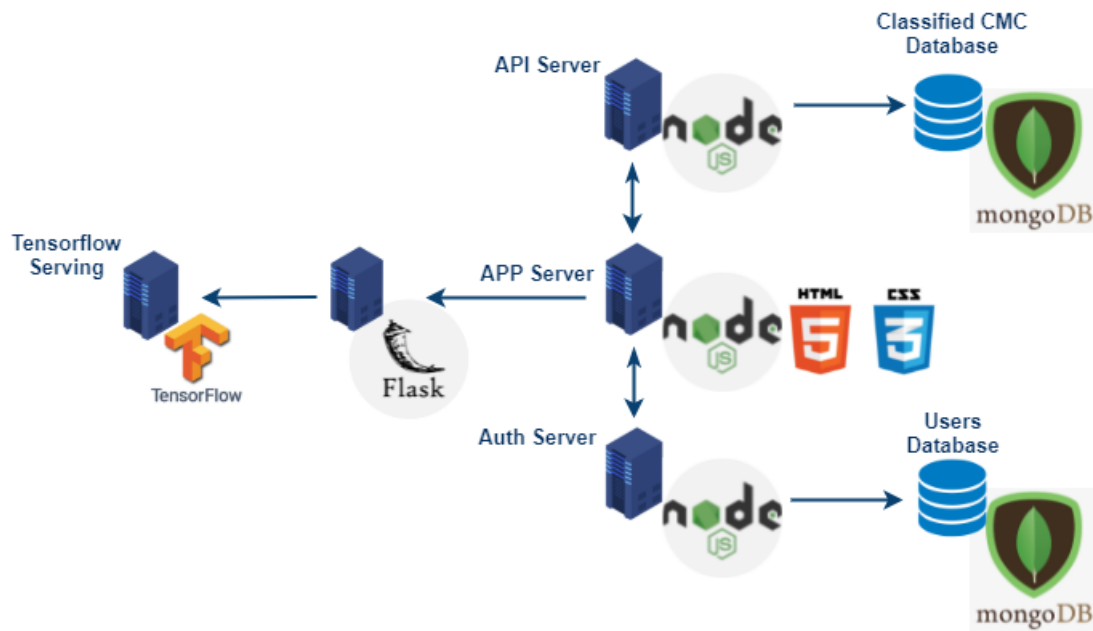


Figura 9.1: Arquitetura da Web App Bombedia.

### 9.3.1 App Server

O servidor da aplicação é o servidor principal e contém todo o *frontend* responsável pela interação com o utilizador. Foi desenvolvido em *NodeJS*.

### 9.3.2 API Server

O *API Server* é responsável por suportar os pedidos feitos à API para obter o histórico dos ficheiros classificados a pedido de um determinado utilizador. Este servidor está conetactado a uma base de dados em *Mongo*.

### 9.3.3 Authentication Server

Este servidor, também desenvolvido em *NodeJS*, permite que os utilizadores da plataforma se possa autenticar de forma a poderem guardar um histórico dos *posts* que forem classificados a pedido dos mesmos. Sendo assim, o *Auth server* possui uma conexão a uma base de dados em *Mongo* que guardar as informações referentes a cada utilizador. Aquando da autenticação, este devolve um *JSON Web Token* ao servidor da aplicação (*app server*) de forma a que as rotas sejam autenticadas.

### 9.3.4 TensorFlow Serving

Para obter as previsões dos modelos, desenvolveu-se inicialmente um *script* em *Python* que importava o *TensorFlow* para poder utilizar o modelo e fazer as previsões. No entanto, percebeu-se que esta solução era extremamente ineficiente. Sempre que se corria este *script* para fazer uma previsão, a inicialização do *TensorFlow* demorava muito tempo, fazendo com que a resposta ao pedido no *site* demorasse cerca de 10 segundos. Assim sendo, foi necessário procurar uma solução para resolver este problema.

Optou-se por utilizar o *TensorFlow Serving*.

O *TensorFlow Serving* é um sistema flexível e de alto desempenho para servir modelos de ML, concebido para ambientes de produção. Este mantém a mesma arquitectura de servidores e APIs. Pensado para a integração com o *TensorFlow*, pode ser alargado para servir outros tipos de modelos. [26].

### 9.3.5 Servidor *Flask*

Uma vez que foi necessário correr os modelos de ML na plataforma *web* e estes estão implementados em *Python*, decidimos exportar a sua utilização para um microserviço em *Flask*, conectando o mesmo à plataforma.

## 9.4 Aplicação Final

Vamos agora passar à apresentação da aplicação desenvolvida, que pode ser acedida em: <http://netlang-corpus.ilch.uminho.pt:10200/>. O *design* da plataforma foi inspirado no *Smash Lite Template* de *UIdeck* <sup>1</sup>.

### 9.4.1 Analisador

A funcionalidade principal da aplicação é o analisador. Este analisador consiste num formulário que recebe um texto de CMC e devolve o resultado da previsão do modelo para a capacidade que o *input* passado tem para gerar controvérsia.

Para obter esta classificação adotou-se o seguinte processo: ao receber um texto como *input*, faz-se uma primeira verificação para averiguar se este contém termos que estejam incluídos nos termos mais procurados nos últimos 3 dias (usando para isso os dados do *Google Trends*<sup>2</sup>). Caso contenha, então é considerado **controverso** e é passado pelo *NRC* para verificar se o sentimento associado é **positivo** ou **negativo**, que indica se os comentários serão maioritariamente a favor ou contra. Caso não contenha, então é submetido ao modelo de *ML* de forma a obter a previsão. De forma a esquematizar este processo, a figura 9.2 contém a *pipeline* de classificação adotada e que foi agora descrita. Inicialmente não era feita esta pré-classificação utilizando o *Google Trends*, mas decidiu-se implementar este passo para melhorar a *performance* de resposta da aplicação.

A figura 9.3 apresenta a página do *Bombedia* que contém o analisador. O resultado que esta ferramenta retorna pode ser **Controverso** ou **Não Controverso**. No caso de ser considerado controverso, então é retornada também uma de duas classificações: *positivo* ou *negativo*. Esta segunda classificação, como já foi sendo explicada ao longo do relatório permite inferir se os comentários gerados serão a favor ou contra o *post* (sendo neste último caso comentários "tóxicos").

---

<sup>1</sup><https://onpagelove.com/smash-lite>

<sup>2</sup><https://trends.google.pt/trends/>

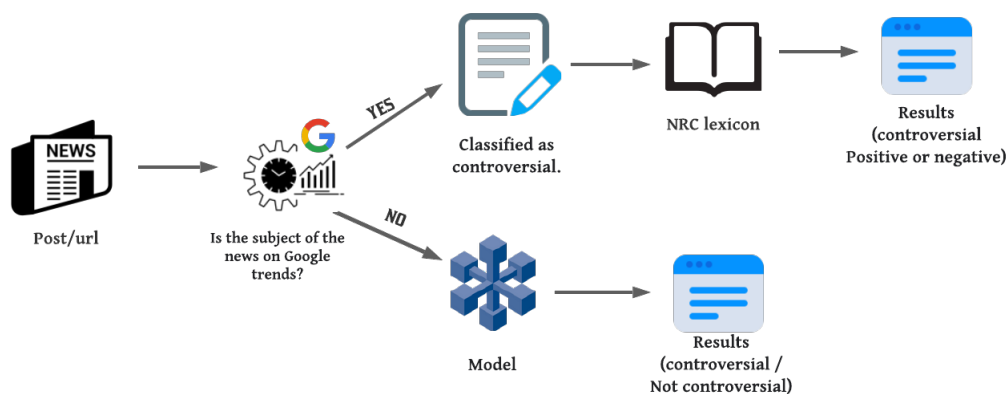


Figura 9.2: *Pipeline* otimizada da classificação de *posts*



Figura 9.3: Analisador de *posts*

Vejamos agora um exemplo de classificação obtida para um *post* em concreto. A figura 9.4 apresenta o resultado obtido para a notícia *Autarca de Borba vai ser julgado por cinco crimes de homicídio* <sup>3</sup>. Neste caso, o analisador classifica esta notícia como controversa no sentido negativo, ou seja, "tóxica", com capacidade de gerar comentários negativos.

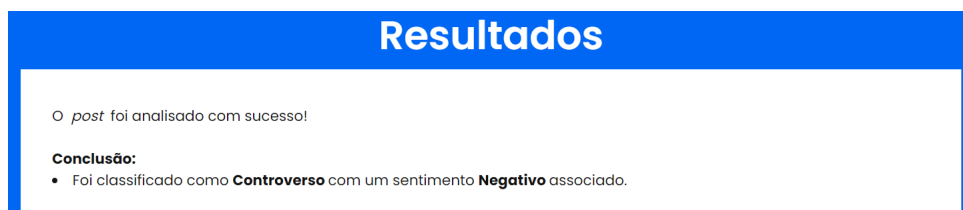


Figura 9.4: Exemplo de um resultado fornecido pelo analisador

<sup>3</sup><https://www.jn.pt/justica/autarca-de-borba-vai-ser-julgado-por-cinco-crimes-de-homicidio-13838733.html>

### 9.4.2 Analisador por *URL*

Opcionalmente, pode-se analisar um *post* de uma notícia utilizando o *URL* da mesma.

Para implementar este analisador recorreu ao módulo `newspaper` do *Python*. Este módulo permite extrair as várias componentes que compõem uma notícia associada a um determinado *URL*, como o título, os autores, a data, e o texto. Desta forma, é extraído o conteúdo da notícia e passado ao classificador.

Para além disso, o `newspaper` é ainda capaz de reconhecer a língua em que vem escrita a notícia na maioria das vezes, o que é uma vantagem caso se pretenda alargar a ferramenta a mais idiomas.

## 9.5 Pesquisa por Palavras Controversas

Para além da funcionalidade principal, o analisador, decidiu-se adicionar uma ferramenta que permitisse obter palavras controversas (isto é, associadas a um sentimento positivo ou negativo) a partir de uma determinada palavra que fosse passada como *input*. Assim, pensando no público alvo da aplicação, um utilizador que queira tornar *post* que seja neutro, ou não controverso, num *post* controverso, pode pesquisar por determinados sinónimos que sejam considerados positivos ou negativos conforme o pretendido.

Para implementar esta funcionalidade recorreu-se aos módulos `spacy` e `gensim` (mais concretamente o `word2vec`) do *Python*. Como dicionário utilizou-se o *NRC*. O módulo `word2vec` utiliza uma rede neuronal capaz de associar palavra semanticamente semelhantes. Ao obter uma lista de palavras associadas à palavra passada como *input* bastou depois apenas verificar se estas eram consideradas pelo *NRC* como positivas ou negativas.

A figura 9.5 apresenta a página do Bombedia que contém esta ferramenta.

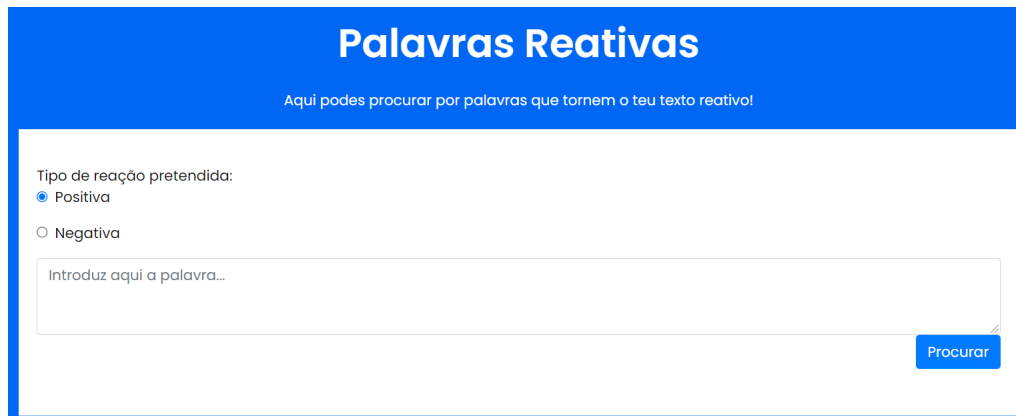


Figura 9.5: Funcionalidade de procura por palavras controversas no Bombedia

## 9.6 Temas Quentes

### 9.6.1 Temas mais procurados

De modo a poder obter uma classificação atualizada no tempo, decidiu-se incorporar os dados fornecidos pelo *Google Trends* na avaliação que é feita, como foi explicado na secção 9.4.1. Para dar ao utilizador a possibilidade de saber quais são os temas mais procurados ou que dão que falar atualmente, decidiu-se incluir na aplicação a funcionalidade que explicaremos de seguida. Através do módulo `pytrends`<sup>4</sup> do *Python*. Este módulo permite obter dados que são fornecidos pelo *Google Trends* e que permitem saber quais os temas mais procurados no *Google* num determinado intervalo de tempo e localização. Uma vez que, atualmente, este módulo deixou de ter a funcionalidade de extrair dados de um determinado dia sem ser o dia em que se efetua a extração, criou-se um *script* que extrair os dados dos temas mais procurados em cada dia e guarda num ficheiro em formato *csv* de modo a resolver o problema.

Na figura 9.6 é possível observar um excerto da página que possui a funcionalidade descrita. Neste caso apresentam-se os temas mais procurados na região de Portugal para o dia 16 de junho de 2021.



The screenshot shows a web interface with a blue header titled "Temas mais procurados". Below the header is a white box containing a date selector labeled "Selecione uma data:" with a text input "dd/mm/yyyy", a calendar icon, and a blue "Procurar" button. Below this is a table with two columns: "Ranking" and "Tópico". The table lists the top 6 search topics for June 16, 2021, in Portugal.

Ranking	Tópico
1	Britney Spears
2	França-Alemanha
3	Portugal
4	SIC
5	Portugal FC
6	TVI

Figura 9.6: Apresentação dos 6 temas mais procurados no *Google* no dia 16 de junho de 2021

### 9.6.2 Análise de Temas ao longo do tempo

Para além da saber quais os temas mais procurados diariamente, o *Google Trends* permite também saber qual foi o interesse que determinado tema teve ao longo do tempo. Esta informação também é bastante relevante para a nossa análise, embora não a estejamos a usar diretamente. Seria uma exploração interessante a nível de trabalho futuro.

Neste separador, o utilizador pode introduzir um tema num formulário e é retornado o gráfico que indica o quanto esse termo foi procurado no *Google* ao longo dos últimos cinco anos.

A figura 9.7 apresenta a página do *Bombedia* que apresenta esta funcionalidade e contém um gráfico como exemplo. Este gráfico indica a popularidade que o termo

<sup>4</sup><https://pypi.org/project/pytrends/>

”coronavírus” teve nas pesquisas do *Google* nos últimos 5 anos, permitindo verificar que este só começa a ser ”quente” no final de 2020. A escala no eixo dos  $y$  baseia-se no volume absoluto de pesquisas do termo num dia relativamente ao volume absoluto de pesquisa no *Google* no mesmo dia e estes valores são normalizados para uma escala de 0 a 100.

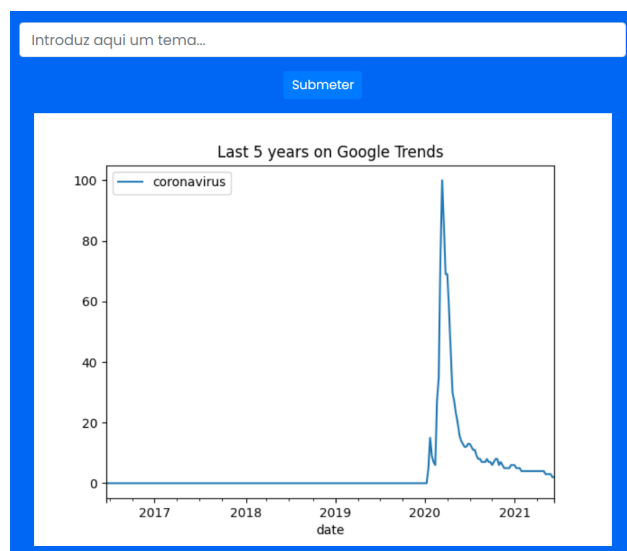


Figura 9.7: Índice de pesquisas para o termo ”Coronavírus” ao longo dos últimos cinco anos

## 9.7 Manter a classificação atualizada

Um dos problemas futuros deste projeto é a capacidade do modelo fazer previsões atualizadas no tempo. Hoje em dia, a informação navega de um canto ao outro do mundo com muita facilidade e com rapidez, sendo que aquilo que é considerado relevante e controverso hoje pode não o ser amanhã. Nesta subsecção pretende-se explicar de que forma é que este problema foi, pelo menos em parte, resolvido.

Como ferramenta de apoio, decidiu utilizar a API do *Google Trends* de modo a conseguir obter informação sobre os temas mais procurados atualmente. Com esta informação e com a ajuda do *Twint*, foram retirados 100 *tweets* relacionados com o tema mais procurado em cada dia da semana. Após isto, cada um dos *tweets* foi classificado com o mesmo método com que o *dataset* é classificado. Adicionados estes *tweets* ao *dataset* utilizado para treinar o modelo, o novo *dataset* fica então atualizado com os temas mais relevantes ao longo do tempo.

Com o *dataset* já atualizado, basta treinar outra vez o modelo para este poder obter resultados com os novos *tweets* recolhidos. Para a execução deste processo, foi criado um simples *script* que trata de carregar o *dataset*, fazer a preparação de dados necessária e de seguida treinar e guardar o modelo.

Em termos da periodicidade com que é feita esta atualização, o grupo considerou em fazer uma atualização diária com os 5 tópicos mais falados de cada dia. No entanto, tendo em conta que o nosso *dataset* tem por volta de 200 mil entradas, seria necessário incluir, no mínimo, 100 *tweets* de cada tópico, ou seja, o *dataset* aumentaria todos os dias 500 entradas. Ao fim de algum tempo, este processo iria aumentar demasiado

o tamanho do *dataset*. Tendo isto em conta, o grupo decidiu fazer uma atualização semanal, recolhendo *tweets* apenas do tópico mais falado de cada dia da semana. Sendo assim, o *dataset* só aumenta 700 entradas por cada semana, sendo este número mais razoável comparativamente ao anterior.

De modo a tornar este processo automático, decidiu-se criar um *script* para *bash* que trata de correr os ficheiros necessários para atualizar o *dataset* e para o treino do nosso modelo com o *dataset* atualizado. Para além disso, utilizou-se o *Cron* para realizar o agendamento semanal da execução deste *script*, tornando esta atualização 100% automática.

A figura 9.8 esquematiza, de forma simples, o processo desenvolvido e explicado para a atualização do *dataset*.

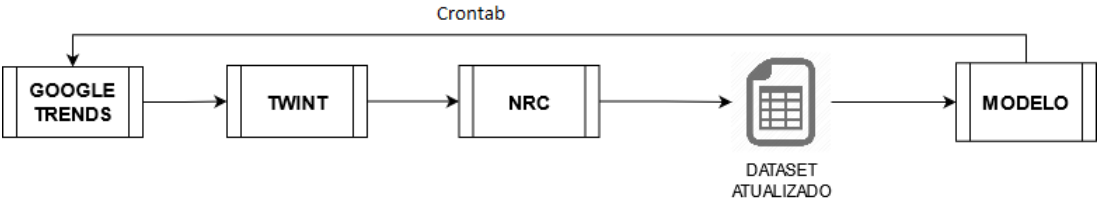


Figura 9.8: *Pipeline* do processo de atualização do *dataset*

A título de exemplo, apresenta-se de seguida a primeira atualização feita ao *dataset* com sucesso. A figura 9.9 mostra um fragmento do *dataset* inicial, antes de ser atualizado, e a figura 9.10 apresenta parte do *dataset* já atualizado, com os temas mais procurados na semana de 14 a 20 de junho. Como se pode verificar nas figuras, o primeiro *dataset* contém 186000 entradas enquanto que o segundo já contém 186694. Desta forma, verifica-se que o *dataset* está a ser atualizado tal como pretendido. Feita esta atualização, o modelo é, de seguida, treinado com o novo *dataset* de modo a obter previsões mais atualizadas.

Unnamed: 0			tweet	classification
0	0	@Ruancosta26	juntão irmão	1
1	1	@namelessapr	PARABÉNS 🍷	2
2	2	25 DE ABRIL CARALHO FELIZ ANIVERSÁRIO TED MOS...		1
3	3	@Vinhai_10 @TiagoLeite015	Se ele se lembrar li...	1
4	4		Amo o Ristovski	2
...	...		...	...
185995	185995	@StarFox_i @FriedHardt @masterchefbr	Se teve a...	0
185996	185996		Ei, Merda!... pede pra fazer xixi e sai de fin...	0
185997	185997		Queixa De Ligt de falta, e acaba de reduzir Sc...	0
185998	185998	@ThPogba06	volta lingard que eu perdôo a facad...	0
185999	185999		No és la primera vegada, ni será la última, qu...	0

186000 rows x 3 columns

Figura 9.9: Dataset antes da atualização

Unnamed: 0		tweet	classification
0	0.0	@Ruancosta26 juntão irmão	1
1	1.0	@namelessapr PARABÉNS 🍷	2
2	2.0	25 DE ABRIL CARALHO FELIZ ANIVERSÁRIO TED MOS...	1
3	3.0	@Vinha1_10 @TiagoLeite015 Se ele se lembrar li...	1
4	4.0	Amo o Ristovski	2
...	...	...	...
186443	NaN	boa noite só p quem faz shows no quarto ao som...	2
186444	NaN	E pronto, agora tenho uma crush na Britney Spears	1
186445	NaN	Depeche mode: don't know :( Britney spears: do...	2
186446	NaN	@gonattago @dicas_de_ingles Britney Spears 🍷	2
186447	NaN	Il est 2h30 du Mat' et Je viens de tomber sur ...	2

186448 rows x 3 columns

Figura 9.10: Dataset após a atualização



## Capítulo 10

# Alternativas, Decisões e Problemas de Implementação

Após a apresentação e descrição das etapas desenvolvidas para obter o produto final, pretende-se, neste capítulo, explicar algumas decisões tomadas que não foram referidas anteriormente. Para além disso, é objetivo desta secção descrever os problemas de implementação que ocorreram assim como refletir sobre as possíveis alternativas para a resolução do projeto. Algumas decisões foram sendo referenciadas ao longo deste relatório, pelo que aqui é feito um resumo destas.

### 10.1 Decisões

#### 10.1.1 Dataset

Relativamente à escolha do *dataset*, a primeira decisão tomada foi utilizar um *dataset* já avaliado para análise de sentimentos. Esta não foi a melhor decisão pois os *datasets* existentes estão mal classificados, apenas de acordo com os *emojis* que continham. Assim sendo, foi necessário tomar a decisão de descartar este *dataset* e criar, de raiz, um novo, classificado por nós, que levou à obtenção de muito melhores resultados. Para a obtenção do novo *dataset*, os conselhos dados pelo Dr. Ricardo Martins foram muito importantes. Estes consistiram em: utilizar a ferramenta *Twint* para a obtenção de *tweets* (inicialmente o plano consistia em usar a API do *Twitter*, mas o Dr. Ricardo explicou-nos que esta não seria a melhor opção para o nosso projeto); e utilizar o dicionário do NRC *Lexicon* para conseguir classificar corretamente o *dataset*, sendo esta, possivelmente, a parte mais importante do nosso projeto. Este processo permitiu que tivéssemos um maior controlo sobre o que o modelo iria aprender, o que fez com que obtivéssemos melhores resultados. Para treinar os modelos de *Machine Learning* é fundamental que o *dataset* utilizado se encontre bem classificado.

#### 10.1.2 Otimização da classificação do *dataset*

Como foi referido no capítulo 5, decidiu-se otimizar a classificação do *dataset* após se ter constatado que algumas previsões feitas pelo modelo para algumas notícias não estavam a ser bem feitas. Para resolver este problema, criou-se um *corpus* com palavras que, apesar de serem neutras, poderão ser consideradas controversas em determinado

contexto. Com esta fase adicional, foi possível obter uma melhor classificação dos *tweets* no que se refere à sua capacidade de gerar controvérsia.

### 10.1.3 Ferramentas

Em termos de ferramentas, decidiu-se desenvolver o trabalho em *Python* recorrendo ao *TensorFlow* para o desenvolvimento da parte de ML, não só por serem as ferramentas mais utilizadas na área de ML no mundo [37], como também pela familiaridade que o grupo tem com estas ferramentas. No que refere ao desenvolvimento da plataforma *Web*, recorreu-se ao *NodeJs* (com *ExpressJS*) com uma conexão a uma base de dados em *Mongo*. Recorreu-se também ao *Flask* para desenvolver um microserviço responsável por executar o código *Python* que tratava da parte do processamento dos *textos* e integração do *Google Trends*, tornando o processo mais eficiente.

### 10.1.4 Modelos de *Machine Learning*

Como já foi explicado anteriormente, os modelos que se decidiu utilizar derivaram da pesquisa que foi feita no que se refere ao estado de arte. Este estudo previamente feito mostrou que estes modelos seriam os melhores para o tipo de classificação que se pretendia prever.

## 10.2 Problemas de Implementação

### 10.2.1 *Encoding*

No que se refere à criação dos ficheiros em formato JSON das notícias que se pretendia avaliar, obtiveram-se vários problemas de *encoding*. Os ficheiros em formato JSON estavam a ser abertos com "ISO-8859" pelo que os pdfs gerados na ferramenta NetAC estavam a aparecer com os acentos mal apresentados. Para colmatar estes problemas na geração dos pdfs, uma vez que a fonte utilizada no *pdflatex* reconhece um número reduzido de caracteres, passou-se a utilizar o *xelatex* e a garantir que os ficheiros eram em guardados em *UTF-8*. Para converter os ficheiros que não estavam codificados desta forma, utilizou-se o comando:

```
iconv -f CP1250 -t UTF-8 < x.txt > x.utf8.txt
```

## Capítulo 11

# Conclusão

Este trabalho pretendeu analisar o impacto que determinado texto de *media*, em português, tem nos seus leitores. Para isso foi feita não só uma análise de sentimentos associados ao conteúdo do texto mas também um confronto desse mesmo conteúdo com os temas atuais, que são considerados "quentes" ou controversos, de forma a obter uma classificação mais credível e atualizada no tempo. Tendo em conta o objetivo do projeto, achou-se adequado a atribuição do nome **Bombedia** (*Bombastic Media*), que faz alusão à funcionalidade de previsão de controvérsia que os *posts* de *media* poderão ou não gerar.

Foi criado um *dataset* e classificado de forma a treinar vários modelos de *Machine Learning* para obter um modelo final. A nível de resultados, a *GRU* foi a que obteve melhores valores, tendo sido selecionada como modelo final que iria servir a aplicação *web* desenvolvida. No que se refere à validação dos resultados utilizando notícias de jornais portugueses, pode-se concluir que a classificação obtida é maioritariamente satisfatória, havendo mais classificações incorretas em *posts* que se relacionam com figuras públicas ou temas envolvendo ciganos, homofobia e transgenerismo. Estes últimos temas foram incluídos no *Corpus* definido para otimizar a classificação, o que permitiu colmatar o problema em parte.

Sendo este um projeto de engenharia, o grupo não ficou pela abordagem da Inteligência Artificial para resolver o problema abordado neste projeto e decidiu efetuar uma classificação com recurso ao *Google Trends* para obter os temas que "dão que falar" no momento, efetuando assim uma análise dos *posts* mais eficiente.

A nível de trabalho futuro, várias otimizações poderiam ser feitas. No que se refere à atualização do *dataset*, de modo a este não ficar saturado com temas antigos que já não são considerados reativos atualmente, poderia ser pensada uma forma de retirar estes dados do *dataset*, melhorando a classificação da ferramenta e diminuindo o tempo de execução do próprio modelo. Para além disso, seria necessário garantir também que o *dataset* continuasse balanceado. Outra questão a estudar futuramente seria o tratamento de comentários com ironia uma vez que são difíceis de serem detetados. No entanto, para o contexto de notícias de jornais, esse caso não era tão comum e por isso não chegou a ser explorado neste trabalho. Seria interessante também estudar o impacto dos *emojis* na classificação do *dataset*, uma vez que foi algo que na classificação efetuada se decidiu retirar. A nível da fonte de dados, poderiam ser obtidos dados de outro tipo de CMC, para além do *Twitter*, para enriquecimento do *dataset*.

Ainda referente à classificação binária que é feita em termos da capacidade que um *post* tem para gerar controvérsia, uma abordagem complementar poderia ser feita. Esta

consistiria em avaliar mais metadados de um determinado *post*, como a data e hora de publicação, os autores, o dia da semana, com inspiração nos artigos analisados no que se refere ao estado de arte. Para além disso, seria também importante avaliar a variação de sentimentos ao longo do *post* (e não apenas obter o sentimento predominante associado ao texto), de forma a melhorar a classificação que é feita (como é demonstrado em [18]).

A nível da definição dos modelos, uma afinação dos hiperparâmetros a utilizar poderia ser feita. No que se refere à língua que foi alvo de estudo, embora o grande foco deste trabalho fosse a análise de textos em português, o projeto foi desenvolvido de forma generalizada de forma a poder ser aplicado a outras línguas, pelo que seria interessante expandir a *web app* para a possibilidade de avaliar *posts* noutras línguas, alargando o âmbito inicial do projeto.

# Bibliografia

- [1] Cruz, Luís Braga da (2020), Prediction of toxicity-generating news using machine learning, Faculdade de Engenharia do Porto, <https://repositorio-aberto.up.pt/bitstream/10216/128539/2/412375.pdf>.
- [2] McQuail, Denis (2005). *McQuail's Mass Communication Theory*. SAGE. ISBN 978-1-4129-0372-1
- [3] Thurlow, Crispin; Lengel, Laura; Tomic, Alice (2004). *Computer Mediated Communication*. SAGE. ISBN 978-0-7619-4954-1
- [4] Walther, Joseph B. (1 February 1996). "Computer-Mediated Communication: Impersonal, Interpersonal, and Hyperpersonal Interaction". *Communication Research*. 23 (1): 3–43. doi:10.1177/009365096023001001. S2CID 152119884
- [5] Walther, Joseph B.; Burgoon, Judee K. (1992). "Relational Communication in Computer-Mediated Interaction". *Human Communication Research*. 19 (1): 50–88. doi:10.1111/j.1468-2958.1992.tb00295.x. hdl:10150/185294
- [6] Skovholt, Karianne; Grønning, Anette; Kankaanranta, Anne (1 July 2014). "The Communicative Functions of Emoticons in Workplace E-Mails: :-)". *Journal of Computer-Mediated Communication*. 19 (4): 780–797. doi:10.1111/jcc4.12063
- [7] Garcia, Angela Cora; Jacobs, Jennifer Baker (1 October 1999). "The Eyes of the Beholder: Understanding the Turn-Taking System in Quasi-Synchronous Computer-Mediated Communication". *Research on Language and Social Interaction*. 32 (4): 337–367. doi:10.1207/S15327973rls3204.2
- [8] Herring, Susan (1 June 1999). "Interactional Coherence in CMC". *Journal of Computer-Mediated Communication*. 4 (4). doi:10.1111/j.1083-6101.1999.tb00106.x. S2CID 5070516
- [9] Lisa Branz and Patricia Brockmann. 2018. Sentiment Analysis of Twitter Data: Towards Filtering, Analyzing and Interpreting Social Network Data. In *Proceedings of the 12th ACM International Conference on Distributed and Event-based Systems (DEBS '18)*. Association for Computing Machinery, New York, NY, USA, 238–241. DOI:<https://doi.org/10.1145/3210284.3219769>
- [10] Carvalho, J., Plastino, A. On the evaluation and combination of state-of-the-art features in Twitter sentiment analysis. *Artif Intell Rev* 54, 1887–1936 (2021). <https://doi.org/10.1007/s10462-020-09895-6>

- [11] Z. Jianqiang, G. Xiaolin and Z. Xuejun, "Deep Convolution Neural Networks for Twitter Sentiment Analysis," in IEEE Access, vol. 6, pp. 23253-23260, 2018, doi: 10.1109/ACCESS.2017.2776930.
- [12] Basile, Angelo & Caselli, Tommaso & Nissim, Malvina. (2017). Predicting Controversial News Using Facebook Reactions. 10.4000/books.aaccademia.2370.
- [13] Gambino OJ, Calvo H. Predicting emotional reactions to news articles in social networks. Comput Speech Lang. 2019;58:280–303.
- [14] He, Lihong & Shen, Chen & Mukherjee, Arjun & Vucetic, Slobodan & Dragut, Eduard. (2020). Cannot Predict Comment Volume of a News Article before (a few) Users Read It.
- [15] Choi, Yoonjung & Jung, Yuchul & Myaeng, Sung-Hyon. (2010). Identifying Controversial Issues and Their Sub-topics in News Articles. 6122. 140-153. 10.1007/978-3-642-13601-6\_16.
- [16] Beelen, Kaspar & Kanoulas, Evangelos & Velde, Bob. (2017). Detecting Controversies in Online News Media. 1069-1072. 10.1145/3077136.3080723.
- [17] A. Sriteja, P. Pandey and V. Pudi, "Controversy Detection Using Reactions on Social Media," 2017 IEEE International Conference on Data Mining Workshops (ICDMW), 2017, pp. 884-889, doi: 10.1109/ICDMW.2017.121.
- [18] Kaplun, Kateryna & Lebeknight, Chris & Feldman, Anna. (2018). Controversy and Sentiment: An Exploratory Study. 10.1145/3200947.3201016.
- [19] W. CHRISTHIE (2015), SENTIMENTALL: Ferramenta para análise de sentimentos em português, Centro Universitário Luterano de Palmas (CEULP/ULBRA)
- [20] DE PELLE, Rogers Prates; MOREIRA, Viviane P. (2017), Offensive Comments in the Brazilian Web: a dataset and baseline results. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BRASNAM), 6. , São Paulo. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação,. ISSN 2595-6094. DOI: <https://doi.org/10.5753/brasnam.2017.3260>.
- [21] Carvalho, C. M. A., Nagano, H., & Barros, A. K. (2017, October). A comparative study for sentiment analysis on election Brazilian news. In Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology (pp. 103-111)
- [22] Martins, Ricardo & Almeida, José & Henriques, Pedro & Novais, Paulo. (2020). Predicting an Election's Outcome Using Sentiment Analysis. 10.1007/978-3-030-45688-7\_14.
- [23] Sayce, David. (2020, December 16), The Number of tweets per day in 2020, <https://www.dsayce.com/social-media/tweets-day/>, acedido em: 16 junho de 2021
- [24] Ahlgren, Matt. (2021, March 2021), 50+ TWITTER STATISTICS & FACTS FOR 2020, *Website Hosting Rating*, <https://www.websitehostingrating.com/twitter-statistics/>, acedido em: 16 junho de 2021

- [25] Morikawa, Rei. (2019, June 3), 12 Best Social Media Datasets for Machine Learning, *Lion Bridge*, <https://lionbridge.ai/datasets/12-best-social-media-datasets/>,  
acedido em: 16 de junho de 2021
- [26] Rawlani, Himanshu. (2018, October 18), *Towards Data Science*,  
<https://towardsdatascience.com/deploying-keras-models-using-tensorflow-serving-and-flask->,  
acedido em: 18 de julho de 2021
- [27] Tsagkias, Manos & Weerkamp, Wouter & Rijke, Maarten. (2009). Predicting the volume of comments on online news stories. *International Journal of Press-politics - INT J PRESS-POLIT.* 1765-1768. 10.1145/1645953.1646225.
- [28] Tsagkias, Manos & Weerkamp, Wouter & Rijke, Maarten. (2010). News Comments: Exploring, Modeling, and Online Prediction. 191-203. 10.1007/978-3-642-12275-0\_19.
- [29] Dornel, Benjamin. (2021, February 3), Predicting Online News Popularity (Part 1), *Towards Data Science*, <https://towardsdatascience.com/predicting-online-news-popularity-part-1-aae9a4f7f1a4>,  
acedido em: 18 de junho de 2021
- [30] Intellipaat Online. (2020, May 28), What is LSTM - Introduction to Long Short Term Memory, <https://intellipaat.com/blog/what-is-lstm/>,  
acedido em: 16 de junho de 2021
- [31] Bushaev, Vitaly. (2018, September 8), Understanding RMS-prop — faster neural network learning, *Towards Data Science*,  
<https://towardsdatascience.com/understanding-rmsprop-faster-neural-network-learning-62e116fcf29a>,  
acedido em: 18 de junho de 2021
- [32] Mani, Kaushik. (2019, February 17), GRU's and LSTM's, *Towards Data Science*,  
<https://towardsdatascience.com/grus-and-lstm-s-741709a9b9b1>,  
acedido em: 18 de junho de 2021
- [33] Wood, Thomas (2019, May 17), Convolutional Neural Network, *DeepAI*,  
<https://deepai.org/machine-learning-glossary-and-terms/convolutional-neural-network>,  
acedido em: 18 de junho de 2021
- [34] IBM Cloud Education (2020, October 20), Convolutional Neural Networks,  
<https://www.ibm.com/cloud/learn/convolutional-neural-networks>,  
acedido em: 18 de junho de 2021
- [35] M. Anjaria and R. M. R. Guddeti, "Influence factor based opinion mining of Twitter data using supervised learning," 2014 Sixth International Conference on Communication Systems and Networks (COMSNETS), 2014, pp. 1-8, doi: 10.1109/COMSNETS.2014.6734907.
- [36] Anjaria M, Guddeti RMR (2014) A novel sentiment analysis of social networks using supervised learning. *Soc Netw Anal Min* 4(1):1-15
- [37] Choudhury, Ambika. (2019, May 29), <https://analyticsindiamag.com/top-7-python-neural-network-libraries-for-developers/>

## Apêndice A

# *Corpus* de Palavras Controversas

---

corpus.txt
------------

---

transgênero  
transgênero  
trans  
joacine katar moreira  
joacine  
katar  
moreira  
joacine katar  
joacine moreira  
katar moreira  
racismo  
racista  
lgbti+  
lgbti  
lgbt  
lgbtq  
lgbtq+

---



## Apêndice B

# Comparação NetAc vs Bombedia

## Jornal Sol:

Tabela 1: Síntese dos resultados por ficheiro

Título	Nº de Comentários	Hate Speech(%) (Resultado NetAC)	Resultado Bombedia	Controverso
Está solteira? Multimilionário britânico procura mulher para construir família e passar férias de luxo	12	1,4433	Neutro	Não
Manuel Luís Goucha escreve carta aberta ao pai	8	2,1672	Neutro	Não
Marido de Goucha deixa provocação a Joacine Katar Moreira: 'Será que é xenofobia?'	24	1,5421	Neutro	Não
"Ele não está na TVI e não estará mais". Marido de Goucha "expulso" da estação de Queluz de Baixo	22	1,6746	Neutro	Não
Marido de Goucha indignado com alojamento de migrantes em Lisboa: "Portugueses sem-abrigo vagueiam ao frio"	40	0,5739	Positivo	Sim
Manuel Luís Goucha defende Judite Sousa: 'Tenho vergonha de seres humanos assim'	20	0,0	Negativo	Sim
Goucha comenta piada do marido: 'Nem eu resisto a um arroz de pato ou a uma canja de pombo'	10	0,0	Neutro	Não
Manuel Luís Goucha "brinca" com telefonema de Marcelo a Cristina Ferreira	3	0,0	Neutro	Não
Cláudio Ramos responde a marido de Goucha: 'Tenho 46 anos, uma filha para criar e uma carreira maior que a tua, Rui'	28	0,7989	Neutro	Não
Manuel Luís Goucha passa fim de semana na Holanda para assistir a provas de equitação	16	1,9608	Negativo	Sim

Telespetador decidiu dizer a Goucha que estava farto dele e o apresentador respondeu à letra	26	1,1881	Negativo	Sim
Marido de Goucha partilha caricatura para 'atacar' Cláudio Ramos: 'Que me desculpem os canídeos de que gosto tanto	4	3,8462	Neutro	Não
Manuel Luís Goucha e Maria Cerqueira Gomes apresentam "Você na TV" ... nus	12	0,0	Negativo	Sim
Goucha. 'Dizem que Bruno Caetano foi afastado do 'Você na TV'! Obrigado por me avisarem'	3	4,0	Neutro	Não
Já o deram como morto. Goucha 'responde' a Cristina Ferreira com convidado surpresa	14	0,3311	Neutro	Não
A próxima temporada do "Você na TV" pode não contar com Manuel Luís Goucha	6	2,027	Negativo	Sim
"Nunca digas adeus", diz Goucha a Cristina Ferreira	10	0,0	Neutro	Não
'O público continua com as pessoas em quem acredita', diz Goucha	2	0,0	Neutro	Não
'Para mim, o Deus da televisão é o Manuel Luís Goucha'	6	0,0	Negativo	Sim
Férias de Cristina Ferreira quase deram vitória a Manuel Luís Goucha	8	0,0	Neutro	Não
Manuel Luís Goucha conta como assumiu homossexualidade perante a mãe	16	0,6006	Neutro	Não
Cristina Ferreira e Manuel Luís Goucha reencontram-se em almoço: 'Não se falou de televisão'	7	1,0204	Neutro	Não
Maria Cerqueira Gomes assumiu programa sem Goucha e aproximou-se de Cristina Ferreira	21	1,0526	Neutro	Não
'Este ano envelheci 10 anos'. Goucha não esquece saída de Cristina	5	0,0	Neutro	Não

Goucha revela que fã se mudou dos EUA para o Alentejo por sua causa	6	1,8182	Negativo	Sim
'Eu não quero ditadores no meu país', diz Manuel Luís Goucha	52	0,4237	Negativo	Sim
Cristina Ferreira derrota Manuel Luís Goucha na estreia do seu programa na SIC	28	0,2033	Neutro	Não
Manuel Luís Goucha confirma 'traição' de Cristina Ferreira em direto	6	0,8696	Neutro	Não
Manuel Luís Goucha não acredita nas Aparições de Fátima	8	0,7692	Negativo	Sim
Goucha confessa: 'quis suicidar-me há muitos anos'	6	0,0	Neutro	Não
Você na TV. 'Nunca mais vai ser como era'? Goucha responde a seguidora — FOTO	4	0,0	Negativo	Sim
Manuel Luís Goucha e Cristina Ferreira iniciam programas de forma semelhante	5	0,0	Neutro	Não
Goucha recorre ao Facebook para criticar revista	5	0,0	Negativo	Sim
Manuel Luís Goucha não foi discriminado pelos tribunais	9	2,2124	Negativo	Sim
'A Cristina Ferreira ganha tanto ou mais que eu', diz Goucha	5	0,0	Neutro	Não
Manuel Luís Goucha 'ataca' novo ministro no Facebook	31	0,3633	Negativo	Sim
Goucha defende concorrente do Masterchef Júnior no Facebook	12	0,5076	Neutro	Não
Terá a amizade entre Teresa Guilherme e Manuel Luís Goucha chegado ao fim?	6	0,0	Neutro	Não
Cristina Ferreira quis "roubar" marido de Manuel Luís Goucha	24	1,129	Neutro	Não
Goucha reage a acusações de agressão no Masterchef Júnior	7	0,2747	Positivo	Sim

Goucha fala sobre novo programa. "Não vou andar lá no fornicançaço"	8	0,0	Neutro	Não
'É muito claro na minha vida que eu já não quero fazer programas diários'	26	1,0116	Negativo	Sim
Manuel Luís Goucha dá 1500 a jovem sobrevivente de cancro	10	1,4493	Neutro	Não
Daniel Oliveira recorda momento em que Mário Machado o ameaçou de morte — VÍDEO	38	0,2412	Positivo	Sim
Cristina Ferreira põe fim a dúvidas sobre a sua amizade com Manuel Luís Goucha	7	1,1111	Neutro	Não
'Odeio touradas e vivo há 20 anos com um homem que vai ver todas'	13	0,4673	Neutro	Não
Goucha. 'Bruno de Carvalho é um homem perturbado?'	6	0,0	Negativo	Sim
Confissões de Verão de Manuel Luís Goucha: 'Não sou dado à praia'	5	1,0526	Negativo	Sim
'É tudo a 7!'. A resposta de Goucha a Cristina Ferreira no Instagram	5	0,5102	Neutro	Não
Marinho Pinto chama 'sirigaita' a Cristina Ferreira	11	0,6993	Neutro	Não
Manuel Luís Goucha revela quem é o seu 'sucessor'	12	1,6471	Neutro	Não
Manuel Luís Goucha partilha imagem do quarto de hospital	5	1,4925	Negativo	Sim
Manuel Luís Goucha já reagiu à saída de Cristina Ferreira da TVI	22	0,8333	Neutro	Não
Lesão condiciona presença de Manuel Luís Goucha	4	0,7576	Negativo	Sim
Cristina Ferreira e Goucha atacados em direto	7	0,3802	Neutro	Não
'Não está contente, vá para a SIC'. As 'bocas' de Manuel Luís Goucha	13	0,8403	Neutro	Não

Suzana Garcia volta a comentar polémica: 'A minha avó é negra e a minha mãe é mulata'	24	3,4351	Positivo	Sim
'Ó senhores deixem-se de mariquices ridículas'	6	2,0134	Negativo	Sim
'Esta é uma fase em que eu e o Manel precisamos deste silêncio'	5	0,0	Negativo	Sim
André Ventura critica mensagem de solidariedade de Costa para Marega: 'Era a esta hipocrisia que me referia'	86	1,7503	Positivo	Sim
Marega e o racismo em Portugal	19	1,3767	Negativo	Sim
Atitude de Bento Rodrigues no Primeiro Jornal está a tornar-se viral nas redes sociais	40	1,2317	Neutro	Não
O Marega foi ao Dubai ser tratado pelo Fisioterapeuta através do medicamento infiltrado Meldonium	26	0,7082	Negativo	Sim
André Ventura dá a entender que não há racismo nos ataques a Marega	81	1,9399	Positivo	Sim
FCP. Marega abandona campo depois de insultos racistas	35	1,5184	Negativo	Sim
Marega reage nas redes sociais e ataca o árbitro	26	1,5816	Positivo	Sim
A fantochada do racismo	104	1,9879	Negativo	Sim
Moussa Marega. O jogador que fez história por dizer 'Basta!' ao racismo	24	3,4413	Neutro	Não
Catarina Martins: 'Adepta de Marega me confesso. Racismo não é opinião. É crime	33	2,8679	Positivo	Sim
PSP diz já ter identificado dez pessoas no caso Marega	71	1,3315	Neutro	Não
Marcelo condena insultos racistas a Marega	73	2,2989	Positivo	Sim
PSP já identificou adeptos que dirigiram insultos racistas a Marega	49	1,4146	Negativo	Sim
Comportamento dos jogadores do FC Porto 'foi nojento'	34	0,8705	Neutro	Não

João Mário sobre o caso Marega: 'Fala-se muito e não se faz nada'	21	2,2417	Neutro	Não
Caso Marega. Conselho de Disciplina abre processo ao Vitória de Guimarães	15	1,5385	Negativo	Sim
Ministro e responsáveis do futebol vão ao Parlamento falar sobre Caso Marega	17	1,8987	Neutro	Não
Hoje em dia falam muito de jogadores como Mbappé, Messi ou CR7 e esquecem-se de Marega	21	0,8547	Neutro	Não
'Mais do que racismo, foi uma prova de estupidez', diz Pinto da Costa sobre caso Marega	38	1,2903	Neutro	Não
Pepe e Marega alvo de processos disciplinares	14	2,8391	Neutro	Não
APCVD abre processo sobre caso Marega para 'averiguar responsabilidades'	10	0,5917	Negativo	Sim
FC Porto. Marega recupera de lesão na China	6	0,0	Negativo	Sim
Rúben Amorim defende Marega: 'Está na hora de passar-se à ação e castigar'	13	1,4535	Neutro	Não
Vitória SC reage a caso Marega e diz que não irá 'vestir a pele do lobo' por um problema 'de dimensão nacional'	16	1,5942	Positivo	Sim
José Carlos Malato sobre a eutanásia: 'Espero que as pessoas possam decidir'	20	0,3356	Neutro	Não
No nono aniversário da legalização do casamento LGBT em Portugal, Malato deixa sugestão à Câmara de Lisboa	6	3,0	Negativo	Sim
Eutanásia. Malato foi ao Parlamento e partilhou a sua opinião	13	0,9709	Negativo	Sim
José Carlos Malato: 'Devia ter morrido no ano passado. Não gosto nada do presente'	18	1,6949	Neutro	Não
Apresentador José Carlos Malato operado de urgência	10	0,5376	Negativo	Sim

Redes sociais. Nuno Markl segue "conselho" de José Carlos Malato	4	0,0	Neutro	Não
José Carlos Malato deixa mensagem sobre homossexualidade nas redes sociais — Foto	31	0,8782	Neutro	Não
Liliana Campos pede desculpa a José Carlos Malato após insinuar que este queria chamar a atenção	8	0,4762	Negativo	Sim
Malato sofre problema cardíaco	8	0,8032	Positivo	Sim
União gay ainda não é permitida nos Casamentos de Santo António	90	1,1825	Neutro	Não
'A minha mãe foi proibida de privar comigo porque sou gay'	22	0,3187	Negativo	Sim
Cara Delevingne assume relação no mês do orgulho LGBT	3	0,0	Negativo	Sim
Vaticano reconhece comunidade LGBT pela primeira vez	8	1,3393	Neutro	Não
Bandeira LGBT hasteada na Câmara Municipal de Lisboa	23	0,4093	Negativo	Sim
Associações de defesa dos direitos LGBT francesas criticam palavras do papa sobre homossexualidade	18	0,3527	Neutro	Não
Estas são as melhores cidades LGBT do mundo	11	4,3243	Neutro	Não
Uma bandeira LGBT 'original' em Moscovo — FOTOS	6	1,3699	Negativo	Sim
Turquia: manifestação LGBT acaba em confrontos com a polícia	10	2,0362	Negativo	Sim
Marcha do Orgulho Gay nos Açores mobiliza pouco mais de dez pessoas	37	0,8574	Neutro	Não
Marcha LGBT. Centenas de pessoas desfilaram em Lisboa contra a discriminação	5	0,0	Negativo	Sim
Escola cristã expulsa aluna por usar camisola com arco-íris nas redes sociais	38	0,9371	Neutro	Não



Gay pride. Todo o orgulho em ser o que se é	18	1,7405	Neutro	Não
Por que razão Julianne Moore é um ícone gay?	6	2,381	Neutro	Não
Editor de revista LGBT do Bangladesh agredido até à morte	6	3,7975	Positivo	Sim
Associação LGBT pede que insultos homofóbicos feitos a Ronaldo sejam investigados	6	3,7975	Neutro	Não
Milhares na marcha LGBT no Porto	3	0,8721	Negativo	Sim
Seja a família como for, o importante é haver amor	19	0,3578	Neutro	Não
Activistas querem abrir a primeira escola 'gay' do Reino Unido	8	3,0822	Neutro	Não
Primeira série LGBT portuguesa começa a ser filmada quinta-feira	6	1,9737	Negativo	Sim
'Eu era gay antes de me curar'	25	2,0797	Neutro	Não
Casal homossexual agredido por grupo no Terreiro do Paço	25	1,9093	Negativo	Sim
Marcha do Orgulho Gay defende direitos de LGBT	13	0,0	Neutro	Não
Futura ministra de Bolsonaro defende que a mulher 'nasceu para ser mãe'	39	1,0668	Neutro	Não
Campolide já tem duas passadeiras com as cores do arco-íris	22	1,1905	Negativo	Sim
Brasil não pode ficar conhecido como paraíso do mundo gay	64	0,8211	Neutro	Não
Barreiro. Deputadas do Bloco apresentam queixa contra deputado do PSD	25	0,5708	Positivo	Sim
Lisboa vai ter passadeiras com cores da bandeira LGBTI contra a homofobia	64	1,0645	Negativo	Sim
Ricardo Araújo Pereira criticado por declarações feitas ao i [vídeo]	12	0,8696	Negativo	Sim
Malta introduziu o divórcio há seis anos. Agora prepara-se para o casamento gay	6	0,0	Neutro	Não

Adeptos do FC Porto detidos em Itália por agressão a agentes da autoridade	37	0,2994	Negativo	Sim
Greta Thunberg poderá estar a caminho de Portugal	50	0,3028	Negativo	Sim
Irmã mais nova de Greta Thunberg chama-se Beata e luta pelo feminismo e contra o bullying	15	2,4735	Negativo	Sim
Mau tempo impede Greta de discursar no Parlamento	71	0,2328	Positivo	Sim
Greta é louca e perigosa. Acho que ela tem de voltar à escola e calar-se	83	0,5747	Negativo	Sim
Erro de atriz leva Greta Thunberg a mudar de nome nas redes sociais	13	1,1696	Positivo	Sim
Depois de ataque de Trump, Michelle Obama deixa mensagem a Greta	63	1,4247	Positivo	Sim
Greta Thunberg envolvida em polémica com empresa de comboios	14	0,738	Negativo	Sim
Greta Thunberg está cada vez mais próxima de Portugal. Agora, em direção aos Açores	16	0,5848	Negativo	Sim
Fernando Rocha testa positivo à covid-19 pela sexta vez	21	0,566	Negativo	Sim
Após teste negativo, Fernando Rocha volta a testar positivo para covid-19	12	0,0	Neutro	Não
Fernando Rocha revela que está infetado com covid-19	30	0,2829	Neutro	Não
Após dois meses em casa, Fernando Rocha revela qual foi a primeira coisa que fez	4	1,0309	Neutro	Não
Já morreram mais pessoas infetadas com covid-19 no Brasil do que na China	57	0,107	Negativo	Sim
Bolsonaro visitou padaria e abraçou e tirou fotografias com funcionários — Vídeo	36	0,5932	Negativo	Sim
Bolsonaro ameaça ministro da Saúde por defender isolamento: Nenhum ministro é indemissível	23	0,4747	Neutro	Não

17 Sou Messias, mas não faço milagres. Bolsonaro sobre recorde de mortes no Brasil	76	0,467	Positivo	Sim
Brasil volta a registrar recorde de novos casos de infecção por covid-19	14	0,4747	Positivo	Sim
Bolsonaro protagoniza mais um momento insólito ao ser questionado sobre mortes: Não sou coveiro	40	0,2681	Positivo	Sim
O dia mais negro do Brasil	48	0,1741	Positivo	Sim
O Brasil é dirigido por um fantoche que é absolutamente ignorante, inimputável, incompetente e cruel	36	0,6923	Positivo	Sim
Bolsonaro Messias não faz milagres, nem os seus discípulos o seguem	13	0,7273	Negativo	Sim
Bolsonaro diz que não é uma gripezinha que o vai derrubar	60	0,2706	Negativo	Sim
Brasil. A longa descida ao inferno do Governo de Jair Bolsonaro	33	0,2662	Neutro	Não
Toda a gente morre um dia. Foi assim que Bolsonaro reagiu às 20 mil mortes por covid-19 no país	87	0,5457	Positivo	Sim
Fiéis de Caminha manifestam-se contra saída de padre motard e sex symbol	38	0,2817	Negativo	Sim
Padre e falsas freiras escravizavam raparigas em Braga	17	0,7937	Positivo	Sim
Padre de Pedrógão diz ser um maroto sem maldade	18	0,5272	Positivo	Sim
Foi um descuido afirma padre após publicar fotografia em cuecas nas redes sociais	7	0,0	Neutro	Não
Padre encontrado morto na praia de São Pedro de Moel	39	0,7194	Negativo	Sim
Vaticano expulsa padre que revelou homossexualidade	8	0,2551	Negativo	Sim

Convidada deixa Fátima Lopes estupefacta: A primeira vez que me prostituí foi com um padre	26	0,7547	Negativo	Sim
Igreja encobriu padre que abusou sexualmente dos filhos	6	0,565	Negativo	Sim
Crianças forçadas a puxar Porsche do padre — Vídeo	15	0,8351	Negativo	Sim
Padre expulsa maestro do coro por ser homossexual	21	0,5545	Negativo	Sim
Padre culpa gays pelos sismos de Itália	10	0,6135	Negativo	Sim
Padre acusado de burlar o Estado	15	0,3409	Positivo	Sim
Padre assume homossexualidade no final da missa	28	2,0408	Negativo	Sim
Igreja: padre denuncia encontros gay em bares e ginásios	5	0,4354	Neutro	Não
Padre gera polémica ao dizer que pedofilia não mata ninguém ao contrário do aborto	29	0,5231	Positivo	Sim
Detido no Algarve padre que abusou de mais de 20 crianças	13	2,2785	Negativo	Sim
Ex-padre casa com modelo 55 anos mais novo	6	3,3898	Neutro	Não
Padre de Pedrógão Grande fotografado em roupa interior muda de religião	14	0,3937	Neutro	Não
Cristina Ferreira mostra mais do que queria em fotografia	16	0,0	Neutro	Não
Cristina Ferreira lança livro com nome chocante. Saiba qual	56	0,4115	Neutro	Não
Tens uma relação amorosa com a Cristina Ferreira? Ruben Rua responde	13	2,6316	Neutro	Não
Costa 'despediu' presidente do TdC por telefone	128	0,6305	Neutro	Não
António Costa deseja as melhoras a Trump	14	2,0243	Positivo	Sim
Donald Trump e Melania testaram positivo para a covid-19	88	0,7235	Negativo	Sim

António Costa reage a entrevista de Ana Leal ao SOL: "É mentira"	99	0,4216	Positivo	Sim
António Costa admite adotar medidas ainda mais restritivas nas próximas semanas	47	0,3435	Neutro	Não
Bazuca de pólvora seca desespera UE	32	0,0677	Negativo	Sim
António Costa diz que manifestação no Porto foi legítima mas condena arremesso de garrafas	56	0,6382	Positivo	Sim
António Costa garante que Portugal não vai usar empréstimos europeus enquanto situação financeira do país não o permitir	59	0,8277	Neutro	Não
António Costa pede aos portugueses que comprem máscaras nacionais	20	0,3311	Neutro	Não
António Costa afirma que o país enfrenta gigantesca responsabilidade	27	0,2336	Neutro	Não
António Costa e Fernando Medina na comissão de honra de Luís Filipe Vieira	119	0,4036	Neutro	Não
Marcelo marca as presidenciais para dia 24 de janeiro	30	0,3929	Negativo	Sim
Costa elogia comportamento exemplar dos portugueses no fim de semana	44	0,0747	Positivo	Sim
Comunistas recusam adiar congresso	58	0,8661	Negativo	Sim
Cláudio Ramos revela: "Odiei trabalhar com a Joana Latino"	22	1,0309	Neutro	Não
Cláudio Ramos confessa a Cristina que lhe custou ficar sem o Big Brother	7	1,4493	Neutro	Não
Cláudio Ramos para Carolina Deslandes: Então burra e quem não lê és tu	15	0,6515	Negativo	Sim
Emocionado, Goucha revela em que momento se sentiu magoado com Cristina Ferreira	9	1,3699	Neutro	Não
Acusado de discriminação, Manuel Luís Goucha responde a seguidora	17	2,3891	Positivo	Sim

Goucha arrasa pessoas que não cumpriram normas de segurança na Nazaré: Imbecis e criminosos	33	1,6706	Negativo	Sim
Rita Blanco dá que falar após 'indiretas' a Cristina Ferreira: Põe lá no canal da outra	9	0,5051	Neutro	Não
Cristina Ferreira compra 2,5 Marta Temido considera que votar contra OE é desistir de melhorar os serviços públicos de saúde	10	0,0	Negativo	Sim

## Jornal Público:

Tabela 1: Síntese dos resultados por ficheiro

Título	Nº de Comentários	Hate Speech(%) (Resultado NetAC)	Resultado Bombedia	Controverso
Supremo Tribunal Federal do Brasil proíbe censura de BD com beijo gay	10	0,3067	Neutro	Não
O PiS tem tudo para ganhar na Polónia, mas a corrida está mais disputada do que parece	9	1,1682	Positivo	Sim
Como Tancos salva a direita de si própria	85	0,1044	Negativo	Sim
Casal sofre agressão homofóbica no Terreiro do Paço	53	0,981	Neutro	Não
Arábia Saudita abre-se ao turismo. Há vistos turísticos pela primeira vez	9	0,0	Neutro	Não
Eles puseram-se Na Pele Dela em nome da igualdade de género	9	1,105	Neutro	Não
Vamos meter medo no coração do homem branco e outras citações de Robert Mugabe	12	1,3129	Neutro	Não
Anti-gender, uma sombra que cobre a Europa	88	0,5397	Neutro	Não
A diversidade deve começar na escola? Sim. Caso contrário, vamos continuar a reproduzir estereótipos	15	0,3979	Neutro	Não
Identidade de género: Estão em causa crianças e jovens que se sentem alvo de chacota	8	0,2053	Neutro	Não
Médicos vão ter guia para atender utentes transgénero e intersexo	13	0,0	Neutro	Não
Parabéns insultuosos	16	0,1451	Negativo	Sim
Esta é a 20ª marcha LGBTI+ em Lisboa. Lei mudou, falta a prática social	42	1,735	Neutro	Não

Milhares levam arco-íris pelas ruas de Lisboa em marcha de orgulho LGBTI+	43	0,3701	Neutro	Não
Investigadoras desmontam conjunto de mentiras sobre ideologia de género	31	0,6524	Positivo	Sim
Numa escola em Taiwan, os rapazes vão poder optar pela saia como uniforme	6	1,3636	Neutro	Não
Casamentos de Santo António ainda Não incluem matrimónios gay	15	0,6547	Neutro	Não
Nem sexo, nem morte. Bruno Maia, o médico sem tabus a caminho da AR	5	1,6949	Neutro	Não
EUA proíbe embaixadas de hastear bandeiras LGBT	9	1,6667	Negativo	Sim
Polícias recebem formação para dar resposta a crimes de ódio contra pessoas LGBTI	7	1,4388	Positivo	Sim
De norte a sul do país, o orgulho LGBTI+ volta a sair à rua	3	0,0	Neutro	Não
Taiwan legaliza casamento entre pessoas do mesmo sexo, uma estreia na Ásia	6	0,0	Neutro	Não
Clima social em Portugal ainda é homofóbico e transfóbico, denuncia ILGA	6	0,0	Neutro	Não
Ser gay é genético ou deve-se a conjunturas externas ? Colégio retira publicação de iniciativa do secundário	52	0,458	Neutro	Não
Brunei pede tolerância com a decisão de punir sexo homossexual com apedrejamento até à morte	18	0,4695	Neutro	Não
Como o Brasil está a contracenar com Bolsonaro	7	1,7143	Negativo	Sim
Fado Bicha: Isto é fado , mesmo que fale do Namorico do André e do Chico	23	1,4019	Negativo	Sim
Uma resposta a Pacheco Pereira: o #MeToo é uma revolução demasiado necessária e já vem tarde	15	0,4921	Positivo	Sim



Lisboa recebe agora a missa Beyoncé a pensar na Igreja e nas mulheres negras	6	0,0	Neutro	Não
No Vaticano quanto mais homofóbico alguém é, mais hipóteses haverá de ser gay	22	1,0989	Neutro	Não
LGBTI nas escolas? Quem está no terreno dispensa discursos apaixonados	4	0,6116	Negativo	Sim
Associações LGBTI em escolas? Depende muito	86	0,7956	Negativo	Sim
Jean Wyllys: O que deu a vitória a Bolsonaro foi a homofobia	126	0,5071	Neutro	Não
PSP identificou dois homens que tentaram atirar ovos contra Jean Wyllys em Coimbra	26	0,1953	Neutro	Não
Primeiro projecto no novo Congresso do Brasil é de "ex-gay" que quer que Bíblia seja património cultural	11	0,5013	Neutro	Não
Homem, mulher ou x? Os passageiros podem escolher a opção nas companhias aéreas	3	2,8986	Negativo	Sim
Nunca haverá um tempo sem Deus ou religião	29	0,0	Neutro	Não
Transição social de género em ambiente escolar atenuar o sofrimento de crianças e jovens	12	0,0	Neutro	Não
Maioria da esquerda reduz poderes de Marcelo, dramatiza Cristas	4	2,4096	Negativo	Sim
A crise da Direita	5	0,2611	Positivo	Sim
O género foi à casa de banho	5	0,0	Neutro	Não
Actor Ângelo Rodrigues terá injectado testosterona e foi hospitalizado. Quais os riscos desta hormona?	29	0,0	Neutro	Não
Alerta para cidadãos confusos	23	0,2562	Neutro	Não
Victoria s Secret cancela o desfile anual dos anjos	9	0,0	Negativo	Sim
E se o seu filho namorasse uma pessoa do mesmo sexo? 42 O CDS é um partido transgénero	19	0,1854	Neutro	Não

Alunos transgênero Não serão mais de 200, adianta secretário de Estado	18	0,6593	Neutro	Não
Inventar uma casa de emergência para vítimas de violência doméstica LGBTI	5	1,4423	Neutro	Não
A culpa e a reparação	22	0,5305	Neutro	Não
Primeiro projecto no novo Congresso do Brasil é de 'ex-gay' que quer que Bíblia seja património cultural	11	0,5013	Neutro	Não
Cristina Ferreira a Presidente? Apresentadora Não descarta candidatar-se a Belém	45	0,4902	Negativo	Sim
Crédito Agrícola vendeu imóvel a mãe de gestor da equipa de Licínio Pina	20	0,2545	Neutro	Não
Cristina Ferreira troca TVI pela SIC e vai ocupar as manhãs	23	0,9098	Neutro	Não
É o fim de uma era. SIC ultrapassa TVI nas audiências mensais pela primeira vez em mais de 12 anos	45	0,3008	Neutro	Não
Joacine Katar Moreira recusa paternalismo dos deputados	25	0,6639	Positivo	Sim
Joacine Katar Moreira exige 'respeito' por parte dos jornalistas	22	0,5859	Neutro	Não
Joacine Katar Moreira: 'Fui eu que ganhei as eleições sozinha'	67	0,6008	Neutro	Não
'Irei manter todas as minhas funções, a mensagem irá ser compreendida', diz Joacine Katar Moreira	39	0,7386	Neutro	Não
Joacine Katar Moreira vs. Daniel Oliveira: polémica acesa nas redes sociais	72	0,5908	Negativo	Sim
Joacine Katar Moreira: 'Sem igualdade Não há liberdade nenhuma'	164	0,6828	Neutro	Não
Livre elege Joacine Katar Moreira, uma activista negra	44	0,9528	Neutro	Não

Joacine Katar Moreira: uma activista negra a caminho do Parlamento?	27	1,625	Neutro	Não
Confronto sobre de tom: direcção do Livre desmente deputada Joacine Moreira	18	0,0	Neutro	Não
Os gritos de 'mentira' de Joacine: 'Senti a vergonha alheia'	18	0,4323	Neutro	Não
Livre preocupado com a sua deputada. Joacine Moreira diz-se apanhada de surpresa	99	0,4457	Neutro	Não
Joacine admite fazer 'cedências necessárias'. Direcção do Livre diz que será preciso 'milagre'	35	0,7225	Negativo	Sim
Nova direcção do Livre sem Joacine eleita com 95 votos a favor e 15 brancos	17	0,8565	Neutro	Não
O direito de resposta (mais moderado) que Joacine entregou ao congresso	5	1,9481	Negativo	Sim
Congresso adia decisão sobre retirada da confiança política a Joacine	39	0,0	Negativo	Sim
Joacine: Elegeram uma mulher negra que gagueja e deu jeito para a subvenção	137	1,2164	Positivo	Sim
Renunciar ao mandato de deputada? Joacine diz que 'está fora de questão'	17	0,3513	Positivo	Sim
O que acontece a Joacine se o Livre aprovar retirada de confiança política?	21	0,2088	Positivo	Sim
Orçamento foi a gota de água que levou Livre a propor retirada de confiança a Joacine	88	0,0872	Neutro	Não
Peço desculpa: Joacine e o Livre discordam em quê?	59	0,8722	Neutro	Não
Escolta a Joacine na AR: GNR só pode intervir se estiver em causa a 'segurança física' de deputados	29	0,1907	Neutro	Não
Joacine perdeu a graça	25	0,4008	Neutro	Não
Assessor de Joacine queixa-se de interrupções permanentes, cerco e mercantilização da informação	31	0,189	Neutro	Não

Joacine garante que tensões Não são por divergências programáticas	29	0,578	Neutro	Não
O erro da escolha de Joacine pelo Livre	43	0,5729	Neutro	Não
Joacine diz que votou 'contra ela própria' e devolve responsabilidades da abstenção à direcção do Livre	51	0,5141	Neutro	Não
Polémica entre Joacine e Livre Não acaba aqui. Caso segue para conselho de jurisdição	54	0,9667	Neutro	Não
Joacine e direcção do Livre trocam acusações. Fundador Rui Tavares critica deputada	28	0,2364	Negativo	Sim
Queixa ou carta aberta: Mulheres Socialistas repudiam programas da SIC e TVI	15	0,2525	Negativo	Sim
Estão os concursos da SIC e da TVI a reproduzir estereótipos femininos?	21	0,6526	Neutro	Não
Rui Pinto assume ser o denunciante do Luanda Leaks	71	0,1076	Positivo	Sim
BE quer Isabel dos Santos impedida de vender participações compradas com dinheiro roubado	28	0,1938	Negativo	Sim
Miguel Relvas rejeita ter ligações a Isabel dos Santos. BE corrige acusação, mas mantém suspeita	10	0,2326	Negativo	Sim
Isabel dos Santos constituída arguida em Angola	48	0,3806	Positivo	Sim
Isabel dos Santos muda de estratégia e negocia devolução de dinheiro a Angola, notícia o Expresso	21	0,0	Negativo	Sim
Advogado de Isabel dos Santos continua inscrito na Ordem, apesar de ter anunciado suspensão	9	0,0	Neutro	Não
Advogados portugueses cobram a offshore de Isabel dos Santos decreto presidencial do pai	25	0,4334	Neutro	Não
Isabel dos Santos: como é que ela construiu um império	35	0,2259	Neutro	Não

Isabel dos Santos terá transferido 115 milhões da Sonangol para o Dubai	46	0,249	Neutro	Não
Isabel dos Santos admite candidatar-se à presidência de Angola	35	0,4092	Neutro	Não
Isabel dos Santos Presidente de Angola? Que em 2027 possa ser candidata é uma ideia	5	1,2121	Negativo	Sim
Lava que se farta! : Isabel dos Santos perde processo contra Ana Gomes	52	0,365	Negativo	Sim
Conan Osiris já ganhou o Festival da Canção?	19	0,3947	Negativo	Sim
Parem de pedir ao Conan para Não ir a Telavive	50	0,2051	Negativo	Sim
Conan Osiris vence Festival da Canção	15	0,0	Neutro	Não
Isabel dos Santos: a empresária, a princesa, o império e 'os pés de barro do pai'	19	0,2732	Neutro	Não
Isabel dos Santos diz que serviços de segurança angolanos lhe entraram nos computadores em Portugal	8	0,0	Positivo	Sim
Divididos, Joacine e Livre já estão de olhos postos no futuro	19	0,6908	Neutro	Não
Joacine Katar Moreira: Vamos continuar a trabalhar com a confiança de uns e sem a confiança de outros	93	0,9954	Negativo	Sim
Livre convocou Joacine para reunião, mas deputada diz que Não recebeu nada	27	0,7625	Neutro	Não
PS condena declarações xenófobas de André Ventura sobre Joacine Katar Moreira	18	0,6631	Neutro	Não
André Ventura propõe que Joacine seja devolvida ao seu país de origem . Livre acusa-o de racismo	240	0,9147	Negativo	Sim
Ventura levado ao colo	142	0,3693	Neutro	Não
Assessor de Joacine retira confiança política ao Livre	9	1,2121	Negativo	Sim

Livre aprova retirada de confiança política a Joacine por maioria	147	0,5434	Negativo	Sim
Joacine deixa de representar o Livre e passa a deputada Não-inscrita a partir de hoje	50	1,0436	Neutro	Não
As duas pestes de 2020: coronavírus e racismo	27	0,7448	Positivo	Sim
Partidos Não vão condenar racismo de Ventura no plenário para Não prolongar polémica basta-lhes as palavras de Ferro	31	0,6524	Negativo	Sim
Anti-racismo. Antifascismo. Anticomunismo	35	1,6488	Negativo	Sim
Novo director da PSP diz que há tanto racismo na polícia como há na sociedade portuguesa	40	0,9877	Neutro	Não
Centenas marcham em Lisboa contra o racismo e a violência policial	15	1,4286	Negativo	Sim
Líder das Mulheres Socialistas acusa deputado do Chega de racismo e sexismo	26	1,4472	Negativo	Sim
Portugal instado a enfrentar racismo contra os ciganos	9	1,8307	Neutro	Não
Até quando haverá racismo contra as mulheres negras em Portugal?	21	1,7488	Positivo	Sim
Mulher acusa polícia de agressão e racismo. PSP chamou bombeiros e disse que era uma queda	171	0,9481	Positivo	Sim
Presidente dos conselheiros das comunidades portuguesas demite-se por causa de André Ventura	60	0,2612	Neutro	Não
Secretária de Estado diz que ciganofobia está no dia-a-dia da sociedade portuguesa	6	0,3289	Neutro	Não
Contas de Isabel dos Santos em Portugal arrestadas a pedido de Angola	24	0,197	Neutro	Não

Youtube:

Tabela 1: Síntese dos resultados por ficheiro

Título	Nº de Comentários	Hate Speech(%) (Resultado NetAC)	Resultado Bombedia	Controverso
O Interrogatório a Fernanda Cândia Ex Namorada de José Sócrates - Especial CMTV - 22 Abril 2018	59	0,9921	Neutro	Não
A seguir aos ciganos, Brasileiros são o maior alvo de discriminação em Portugal	155	1,4893	Negativo	Sim
André Ventura visita Quinta da Fonte e ignora os ciganos...	319	1,8509	Negativo	Sim
FAMILIA CIGANA IMPEDIDA DE JANTAR NUM RESTAURANTE	139	1,667	Neutro	Não
Almeirim - Este concelho não é para ciganos	77	2,4209	Negativo	Sim
Racismo em Portugal	426	1,9299	Negativo	Sim
Os Pretos Todos Daqui Para Fora!	451	1,8257	Negativo	Sim
Tabu Brasil: Mudança de Sexo (Dublado) - Documentário National Geographic	277	1,3985	Neutro	Não
Mulher faz cirurgia para virar homem	310	1,4461	Positivo	Sim
Mesquita Nunes desmascarou as mentiras do Bloco e Mariana Mortágua entrou em desespero	108	1,2544	Positivo	Sim
Maria Capaz entrevista Fernanda Cândia	40	2,0765	Negativo	Sim
Fernanda Cândia TVI Jornal das 8 12-05-16	7	3,4653	Negativo	Sim
Mariana Mortagua perde as estribeiras quando Mesquita Nunes compara o Bloco a Marine Le Pen	382	0,5618	Positivo	Sim
ÁRABE PORCO E COVARDE BATENDO NA NOIVA	429	1,5164	Positivo	Sim

Crianças Transgênero DE SUA OPINIÃO	333	0,8992	Negativo	Sim
Você decidiu ser menina? - Transgênero na infância (OFICIAL)	10442	0,9384	Neutro	Não
SOU TRANS, MAS A CIRURGIA NÃO ME FEZ MAIS MULHER — PAPO KABELO COM KAROL PINHEIRO — Salon Line	1312	1,4085	Negativo	Sim
O casal transgênero em que o pai deu à luz um menino	211	3,2193	Neutro	Não
A saga de ter um filho transgênero	54	0,793	Negativo	Sim
Ex-gay que tirou o pênis explica por que foi fácil se tornar 'homem hétero'	5391	1,6146	Neutro	Não
Domingo Espetacular conta o drama de quem se arrependeu de mudar de sexo	4655	1,0523	Neutro	Não
Menino que mudará de gênero e nome faz planos: 'Quero ter marido e 3 filhas'	314	0,7344	Neutro	Não
Profissão Repórter - Transgêneros - 01 08 2018	114	1,385	Negativo	Sim
Miguel Neto no bairro da Jamaica, Portugal [Webnível 95]	357	0,7711	Neutro	Não
Em Portugal não há racismo, há racismozinho :: Inferno T4 Ep.5	1125	2,0671	Negativo	Sim
Racista fala que homem nenhum gosta de Negra.	1367	3,2168	Negativo	Sim
Mulher é presa em flagrante por racismo.	1572	3,9361	Positivo	Sim
BALANÇO GERAL - Racismo e prisão! PM negro é chamado de macaco	543	4,3715	Positivo	Sim
Mulher racista pratica atos de racismo contra os Africanos durante a manifestação	39	2,6114	Positivo	Sim
Racismo cigano no você na TV quintino Aires rassita cristina ferreira racista	158	1,9827	Positivo	Sim
Racismo de ciganos no Lumiar	69	1,8484	Negativo	Sim
Negra vai em protesto Neonazista e se encontra com membro da Klu Klux Klan [Legendado Português]	899	1,9975	Positivo	Sim



'Se for negro, não entra': Polícia italiana impede refugiados de embarcar em trem para Alemanha	574	1,2255	Positivo	Sim
Adolf Hitler fala sobre os Judeus e os Aliados.	1163	0,667	Negativo	Sim
5 Frases de Adolf Hitler	2511	0,5226	Positivo	Sim
Mulher é presa em flagrante após usar termos racistas contra gerente de supermercado	3861	4,1268	Positivo	Sim
O anti-Papa: Governo Bolsonaro quer espionar igrejas católicas	228	0,9688	Negativo	Sim
A apresentadora de rádio Katie Hopkins humilha protestantes anti-Trump	62	0,5324	Negativo	Sim
Preconceito contra Pobre!!! Absurdo dos absurdos!!!	147	0,7468	Positivo	Sim
Comentário RACISTA de Lula.	216	4,2659	Negativo	Sim
Ciganos portugueses no Brasil	32	1,6293	Negativo	Sim
bairro sao joao de deus...tarrafal	99	0,8828	Neutro	Não
Desacatos entre CIGANOS e PRETOS 2010 08 28	20	1,0753	Negativo	Sim
Dois ciganos são executados na porta de casa no bairro Soledade - BALANÇO GERAL	70	2,4062	Neutro	Não
CMTV mostra vídeo do tiroteio em Lisboa	48	2,1088	Negativo	Sim
Mais uma criança retirada a família de etnia cigana	16	2,8205	Negativo	Sim
Zézinho só estava a fazer uma ganza, levou na boca (Ameixoeira)	1016	1,6603	Negativo	Sim
José Berardo mudou estatutos da associação à revelia dos credores para defender interesses pessoais	97	0,5004	Neutro	Não
Mariana Mortágua mais uma vez humilhada por Leitão Amaro	51	0,5249	Negativo	Sim
Cláudio Ramos PASSA-SE DA CABEÇA COM COLEGAS EM DIRETO - Junho 2018	104	0,8543	Neutro	Não

Cláudio Ramos Arrasa famosos que defenderam Júlia Palha	67	0,6757	Negativo	Sim
Maria Leal hoje aqui só para ti- Você na TV	220	1,1324	Negativo	Sim
Maria Leal e o desafio de cultura geral - 5 Para a Meia Noite	300	0,6114	Negativo	Sim
DENTISTA DROGADA, BÊBADA, RACISTA E HOMOFÓBICA	12	3,3473	Negativo	Sim
Maria Leal acusada de roubo Responde e é Arrasada	59	1,9745	Positivo	Sim
Sérgio Henriques Responde a Acusações de Maria Leal (Ex Namorada) no Manhã CMTV 04.01.2017	70	2,2046	Negativo	Sim
MARIA LEAL — O AMOR É CEGO...SURDO E DESDENTADO	67	1,3975	Neutro	Não
Briga no autocarro em Portugal com um velho racista	65	3,9568	Positivo	Sim
Portuguese fights /police fights/ range Portuguese compilation	105	1,6958	Negativo	Reati vo
TOUR PELO MEU CORPO TRANS + antes e depois	1831	0,9323	Negativo	Sim
Mudanças de sexo que ficaram INCRÍVEIS	3167	1,9726	Neutro	Não
Racismo entre Portugal e Angola,a 3ª guerra mundial vai começar	658	1,7485	Positivo	Sim
SIC - Bairros Sociais e violência em Portugal	33	0,2805	Negativo	Sim
PORTUGAL FUNDADOR MUNDIAL DO RACISMO E COMERCIO DE ESCRAVO -ANGOLANOS DAM RESPOSTA A PORTUGAL	296	1,3373	Neutro	Não
'Somos negros. Portugal ainda não dá valor como gente'	256	1,0941	Negativo	Sim
O Racismo em Portugal existe? Como enfrentá lo?	51	0,9989	Negativo	Sim

Racismo ao vivo em plena televisão portuguesa SIC	14	2,0548	Neutro	Não
Ataque racista dentro de avião gera críticas à companhia aérea	140	3,7453	Positivo	Sim
Piruka: 'Não acho as gansas o lado negro! Mas nunca dei um risco de coca!'	89	0,5691	Positivo	Sim
TOY FUMA UM CHARRO NA TVI	217	0,511	Negativo	Sim
travesti e abusado e deixado na mão	5280	1,0535	Negativo	Sim
Sensivelmente Idiota - Concorda com o acolhimento de refugiados em Portugal?	379	0,584	Positivo	Sim
Imigrante brasileiro é contra a vinda de refugiados para Portugal	294	0,6618	Negativo	Sim
A feia verdade sobre os Refugiados em Portugal	171	0,5094	Negativo	Sim
Refugiados venezuelanos deixam Roraima e chegam a São Paulo	1971	0,4789	Neutro	Não
Refugiados em Portugal	5	3,6036	Negativo	Sim
Refugiadoganhammai que os Portugueses	7	0,3846	Neutro	Não
'refugiados' fogem de Portugal	18	0,4975	Negativo	Sim
PNR - CONTRA CHEGADA DE REFUGIADOS A PORTUGAL	152	0,5234	Negativo	Sim
Refugiados em Portugal	4	0,0	Negativo	Sim
REFUGIADO EM PORTUGAL	7	0,8333	Negativo	Sim
'Refugiados' Vs Portugueses (reformados, sem-abrigo, etc)	29	0,3727	Negativo	Sim
'refugiados' fogem, mas os Portugueses PAGAM regresso	13	0,6006	Negativo	Sim
refugiado viola Portuguesa	23	1,1096	Negativo	Sim
Refugiados reconstroem a vida em Portugal	22	0,0	Negativo	Sim
A questão dos refugiados na Europa	29	0,5476	Negativo	Sim
O drama dos refugiados sírios e africanos que chegam a Calais, França	6	0,0	Neutro	Não

Refugiados em Portugal - Muçulmanos, Informações, Árabes, Abandono e nova vida	245	0,442	Negativo	Sim
A crise de refugiados: Amnistia Portugal @RTP1	26	0,2352	Positivo	Sim
Portugal refugiados a chegar	7	0,7812	Neutro	Não
Em Portugal: 'refugiado' eritreu violou uma mulher sem-abrigo de 67 anos	38	0,6897	Negativo	Sim
Na Europa, Jean Wyllys está dizendo que é Refugiado do Brasil.	488	1,1894	Negativo	Sim
Estes são os 'refugiados' que a Europa recebe diariamente	193	0,4984	Negativo	Sim
Muçulmanos de segunda geração estão completamente integrados em Portugal	77	0,5199	Negativo	Sim
Pedidos de asilo aumentam em Portugal	54	0,927	Negativo	Sim
Portugueses em Angola vivem bem mas não gostam dos pretos no país deles	2928	0,9834	Neutro	Não
NA SOMBRA DO PECADO - As Testemunhas de Jeová no documentário da TV de Portugal	51	0,2237	Positivo	Sim
Os portugueses são racista? (xenóforo) (Brazucas em Portugal)	67	0,9893	Negativo	Sim
COMO AS MULHERES BRASILEIRAS SÃO VISTAS EM PORTUGAL?	493	1,0802	Negativo	Sim