Politécnico
Castelo Branco
Escola Superior
de Tecnologia

Qualidade do Ar em Zonas Urbanas: Recolha de Dados e o seu Reflexo no Comportamento Online

Antonino Carlos Gonçalves Candeias

20170292

Orientadores

Professor Doutor Fernando Reinaldo Silva Garcia Ribeiro

Professor Doutor Rogério País Dionísio

Dissertação apresentada à Escola Superior de Tecnologia do Instituto Politécnico de Castelo Branco para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Informática - Área de Especialização em Desenvolvimento de Software e Sistemas Interativos, realizada sob a orientação científica do Professor Adjunto, Doutor Fernando Reinaldo Silva Garcia Ribeiro e do Professor Coordenador, Doutor Rogério País Dionísio, do Instituto Politécnico de Castelo Branco.

Janeiro de 2026

Composição do júri

Presidente do júri

Doutor, José Carlos Meireles Monteiro Metrolho

Prof. Coordenador, Escola Superior de Tecnologia do Instituto Politécnico Castelo Branco

Vogais

Doutor, Vítor Manuel Jesus Filipe

Prof. Associado, Universidade de Trás-os-Montes e Alto Douro

Doutor, João Manuel Leitão Pires Caldeira

Prof. Adjunto, Escola Superior de Tecnologia do Instituto Politécnico Castelo Branco

Doutor, Fernando Reinaldo da Silva Garcia Ribeiro

Prof. Adjunto, Escola Superior de Tecnologia do Instituto Politécnico Castelo Branco

Agradecimentos

Com a finalização desta dissertação de mestrado, e consequentemente do curso, gostaria de expressar a minha sincera gratidão às pessoas que mais me ajudaram e contribuíram para o meu desenvolvimento, tanto a nível pessoal como académico.

Em primeiro lugar, quero agradecer aos meus orientadores, Professor Adjunto Doutor Fernando Reinaldo Silva Garcia Ribeiro e Professor Coordenador Doutor Rogério Pais Dionísio, pelo apoio contínuo, paciência e grande disponibilidade demonstrados ao longo deste processo.

Agradeço também a todos os professores que me lecionaram durante o curso e que contribuíram significativamente para a minha formação académica.

Quero, igualmente, agradecer aos meus familiares aos meus pais, à minha irmã e aos meus avós pelo apoio e incentivo, que foram essenciais ao longo de todo este percurso.

Por último, mas não menos importante, agradeço aos meus amigos, pela companhia, motivação e encorajamento ao longo de toda a minha vida académica.

Resumo

Esta dissertação tem como objetivo analisar a relação entre qualidade do ar e os sentimentos expressos nas redes sociais, com foco no episódio de poeiras do deserto do Saara que afetou Lisboa em março de 2022. Para isto, foram recolhidos dados ambientais da plataforma QualAR, e dados meteorológicos do Meteomanz, complementados com a recolha de publicações do Twitter (plataforma atualmente conhecida como X) geolocalizadas em Lisboa. Os textos das publicações foram traduzidos e analisados com o algoritmo Valence Aware Dictionary and sEntiment Reasoner (VADER), permitindo identificar padrões emocionais e avaliar possíveis correlações com a variação dos poluentes atmosféricos. Foi também realizada uma análise preditiva no estudo, usando algoritmos de *Machine Learning*, nomeadamente Decision Tree e Random Forest, de forma a explorar relações entre os diferentes parâmetros e apoiar a interpretação dos resultados.

Os resultados mostram que, apesar de o evento ter provocado níveis excecionalmente elevados de partículas inaláveis com diâmetro < 10 μm (PM10), não se verificou uma alteração significativa nos sentimentos expressos online, que permaneceram maioritariamente neutros ou ligeiramente positivos. Estes resultados sugerem que a perceção ambiental da população não se reflete de forma clara nas redes sociais, possivelmente devido a fatores culturais, sociais ou ao reduzido impacto do Twitter no contexto português.

Conclui-se que a integração entre dados ambientais e sociais é metodologicamente viável e relevante, mas requer mais investigação, com períodos mais longos, diferentes cidades e técnicas de análise mais avançadas. Este estudo contribuiu para a compreensão da relação entre poluição atmosférica e perceções públicas em Portugal, colmatando uma lacuna existente na literatura.

Palavras-chave

Qualidade do ar; Redes sociais; Análise de sentimentos; Twitter; Poeiras do Saara

Abstract

This dissertation aims to analyze the relationship between air quality and the sentiments expressed on social media, focusing on the Saharan dust episode that affected Lisbon in March 2022. To achieve this, environmental data were collected from the QualAR platform, and meteorological data were obtained from Meteomanz, complemented by the collection of geolocated Twitter *posts* (now known as X) in the Lisbon area. The content of these *posts* was translated and analyzed using the Valence Aware Dictionary and sEntiment Reasoner (VADER) algorithm, enabling the identification of emotional patterns and the assessment of possible correlations with fluctuations in atmospheric pollutants. A predictive analysis was also conducted using Machine Learning algorithms, specifically Decision Tree and Random Forest, to explore relationships between the various parameters and support the interpretation of the results.

The findings indicate that, although the event caused exceptionally high levels of inhalable particulate matter with a diameter < 10 μm (PM10), there was no significant change in the sentiments expressed online, which remained mostly neutral or slightly positive. These results suggest that the public's environmental perception is not clearly reflected on social media, possibly due to cultural or social factors, or the relatively limited influence of Twitter in the Portuguese context.

It is concluded that the integration of environmental and social data is both methodologically feasible and relevant, but further research is needed, involving longer time periods, different cities, and more advanced analytical techniques. This study contributes to the understanding of the relationship between air pollution and public perception in Portugal, addressing a gap in the existing literature.

Keywords

Air quality; social media; Sentiment analysis; Twitter; Sahara dust

Índice geral

Índice de figuras

Lista de tabelas

Lista de abreviaturas, siglas e acrónimos

- APA – Agência Portuguesa do Ambiente
- C6H6 – Benzeno
- CO – Monóxido de Carbono
- CSV – Comma-Separated Values
- DT – Decision Tree
- HTML – HyperText Markup Language
- IoT – Internet of Things
- LoRaWan – Long Range Wide Area Network
- MLK – Machine Learning
- MQTT – Message Queuing Telemetry Transport
- MSE – Mean Squared Error
- NO2 – Dióxido de Azoto
- O3 – Ozono troposférico
- PLN – Processamento de Linguagem Natural
- PM10 – Partículas inaláveis com diâmetro < 10 μm
- PM2.5 – Partículas inaláveis finas com diâmetro < 2.5 μm
- QI – Questões de Investigação
- R^2 – Coeficiente de Determinação
- RF – Random Forest
- SO2 – Dióxido de Enxofre
- TTN – The Things Network
- VADER – Valence Aware Dictionary and sEntiment Reasoner
- WoS – Web of Science

1 Introdução

Nas últimas décadas, a crescente preocupação com os impactos da poluição atmosférica na saúde pública e no bem-estar das populações urbanas tem impulsionado o desenvolvimento de novas abordagens de monitorização ambiental. Paralelamente, o aumento exponencial da utilização das redes sociais como meio de expressão pública oferece uma oportunidade única para a análise indireta de fenómenos ambientais, através da observação das emoções manifestadas online.

No entanto, a convergência entre qualidade do ar e análise de dados provenientes de redes sociais permanece uma área de investigação ainda pouco explorada. A maioria dos estudos existentes concentra-se em abordagens tradicionais de monitorização ambiental, recorrendo a sensores Internet of Things (IoT) e a dados de satélite, ou em análises isoladas que não integram estas duas dimensões. Essa limitação evidencia a necessidade de estudos que combinem dados ambientais e sociais, explorando novas metodologias capazes de oferecer uma perspetiva mais integrada sobre o fenómeno.

Neste contexto, a presente investigação procura explorar a correlação entre os níveis de qualidade do ar e os sentimentos expressos nas redes sociais, com foco em especial na análise de dados recolhidos durante o mês de março de 2022, um período marcado por um episódio atmosférico extremo associado à passagem de poeiras do deserto do Saara pela Península Ibérica. Através da integração de dados ambientais e meteorológicos com publicações no Twitter, é possível compreender tanto a variação dos níveis de poluição como os seus reflexos emocionais na população.

O estudo baseia-se na aplicação de técnicas de Processamento de Linguagem Natural (PLN) e análise de sentimentos, com o objetivo de identificar padrões emocionais dominantes e avaliar em que medida estes se relacionam com as variações na qualidade do ar. Para tal, foram recolhidos dados ambientais de fontes oficiais e cruzados com *posts* geolocalizados na cidade de Lisboa. O trabalho procura igualmente identificar limitações e desafios metodológicos associados à utilização de redes sociais como ferramenta de inferência ambiental, propondo abordagens que contribuam para uma análise mais representativa e fiável.

Adicionalmente, o estudo da vertente analítica baseada em algoritmos de *Machine Learning* (ML), que poderá permitir identificar padrões de comportamento, avaliar a importância relativa das variáveis ambientais e estimar possíveis relações entre os parâmetros atmosféricos e os sentimentos expressos nas redes sociais. A integração desta abordagem quantitativa pode contribuir para uma compreensão mais abrangente e rigorosa do fenómeno analisado.

Deste modo, a presente dissertação procura contribuir para o avanço do conhecimento sobre as relações entre perceção social e poluição atmosférica, um tema ainda pouco explorado no contexto português.

1.1 Objetivos

O presente trabalho tem como principal objetivo analisar a relação entre a qualidade do ar e os sentimentos expressos nas redes sociais, procurando compreender de que forma

fenómenos ambientais, como a passagem de poeiras do deserto do Saara, podem influenciar o comportamento e a perceção pública.

De forma mais detalhada, este estudo visa:

- Examinar os principais parâmetros de qualidade do ar ao longo de um período temporal específico, identificando padrões e variações que possam refletir alterações nas condições atmosféricas;
- Investigar as manifestações emocionais e comportamentais expressas pelos utilizadores nas redes sociais, associando-as a potenciais flutuações nos níveis de poluentes atmosféricos;
- Avaliar a existência de correlações entre os indicadores ambientais e o sentimento coletivo digital, com vista à identificação de possíveis relações de causa e efeito;
- Explorar a viabilidade de modelos preditivos que permitam antecipar variações no sentimento público com base em condições ambientais observadas;
- Contribuir para uma compreensão integrada entre fatores ambientais e sociais, reforçando a importância da comunicação e sensibilização relativamente aos impactos da poluição atmosférica na vida quotidiana e no bem-estar coletivo.

1.2 Metodologia e Planeamento

O desenvolvimento deste trabalho seguirá uma sequência de etapas com o objetivo de analisar a possível relação entre a qualidade do ar e os sentimentos expressos nas redes sociais. Numa fase inicial, será realizada uma pesquisa e revisão da literatura existente sobre qualidade do ar, análise de sentimentos e perceção ambiental, de forma a compreender o enquadramento científico e identificar abordagens metodológicas relevantes para o estudo.

De seguida, proceder-se-á à recolha de dados provenientes de fontes oficiais e plataformas digitais. Espera-se obter dados ambientais relacionados com a concentração de poluentes atmosféricos, bem como dados meteorológicos de estações próximas à área de estudo. Paralelamente, serão recolhidas publicações em redes sociais geolocalizadas em Lisboa.

Após a recolha, os dados serão tratados e integrados de forma a garantir a sua consistência e permitir análises conjuntas. Serão aplicadas técnicas de limpeza, normalização e cruzamento dos diferentes conjuntos de dados, assegurando a correspondência temporal entre variáveis.

Está também prevista a aplicação de métodos de análise exploratória e de visualização gráfica para identificar padrões nos dados. Adicionalmente, será realizada uma análise de sentimentos aos conteúdos recolhidos nas redes sociais, recorrendo a ferramentas adequadas para esse fim.

Por fim, considera-se a possibilidade de aplicar técnicas de análise preditiva, com recurso a algoritmos de aprendizagem automática, com o objetivo de explorar possíveis relações entre os dados ambientais e os sentimentos expressos online. Os resultados obtidos serão analisados e representados graficamente, de forma a apoiar a formulação de conclusões.

Assim, o desenvolvimento desta dissertação foi organizado em sete fases distintas, distribuídas, alinhadas com a metodologia, ao longo de 12 meses. Cada fase corresponde a um conjunto de atividades fundamentais para a realização do trabalho, desde a conceção inicial até à redação final do documento.

Este cronograma permitiu garantir uma gestão eficiente do tempo e o cumprimento dos objetivos definidos para cada etapa do projeto, assegurando a progressão sequencial e lógica das fases do trabalho, **Figura 1**.

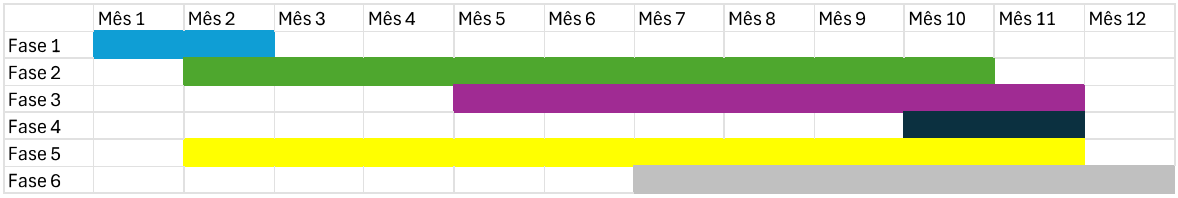

Figura 1 - Cronograma da dissertação.

Fase 1: Revisão do Estado da Arte: Sobre a monitorização da qualidade do ar e análise do comportamento em redes sociais
Fase 2: Implementação Técnica: Recolha e tratamento de dados, e testes;
Fase 3: Análise Computacional: Aplicação de algoritmos para estudar correlações entre poluição e emoções;
Fase 4: Avaliação e Discussão dos Resultados;
Fase 5: Redação da Dissertação Final;
Fase 6: Disseminação Científica: Preparação de artigos, submissão a conferências e revistas.

1.3 Estrutura do Documento

A estrutura adotada para a presente dissertação segue uma organização clássica, orientada para a clareza na apresentação dos dados, da metodologia e das conclusões obtidas. Está dividida nos seguintes capítulos:

Capítulo 1: Apresenta o enquadramento geral do tema, a motivação que conduziu à realização do estudo, os objetivos da investigação e a estrutura do documento.

Capítulo 2: Reúne e discute estudos anteriores que exploram a relação entre a qualidade do ar, as emoções expressas nas redes sociais e a análise de sentimentos.

Capítulo 3: Descreve a abordagem seguida para a recolha, tratamento e integração dos dados, bem como as ferramentas e tecnologias utilizadas.

Capítulo 4: Apresenta as análises realizadas e os resultados obtidos, com destaque para a visualização dos dados, interpretação dos comportamentos observados e avaliação das correlações entre variáveis ambientais e sentimentos.

Capítulo 5: Complementa o trabalho com uma análise complementar com algoritmos de ML.

Capítulo 6: Integra e interpreta criticamente os resultados comparando com os estudos relacionados, destacando semelhanças, divergências e possíveis explicações para os fenómenos observados. São também abordadas limitações e perspetivas de melhoria.

Capítulo 7: Resume as principais conclusões do estudo, os contributos alcançados e apresenta sugestões para investigações futuras.

2 Análise de Trabalhos Relacionados

A relação entre redes sociais e qualidade do ar tem sido explorada sob diversas perspetivas na literatura científica, incluindo a análise de sentimentos, previsão de poluição e perceção pública sobre os impactos ambientais e de saúde. Neste capítulo, são discutidos os trabalhos mais relevantes dentro dessas categorias, destacando as suas interseções e contribuições.

2.1 Metodologia

Para identificar artigos relevantes, seguiu-se uma adaptação das diretrizes PRISMA [1]. As subseções a seguir detalham o objetivo da revisão de literatura, os critérios de inclusão e exclusão, o procedimento de busca e seleção e a extração de dados.

O objetivo desta análise é contextualizar o presente trabalho dentro do estado da arte, identificando as abordagens mais utilizadas e os resultados obtidos em estudos anteriores. Pretende-se compreender como as redes sociais têm sido utilizadas para monitorizar ou prever a qualidade do ar, de que forma os métodos de análise de sentimentos contribuem para a perceção pública dos fenómenos ambientais e quais as principais limitações apontadas pela literatura. Esta contextualização permite não só justificar a relevância da investigação realizada, mas também evidenciar as áreas nesses estudos que não foram tão exploradas.

Estes estudos e o trabalho feito têm como objetivo explorar a relação entre qualidade do ar e emoções expressas nas redes sociais, definida pelas seguintes Questões de Investigação (QI):

QI1: Como as variações da qualidade do ar se correlacionam com as emoções expressas nas redes sociais.

QI2: Quais as principais emoções expressas em função da qualidade do ar (boa, moderada, má)?

QI3: Quais são as limitações da análise de emoções em redes sociais como indicador indireto da qualidade do ar?

2.2 Critérios de Inclusão e Exclusão

Foram definidos os seguintes critérios de inclusão:

- Estudos que abordem a influência da qualidade do ar nos sentimentos expressos pelas pessoas nas redes sociais;
- Estudos publicados nos últimos 5 anos;
- Fontes de literatura relevantes, como teses, relatórios de projetos e estudos técnicos disponíveis em repositórios académicos.

Foram excluídos os trabalhos que:

- Estudos que não estão disponíveis em inglês ou português.

2.3 Fontes de Informação

A busca por artigos foi conduzida nas bases de dados Scopus e Web of Science (WoS).

A *string* de pesquisa foi construída com base na junção de *querys* que tem interesse para este estudo, nomeadamente relacionadas com: qualidade do ar; influência ou impacto; saúde, sentimentos e bem-estar; e redes sociais.

("air quality" OR aqi OR polution) AND (perception OR impact OR influenc*) AND (health OR wellbeing OR happiness OR consumer* OR sentiment*) AND ("social networks" OR twitter OR facebook OR weibo OR "social media" OR "X Social Media " OR "X Social network "*)

A pesquisa inicial foi realizada com base nos títulos, resumos e palavras-chave dos artigos. Quando necessário, foram analisadas a estrutura, introdução, metodologia, resultados e conclusões para determinar a relevância do estudo.

2.4 Processo de Seleção

A **Figura 2** detalha o processo de seleção, o número de artigos obtidos e removidos em cada fase do processo. Na pesquisa inicial foram encontrados 65 artigos na WoS e 94 na Scopus. Analisados os títulos e resumos dos artigos, foram excluídos 47 artigos dos resultantes da WoS e 64 da Scopus. Juntando os resultados de ambas as pesquisas, foram removidos 12 registos duplicados. Resultando em 36 artigos para incluir na análise.

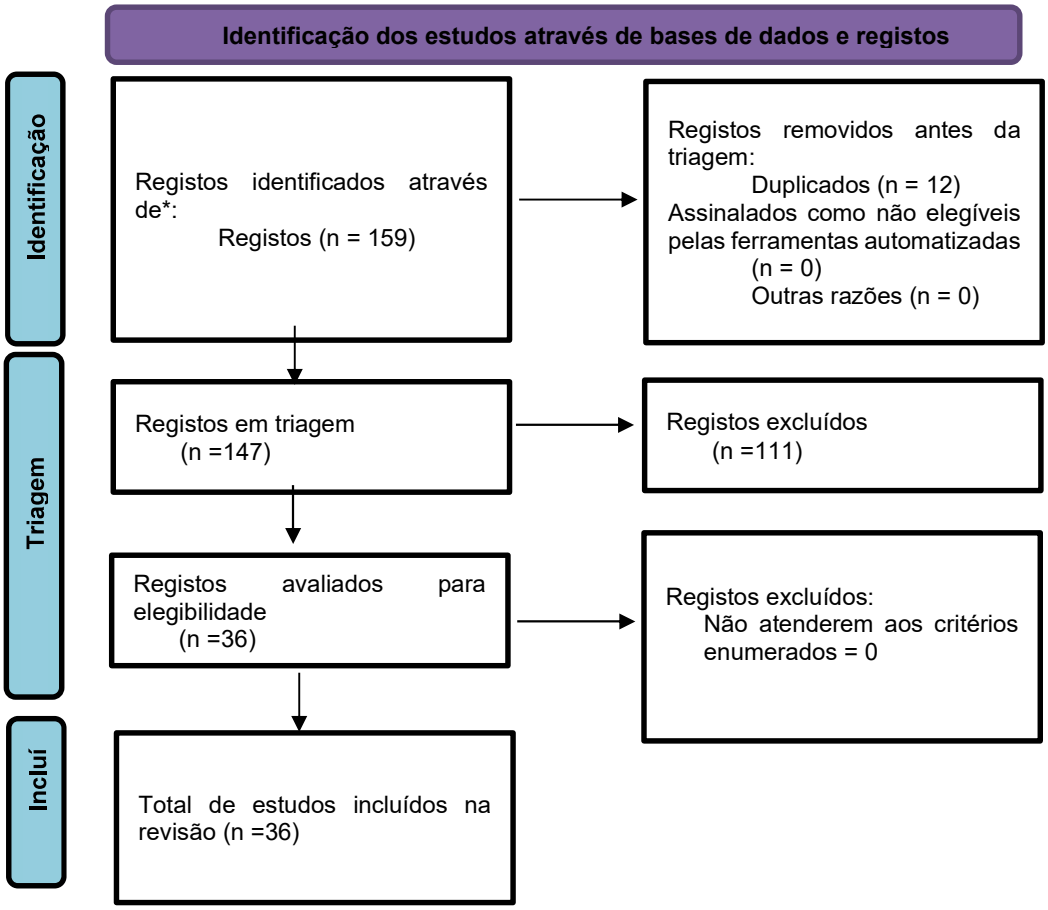

Figura 2 - Diagrama de seleção dos artigos.

A seguir, apresentam-se as análises quantitativas sobre a produção científica referente ao tema estudado. A **Figura 3** mostra a distribuição do número de artigos publicados por ano, evidenciando a evolução do interesse ao longo do tempo.

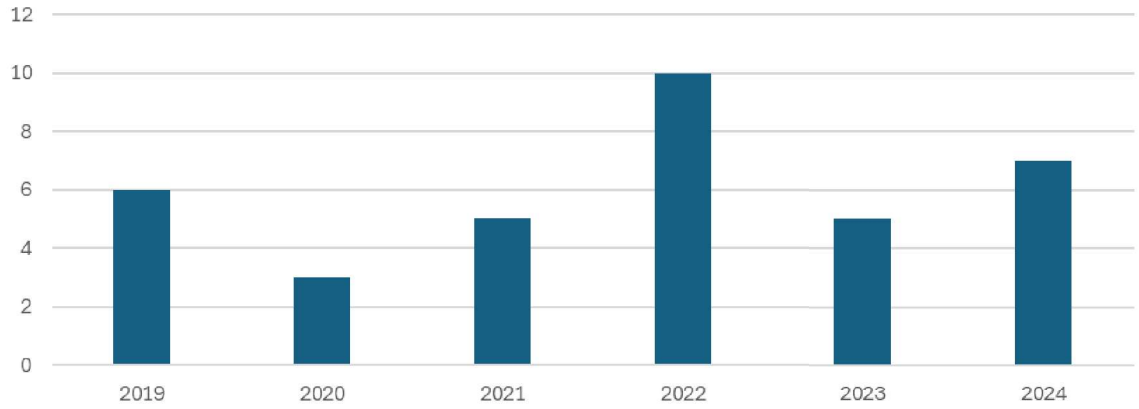

Figura 3 - Número de artigos publicados por ano.

A **Figura 4** detalha a distribuição dos artigos por país ou região, refletindo as áreas geográficas com maior produção sobre o assunto, com destaque para a China que é o país que tem substancialmente mais estudos sobre estas questões.

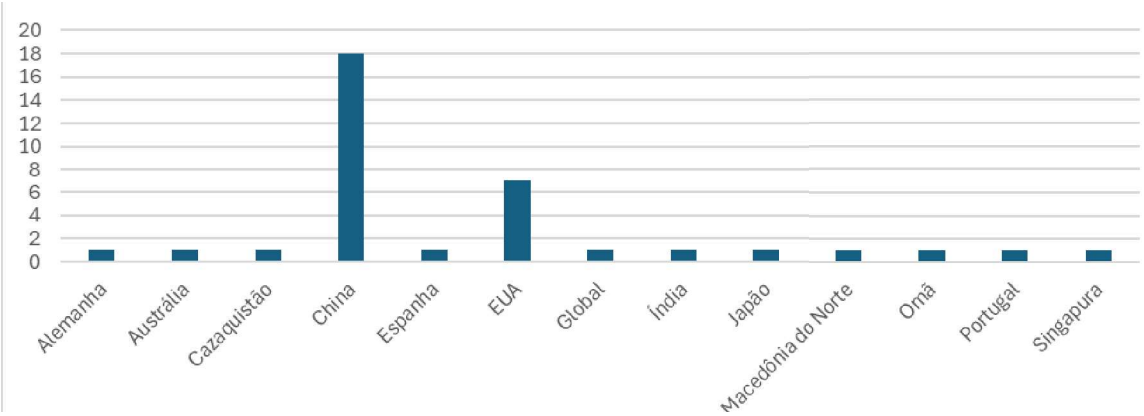

Figura 4 - Distribuição dos artigos por país ou região.

2.5 Extração de Dados

A **Tabela 1** resume os parâmetros extraídos de cada um dos estudos selecionados.

Tabela 1 - Resumo dos artigos analisados.

Artigo	Parâmetros de entrada	Algoritmo	Local	Correlação Qualidade do Ar-Emoções	Principal Limitação deste tipo de abordagem
[2]	AQI, PM2.5, Dados do Twitter	ROST EA (*Emotion Analysis) artificial neural network* (ANN)	China	✓	Estudo feito apenas em turistas.
[3]	PM2.5, Dados do Twitter,	LSTM (*Long Short-Term*	Pequim, China	✓	Estudo feito apenas num local

Artigo	Parâmetros de entrada	Algoritmo	Local	Correlação Qualidade do Ar-Emoções	Principal Limitação deste tipo de abordagem
	Weibo e Facebook	*Memory*), SVM (*Support Vector Machine*)			
[4]	PM2.5, Dados sobre incêndios, Twitter	*Linguistic Inquiry and Word Count* (LIWC)	Califórnia, EUA	✓	Estudo feito sobre um evento específico.
[5]	AQI, Weibo	Regressão, PLN	China	✓	Tendências online, podem ter efeitos nos dados.
[6]	PM2.5, Weibo	Louvain algorithm	China	✓	Sem validação com dados de sensores
[7]	AQI, redes sociais	LSTM (*deep learning*)	China	✓	Modelos complexos difíceis de interpretar
[8]	PM2.5, opiniões online, Weibo	Análise Fatorial	China	✓	Incongruências nos dados de redes socias
[9]	AQI, Weibo	Modelos preditivos	China	✓	Dados insuficientes para causalidade direta
[10]	AQI, Weibo	Análise de sentimentos	China	✓	Uso de um único indicador emocional
[11]	*Posts* sobre incêndios no Twitter	Análise temática	EUA	✓	Análise qualitativa pode ser subjetiva
[12]	AQI, pesquisas online	Modelos estatísticos	China	✓	Diferenças culturais podem afetar os resultados
[13]	Dados ambientais, Twitter, reddit, youtube	Pointwise Mutual Information (PMI)	Global	✓	Dados não representam toda a população
[14]	Twitter	Análise de conteúdo	EUA	✓	Falta de dados longitudinais para avaliar mudanças ao longo do tempo
[15]	Dados IoT, qualidade do ar	ML	Portugal	X	Foco na qualidade do ar, não nas emoções

Artigo	Parâmetros de entrada	Algoritmo	Local	Correlação Qualidade do Ar-Emoções	Principal Limitação deste tipo de abordagem
[16]	PM2.5, entrevistas, redes sociais	Análise Qualitativa	Austrália	Existe a correlação entre as entrevistas	Difícil medir impacto real com esta abordagem
[17]	AQI, pesquisas online	Modelos de redes sociais	China	✓	Fatores sociais podem influenciar mais que o ar
[18]	AQI, Weibo	Análise estatística	China	✓	Apenas dados urbanos analisados
[19]	AQI, dados migração	Análise de regressão Probit	China	✓	Outros fatores econômicos não considerados
[20]	AQI, redes sociais	Modelos de análise temporal3	Hong Kong	✓	Evento específico
[21]	Redes socias	Modelos preditivos	EUA	✓	Difícil prever efeitos a longo prazo
[22]	AQI, tráfego, Twitter	Modelos combinados	Flórida EUA	✓	Foco na interação entre fatores ambientais.
[23]	Weibo	Modelos de inferência	Província de Shandong, China	✓	Dados podem ser influenciados por mídia
[24]	Dados de redes sociais, dados internos	*Deep Learning*, PLN	Global	✓	Foco em saúde indoor.
[25]	Twiter	PLN	Japão e Itália	✓	Outros fatores externos podem influenciar o estudo
[26]	Redes socias, AQI	PLN	Balcãs	✓	Diferenças regionais não totalmente exploradas
[27]	AQI, redes sociais	Modelagem espacial,	Wuhan, China	✓	Dados podem não capturar áreas rurais
[28]	AQI, Twitter	Análise de sentimentos	Paris, Londres, Nova Délhi	✓	Representatividade dos dados limitada
[29]	AQI, dados dispositivos moveis	ML	Singapura	✓	Difícil diferenciar efeito da poluição de outros fatores.
[30]	Dados Twitter, AQI	PLN3	Minneapolis, EUA	✓	Apenas publicações públicas analisadas.

Artigo	Parâmetros de entrada	Algoritmo	Local	Correlação Qualidade do Ar-Emoções	Principal Limitação deste tipo de abordagem
[31]	AQI, *reviews* nas redes sociais	Análise Temática	China	✓	Pode haver erros na seleção de reviews
[32]	AQI, redes socias	Análise Fatorial	Muscat, Omã	✓	Diferenças demográficas podem influenciar respostas
[33]	Dados saúde, ambiente, AQI	Modelos estáticos	Hamburgo, Alemanha	✗	Relação complexa com outros fatores urbanos.
[34]	AQI, Dados de Grupos vulneráveis e preocupações ambientais	Modelos preditivo	China	✓	Fatores socioeconómicos podem ser mais influentes
[35]	AQI, Twitter	PLN	Delhi, Kolkata, Mumbai, Hyderabad, India	✓	Diferenças regionais podem afetar resultados.
[36]	Dados incêndios, AQI Twitter	Análise Sentimentos	Incêndios florestais na Califórnia, EUA	✓	Evento específico
[37]	Dados espaciais e PM2.5	Modelo BERT para extração de dados do Weibo; Regressão Geograficamente Ponderada (GWR)	Pequim, China	✗	Difícil captar impactos individuais.

2.6 Análise e Discussão de Resultados

Podemos então encontrar vários pontos em que estes documentos se intercetam tendo como base as nossas questões de investigação:

- Correlação entre a Qualidade do Ar e Emoções Expressas nas Redes Sociais

Vários estudos, como os de [1], [2], [3], [9], [27] e [34], demonstram que há uma forte correlação entre os níveis de poluição atmosférica e as emoções expressas pelos utilizadores nas redes sociais. A metodologia adotada por esses trabalhos envolve a extração de dados de plataformas como Twitter e Weibo (equivalente ao Twitter na China), seguida da comparação com medições oficiais da qualidade do ar, utilizando técnicas de PLN e análise de sentimentos. Os resultados indicam que picos de poluição tendem a aumentar a prevalência de sentimentos negativos, enquanto períodos de baixa poluição estão mais associados a sentimentos neutros ou positivos.

Em particular, o estudo [9] analisa publicações em redes sociais na China e identifica uma redução na felicidade expressa pelos utilizadores durante episódios de piora na qualidade do ar, sugerindo que a poluição influencia diretamente o humor coletivo da

população urbana. Da mesma forma, em [27] exploraram as reações emocionais ao ar poluído em diversas cidades, concluindo que regiões com pior qualidade do ar apresentam maior incidência de sentimentos negativos. Estudos mais recentes, como e, [34], ampliam esta análise, verificando que este impacto emocional é consistente em várias metrópoles a nível global.

- Principais Emoções Expressas em Relação à Qualidade do Ar

A classificação das emoções predominantes em função da qualidade do ar é um aspeto central em diversos estudos. Trabalhos como os de [4], [5], [10], [11], [19], [28] e [35] apontam que, em períodos de boa qualidade do ar, predominam emoções positivas ou neutras, enquanto episódios de elevada poluição estão fortemente associados a emoções negativas. As emoções mais comuns nesses momentos incluem medo, frustração e indignação, refletindo a preocupação da população com os impactos ambientais e na saúde pública.

Fatores culturais e regionais também desempenham um papel crucial na forma como essas emoções são expressas. [7], [17] e [31] analisaram essa questão, indicando que a perceção da poluição pode variar dependendo do contexto social, do nível de informação disponível e da consciência ambiental das populações estudadas. Enquanto algumas comunidades expressam mais indignação e exigem ações governamentais, outras demonstram resignação ou adaptação às condições ambientais adversas.

Outro fator relevante é abordado por [8], [15] e [20], que destacam que eventos de poluição extrema, como incêndios florestais ou períodos prolongados de nevoeiro tóxico, podem gerar um aumento na mobilização social e na pressão por políticas públicas mais eficazes. Isso demonstra que as emoções expressas online não se limitam ao ambiente digital, podendo influenciar debates políticos e iniciativas ambientais concretas.

- Limitações do Uso das Emoções Expressas para Medir a Qualidade do Ar

Apesar do potencial da análise emocional para inferir perceções ambientais, diversas limitações devem ser consideradas. Estudos como os de [16], [18], [22], [26] e [33] apontam que as emoções expressas nas redes sociais podem ser influenciadas por fatores externos, como eventos sociais, crises políticas ou até mesmo acontecimentos desportivos, o que pode comprometer a precisão na correlação entre emoções e qualidade do ar. Por exemplo, períodos de grande atividade política podem gerar um aumento de publicações negativas, independentemente das condições ambientais.

Outro desafio significativo está relacionado à representatividade dos dados. Os estudos [2],[14], [23] e [30] alertam que a amostragem obtida a partir das redes sociais pode não refletir a totalidade da população afetada pela poluição, já que a maior parte foca apenas em um determinado local onde grupos sociais podem estar sub-representados no ambiente digital. Além disso, há dificuldades em distinguir emoções diretamente relacionadas à qualidade do ar daquelas que surgem de interações sociais ou outras preocupações do quotidiano. Para mitigar esses desafios, alguns estudos propõem abordagens mais sofisticadas de análise de sentimentos. Os estudos [12], [21] e [29] sugerem a combinação de análise emocional com modelos contextuais mais avançados, que consideram variáveis como localização geográfica, condições meteorológicas e padrões de mobilidade urbana. Essas metodologias permitem uma inferência mais precisa, reduzindo os erros e aumentando a robustez das conclusões obtidas.

Embora esses estudos forneçam pontos de vista valiosos, a maior parte deles concentram-se em locais bastante distantes de Portugal com grandes diferenças culturais e socioeconómicas e normalmente em centros muito populosos, onde poderá ser difícil ter uma comparação coerente. O nosso estudo pretende então colmatar essa lacuna geográfica identificando localmente as convergências entre os dados e avaliar quais os efeitos na opinião pública.

3 Criação do Dataset

Este capítulo descreve o processo de recolha e preparação dos dados utilizados neste estudo. A análise centra-se num fenómeno atmosférico específico ocorrido em março de 2022, quando nuvens de poeiras provenientes do deserto do Saara atravessaram a Península Ibérica, afetando significativamente a qualidade do ar em Portugal. Este estudo foca-se em Lisboa, devido a maior facilidade da recolha de dados e possibilidade de obter um maior número de publicações no Twitter.

Dado o impacto ambiental e social deste evento, pretende-se estudar se essas condições atmosféricas extremas tiveram algum reflexo nos sentimentos expressos pelas pessoas, através de uma análise de sentimento aplicada a publicações do Twitter. Para isso, procedeu-se à recolha de dados ambientais e meteorológicos, bem como de *posts* publicados durante todo o mês de março de 2022. O objetivo é cruzar estas informações e avaliar eventuais correlações entre os níveis de poluição atmosférica e o estado emocional da população.

Na Secção 3.1, descreve-se a aquisição de dados ambientais provenientes da plataforma QualAR [38] da Agência Portuguesa do Ambiente (APA), com registos de parâmetros como partículas inaláveis (PM10), partículas inaláveis finas com diâmetro < 2.5 µm (PM2.5) e Dióxido de Azoto (NO2), recolhidos com granularidade horária. Para complementar estes dados, foram também recolhidas variáveis meteorológicas (como temperatura, humidade e vento) através da plataforma Meteomanz [39] com o objetivo de enriquecer o contexto de análise.

Na Secção 3.2, explica-se o processo de recolha de dados do Twitter, utilizando a ferramenta Power Automate da Microsoft,[40]. Através desta automação, foram recolhidos *posts* publicados ao longo de todo o mês de março de 2022, com base em critérios geográficos e temporais, com o intuito de posteriormente aplicar uma análise de sentimento ao conteúdo textual.

Na Secção 3.3, descreve-se o processo de junção dos dados ambientais e dos *posts* recolhidos, com o objetivo de preparar a base de dados para a análise no capítulo seguinte. Ambos os conjuntos de dados possuem registos com granularidade horária, o que permitiu alinhar cronologicamente os valores de poluentes atmosféricos com os *posts* publicados nas mesmas faixas horárias.

3.1 Aquisição de Dados e Parâmetros da Qualidade do Ar

Inicialmente, a recolha de dados ambientais foi planeada com uma abordagem prática e descentralizada, através da instalação de sensores físicos de qualidade do ar, com o objetivo de obter um período adicional de dados para análise. No entanto, esta estratégia foi posteriormente descartada, tendo-se optado por utilizar dados de um período específico março de 2022 para garantir uma maior consistência e foco na análise. Na abordagem descentralizada os sensores seriam conectados a uma rede Long Range Wide Area Network (LoRaWAN) [41], via servidor de rede The Things Network (TTN) [42], permitindo a transmissão de dados em tempo real. A informação recolhida era posteriormente encaminhada para um servidor de aplicação executando o *software* Node-RED [43] utilizando o protocolo Message Queuing Telemetry Transport (MQTT) [44], onde era armazenada numa base de dados para posterior análise. Esta arquitetura, apresentada na

Figura 5 foi concebida de modo a garantir autonomia do sistema, permitindo uma recolha contínua e de dados ambientais.

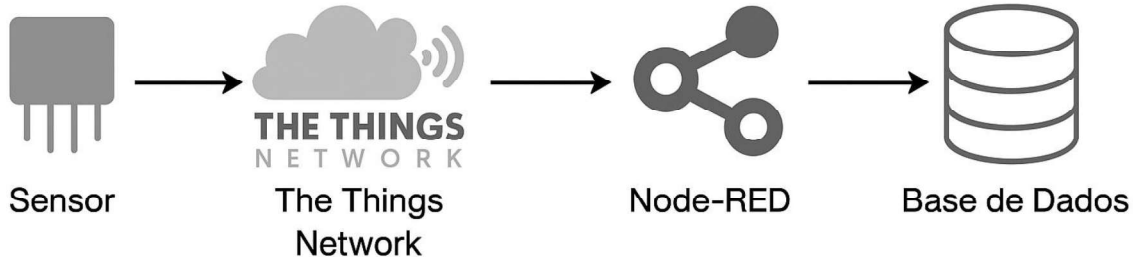

Figura 5 - Diagrama da arquitetura de recolha de dados com sensor + TTN + MQTT + Node-RED + Base de Dados.

Apesar das vantagens desta solução, nomeadamente a independência de fontes externas, e a possibilidade de expansão para diferentes localizações a sua implementação prática revelou limitações técnicas. Problemas de conectividade e instabilidade na integração com a rede TTN impediram a recolha fiável e contínua dos dados. Também se verificaram limitações significativas devido a problemas de calibração e fiabilidade dos sensores utilizados, comprometendo assim a consistência necessária para a análise proposta.

3.1.1 Recolha de Dados

Apesar de uma primeira abordagem mais descentralizada, para a construção do dataset principal, foi essencial incluir dados ambientais e meteorológicos com resolução horária, uma vez que estes servem como base de contexto para análises posteriores, como a associação entre condições ambientais e os sentimentos expressos nas redes sociais.

Os dados de qualidade do ar foram recolhidos a partir da plataforma oficial da APA [38]. Foi selecionado o relatório anual de 2022 da estação de Entrecampos, pela sua localização central em Lisboa, com registo de parâmetros como índice global, PM10, PM2.5, NO2, entre outros. Posteriormente, foi feito um filtro para manter apenas os dados referentes ao mês de março de 2022, mantendo a granularidade hora a hora.

Em complemento, os dados meteorológicos foram extraídos da plataforma Meteomanz [39], onde foi identificada uma estação próxima da zona de Entrecampos, assegurando coerência geográfica com os dados de qualidade do ar. A recolha abrangeu igualmente todo o mês de março de 2022, com variáveis como temperatura, humidade relativa, velocidade do vento entre outras, estas também contendo uma resolução horária semelhante.

3.1.2 Tratamento do Dataset

Após a recolha dos dados ambientais e meteorológicos, foi necessário realizar um processo de tratamento e preparação dos mesmos para garantir a sua consistência e integridade, de forma a permitir uma análise fiável.

Relativamente aos dados ambientais obtidos através da plataforma QualAR, os ficheiros originais incluíam registos de diferentes estações e de um período alargado (abrangendo todo o ano de 2022 e anos anteriores). Assim, recorreu-se ao Microsoft Excel para proceder à filtragem dos dados, selecionando unicamente os registos correspondentes ao mês de março de 2022 e à estação de medição situada em Entrecampos. Esta escolha foi motivada pela sua proximidade com a zona urbana de maior

densidade populacional e atividade. Após a filtragem, foram removidas colunas irrelevantes e garantido que todos os valores numéricos estavam no formato adequado, verificando a ausência de caracteres incorretos, células vazias ou erros de leitura. Este passo foi fundamental para assegurar que as médias e comparações futuras fossem baseadas em dados válidos.

No que diz respeito aos dados meteorológicos, estes foram recolhidos a partir da estação mais próxima da localização geográfica da estação de Entrecampos, com o objetivo de manter uma correspondência espacial coerente entre os dois conjuntos de dados. Foram selecionados apenas os registos referentes ao mês de março de 2022. Para além de ter sido feita a verificação de dados como anteriormente, filtragens e verificação de formatos, ainda se retificaram os dados meteorológicos que apresentavam um formato horário invertido, em que o dia começava às 24h e não às 00h, o que dificultava a correspondência direta com os dados ambientais. Para resolver esta inconsistência, foi necessário reestruturar os dados, reorganizando os horários e alinhando corretamente cada registo com o respetivo dia e hora. Adicionalmente, foi efetuada uma verificação de valores em falta, que poderiam comprometer a precisão da análise temporal. Nestes casos, procedeu-se ao ajustamento dos registos de forma a garantir que, para cada hora do mês, existisse uma correspondência direta entre dados ambientais e meteorológicos.

Este trabalho de tratamento e normalização permitiu assegurar que todos os elementos do *dataset* estivessem harmonizados em termos de estrutura temporal e prontos para integração na análise exploratória.

3.1.3 Estrutura e Descrição do Dataset

O conjunto de dados meteorológicos e de qualidade do ar contém várias colunas, **Tabela 2**, cada uma representando uma variável específica recolhida de forma horária durante o mês de março de 2022. A coluna Data corresponde à data e hora da medição, no formato YYYY-MM-DD HH:MM, permitindo associar os restantes valores a um instante temporal exato, neste caso cada dia tem 24 valores corresponde as 24 horas diárias.

A variável Dióxido de Enxofre (µg/m³) representa a concentração de SO2 no ar. A coluna Partículas < 10 µm (µg/m³) indica a concentração de partículas inaláveis com diâmetro inferior a 10 micrómetros (PM10), enquanto Partículas < 2.5 µm (µg/m³) corresponde às partículas finas (PM2.5), ainda mais pequenas e potencialmente mais prejudiciais para a saúde humana. A variável Ozono (µg/m³) representa a concentração de ozono troposférico (O3), e Dióxido de Azoto (µg/m³) refere-se à presença de dióxido de azoto (NO2), um poluente típico do tráfego urbano. A coluna Monóxido de Carbono (mg/m³) contém os valores de monóxido de carbono (CO), enquanto Benzeno (µg/m³) regista a concentração de Benzeno (C6H6), um composto orgânico volátil com efeitos tóxicos.

Relativamente às variáveis meteorológicas, Temp. (ºC) indica a temperatura do ar em graus Celsius, e H.Rel (%) corresponde à humidade relativa do ar expressa em percentagem. A Vel. vi. (km/h) refere-se à velocidade do vento em quilómetros por hora. Por fim, Prec. (mm) apresenta a precipitação acumulada na hora anterior, medida em milímetros.

Entre todas as variáveis, destaca-se especialmente a coluna de Partículas < 10 µm (PM10), que assumiu particular relevância neste estudo devido ao seu comportamento de grande variação durante o mês analisado. Essa variação esteve diretamente relacionada

com episódios de poeiras em suspensão provenientes do Saara, que afetaram de forma notória a qualidade do ar em Lisboa durante o período observado.

Por fim, importa sublinhar que a presença da informação horária será fundamental para análises futuras facilitando a integração direta com os dados provenientes das redes sociais e permitindo análises comparativas mais precisas.

Tabela 2 - Estrutura do dataset de parâmetros de qualidade do ar e meteorológicos.

Variável/Parâmetro	Descrição
Data	Data e hora da medição no formato YYYY-MM-DD HH:MM, com valores horários (24 por dia).
Dióxido de Enxofre (µg/m³)	Concentração de dióxido de enxofre (SO2) no ar.
Partículas < 10 µm (µg/m³)	Concentração de partículas inaláveis com diâmetro inferior a 10 micrómetros (PM10). Teve especial relevância devido à variabilidade causada por poeiras do Saara.
Partículas < 2.5 µm (µg/m³)	Concentração de partículas finas (PM2.5), mais pequenas e prejudiciais para a saúde.
Ozono (µg/m³)	Concentração de ozono troposférico (O3).
Dióxido de Azoto (µg/m³)	Presença de dióxido de azoto (NO2), poluente típico do tráfego urbano.
Monóxido de Carbono (mg/m³)	Concentração de monóxido de carbono (CO).
Benzeno (µg/m³)	Concentração de C6H6, composto orgânico volátil com efeitos tóxicos.
Temp. (ºC)	Temperatura do ar em graus Celsius.
H.Rel (%)	Humidade relativa do ar, expressa em percentagem.
Vel. v. (km/h)	Velocidade do vento em quilómetros por hora.
Prec. (mm)	Precipitação acumulada na hora anterior, em milímetros.

3.1.4 Apresentação Gráfica do Dataset

A classificação da qualidade do ar apresentada na **Tabela 3**, retirada da plataforma QualAr, permite contextualizar os valores registados ao longo do mês de março. Com base nestes intervalos, é possível compreender em que categoria (de "Muito Bom" a "Mau") se enquadram os níveis de concentração observados para cada poluente analisado.

Tabela 3 - Classificação dos poluentes.

Classificação	PM10	PM2.5	NO2	O3	SO2
Muito Bom	0-20	0-10	0-40	0-80	0-100
Bom	21-35	11-20	41-100	81-100	101-200
Médio	36-50	21-25	101-200	101-180	201-350
Fraco	51-100	26-50	201-400	181-240	351-500
Mau	101-1200	51-800	401-1000	241-600	501-1250

Para os restantes poluentes presentes no *dataset*, foi construída uma tabela, **Tabela 4** a partir de estudos encontrados em [45], para o CO e [46], para o C6H6.

Tabela 4 - Classificação dos restantes poluentes.

Classificação	CO	C6H6
Muito Bom	0 – 2	0 – 1
Bom	2 – 4	1 – 2

Médio	4 – 6	2 – 3
Fraco	6 – 10	3 – 5
Mau	> 10	> 5

Apresenta-se de seguida uma descrição dos dados recolhidos ao longo do mês de março de 2022, com foco em parâmetros relacionados com a qualidade do ar, variáveis meteorológicas e o indicador de análise de sentimentos. Estes dados, como descrito acima, foram organizados pela data pretendida permitindo uma leitura detalhada da variação destes fenómenos ao longo do tempo.

A **Figura 6** mostra os dados relativos à qualidade do ar, começando com as partículas inaláveis (PM10).

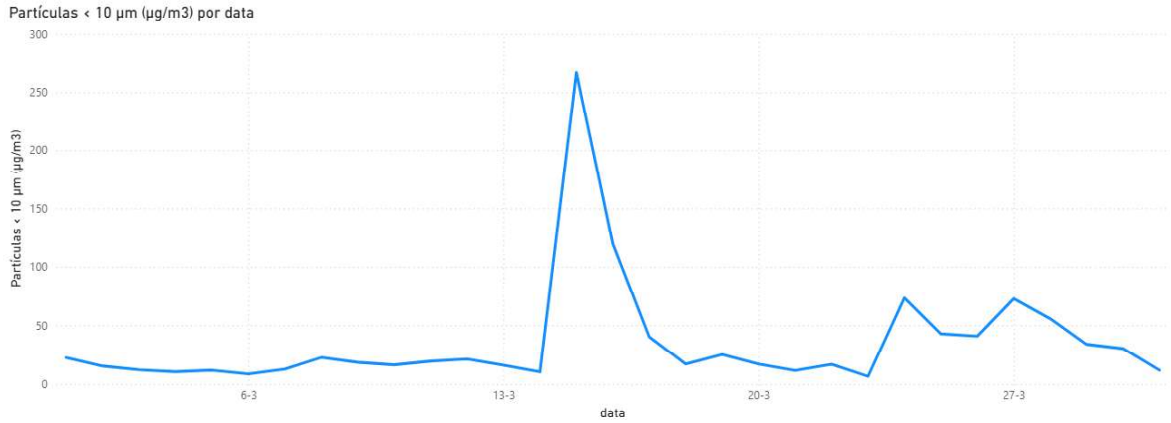

Figura 6 - Partículas PM10 x data.

Nos primeiros 14 dias do mês, os níveis de PM10 mantiveram-se relativamente baixos, geralmente abaixo de 30 µg/m³, o que se enquadra nas categorias de qualidade do ar "Muito Bom" (0–20 µg/m³) a "Bom" (21–35 µg/m³). No entanto, a partir de 15 de março, observa-se um aumento abrupto e significativo nos valores registados. Destaque para os dias 15 e 16 de março, com concentrações que ultrapassam largamente os níveis habituais, atingindo máximos de 407.08 µg/m³ (15/mar, depois das 18h) e 192.76 µg/m³ (16/mar, antes das 18h). Esses valores inserem-se na categoria "Mau" (101–1200 µg/m³), segundo os critérios classificação da qualidade do ar da referência utilizada, e aqui podemos claramente ver a influência do episódio de poeiras do Saara, que afetou diversas regiões do país, reduzindo a qualidade do ar e tornando o fenómeno visível a olho nu, com tons alaranjados no céu e poeira acumulada em superfícies.

O gráfico apresentado na **Figura 7** apresenta as três variáveis relacionadas à qualidade do ar: ozono, dióxido de azoto e partículas com diâmetro inferior a 2.5 µm. Estas três variáveis apresentam a mesma unidade de medida (µg/m³) e escalas semelhantes, o que permite uma comparação direta e facilita a visualização das suas dinâmicas conjuntas ao longo do tempo. A sobreposição das linhas azul-claro (ozono), azul-escuro (dióxido de azoto) e laranja (partículas < 2.5 µm) ajuda a identificar padrões e eventos que possam influenciar simultaneamente estes poluentes. Importa ainda referir que, em alguns dias, não houve registo de leituras de ozono, possivelmente devido a falhas ou limitações na fonte dos dados.

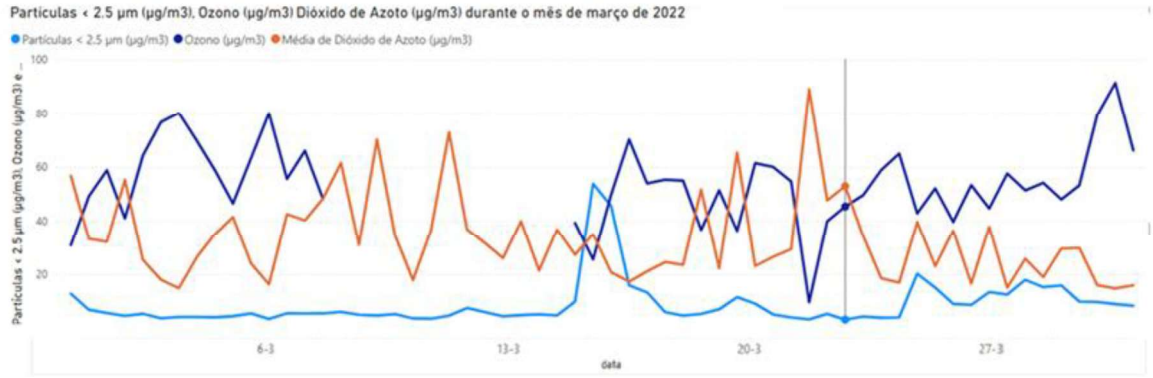

Figura 7 - Partículas PM2.5, O3, NO2 x data.

Nos primeiros dias de março, o ozono apresentou valores maioritariamente baixos, mas com picos a 3 e 4 de março que atingiram os limites da categoria "Bom". Ao longo do mês, registaram-se oscilações, com destaque para o final de março, onde os valores voltaram a subir, aproximando-se novamente do limiar superior da categoria "Bom" (por exemplo, 91.39 µg/m³ no dia 30 de março, antes das 18h). O dióxido de azoto, apesar da sua variabilidade, manteve-se quase sempre em níveis considerados "Muito Bom" ou "Bom", com apenas uma exceção mais notável: no dia 21 de março (depois das 18h), foi registado um pico de 89.05 µg/m³, a rondar o limite superior da categoria "Bom". Já as partículas com diâmetro inferior a 2.5 µm (PM2.5) mantiveram-se em níveis baixos durante a maior parte do mês, mas apresentaram picos classificados como "Fraco" em dias específicos nomeadamente a 15 de março (53.75 µg/m³, depois das 18h) e 16 de março (45.60 µg/m³, antes das 18h). Como já referido anteriormente, esses são os picos que coincidiram com a passagem de poeiras do Saara sobre o território português.

Para o dióxido de enxofre foi feito um gráfico individual, apresentado na **Figura 8** já que a escala não era compatível como nenhum dos outros parâmetros.

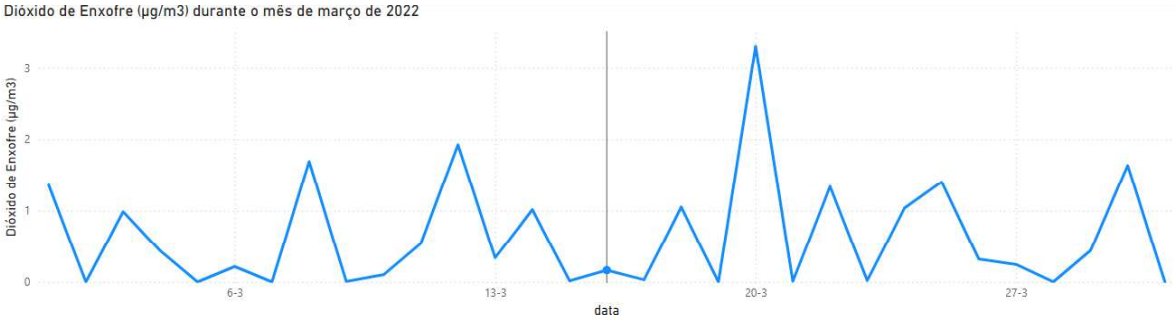

Figura 8 - SO2 x data.

Durante o mês de março, os valores de dióxido de enxofre (SO_2) mantiveram-se consistentemente baixos, oscilando na maioria dos períodos entre 0 e 3 µg/m³. Segundo os critérios nacionais de qualidade do ar, estes valores correspondem à categoria de qualidade “Muito Bom” (0–100 µg/m³). Estes resultados confirmam a baixa presença de SO2 no ar ao longo do mês, sem episódios significativos de poluição por este poluente.

No gráfico da **Figura 9** são apresentadas duas variáveis relacionadas com a qualidade do ar: CO e C6H6. Estas substâncias são medidas em unidades diferentes mg/m³ para o CO (linha azul-escura) e µg/m³ para o C6H6 (linha azul-clara).

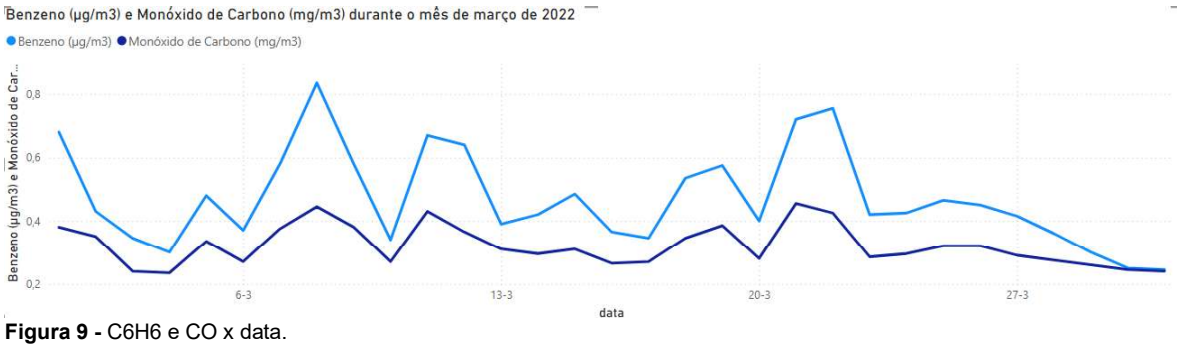

Figura 9 - C6H6 e CO x data.

Durante o mês de março, as concentrações de CO e C6H6 mantiveram-se estáveis e em níveis baixos, sugerindo um impacto reduzido destes poluentes na qualidade do ar local. As pequenas variações observadas entre os períodos, antes e depois das 18 horas, podem estar relacionadas com dias de maior tráfego e atividades urbanas, sem, contudo, comprometer a segurança ambiental.

Na **Figura 10** estão incluídas três variáveis meteorológicas relevantes para a análise da qualidade do ar: velocidade do vento (km/h), precipitação (mm) e temperatura (ºC). Estas variáveis partilham escalas compatíveis, o que permite uma visualização conjunta e facilita a comparação das suas flutuações ao longo do tempo.

Figura 10 - Temp, Prec, Vel.V x data.

Podemos então observar que a velocidade do vento não variou significativamente ao longo do mês, sendo que a sua influência na dispersão dos poluentes atmosféricos foi reduzida. A precipitação apresentou episódios intermitentes, estando ausente em vários dias. A temperatura manteve-se dentro de uma faixa relativamente estável, com pequenas variações entre os vários dias.

Para a humidade relativa foi feito um gráfico individual já que a escala não era compatível como nenhum dos outros parâmetros, **Figura 11**.

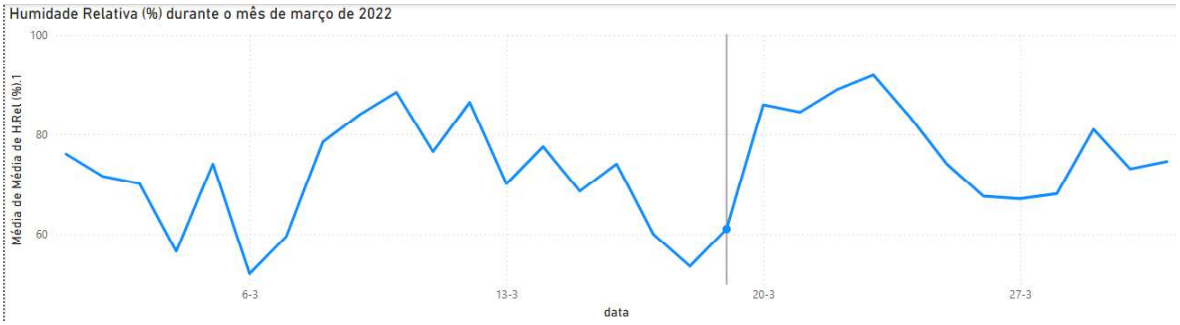

Figura 11 - Humidade Relativa x data.

Ao longo do mês, observa-se que a humidade relativa não se mantém constante, variando de forma intermitente entre valores mais baixos, próximos dos 50%, e valores mais elevados, que ultrapassam os 90% em alguns momentos. Estas variações mostram uma dinâmica diária da humidade, que se mantém dentro de um intervalo relativamente amplo.

3.2 Aquisição de Dados de Redes Sociais

A recolha de dados para a análise de sentimentos foi realizada na plataforma Twitter, com o apoio da ferramenta Power Automate, que permite a automação da extração de conteúdos da *web*. Esta abordagem foi fundamental para garantir uma recolha consistente e estruturada dos dados ao longo de todo o mês de março de 2022.

Os *posts* foram extraídos através da abertura manual de uma *query* de pesquisa no Twitter, especificamente formatada para recolher *posts* geolocalizados na área de Lisboa (raio de 200 km). Um exemplo da *query* de pesquisa utilizada foi:

geocode:38.7169,-9.1399,200km since:2022-03-31 until:2022-04-01 -filter:replies

Esta *query*, restringe os resultados ao intervalo entre as 00h00 e as 23h59 de cada dia, filtra respostas (*replies*) para manter apenas os *posts* principais, e aponta a uma área geográfica relevante para o estudo, neste caso foi usado o geocode de Lisboa num raio de 200 km.

3.2.1 Recolha de Dados

O Power Automate é uma ferramenta da Microsoft que permite automatizar fluxos de trabalho entre aplicações. No contexto deste projeto, foi utilizado para automatizar a extração de *posts* diretamente da interface do Twitter. O fluxo inicia-se com a abertura de um ficheiro Excel, onde os dados recolhidos serão armazenados. Em seguida, acede-se a uma página do Twitter previamente aberta com a *query* específica para o dia pretendido. Dentro de um *loop*, o Power Automate extrai os dados visíveis da página *web* (como utilizador, conteúdo do *post* e data), escreve esses dados numa linha do Excel, e move-se para a linha seguinte. Após cada iteração, o fluxo faz um *scroll* vertical equivalente ao tamanho do ecrã e aguarda cerca de três segundos antes de repetir, para evitar sobrecarga ou falhas na execução. A **Figura 12** apresenta o fluxo de Funcionamento no Power Automate.

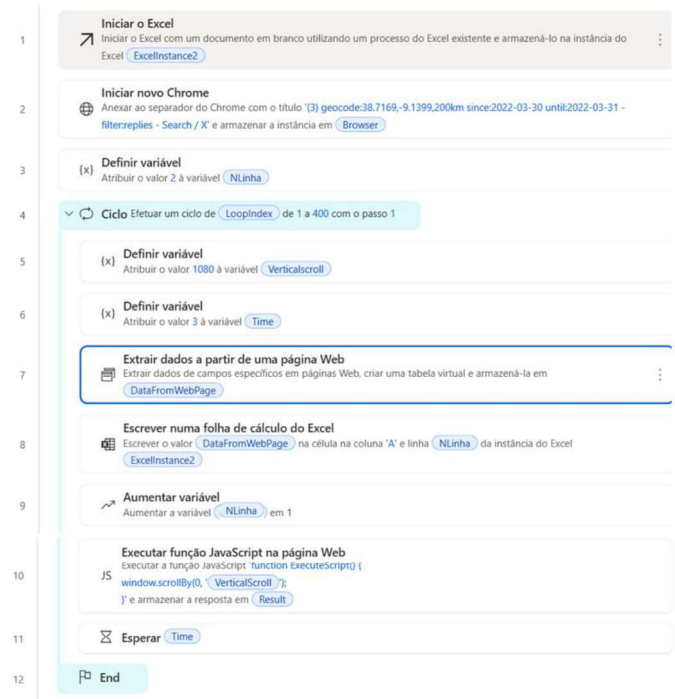

Figura 12 - Fluxo de funcionamento no power automate.

A Automatização da recolha dos campos relevantes (utilizador, conteúdo dos *posts* e data/hora da publicação), é feita na parte do fluxo, "extrair de dados a partir de uma página web", identificando os elementos HyperText Markup Language (HTML) da página. Primeiro o nome do utilizador é extraído marcando os elementos de dois nomes de utilizador seguidos, assim o programa reconhece automaticamente e armazena naquele valor todos os nomes de utilizador, o mesmo processo é feito para os *posts* e para data com a diferença em que esta é marcado o elemento <*time*>, **Figura 13**.

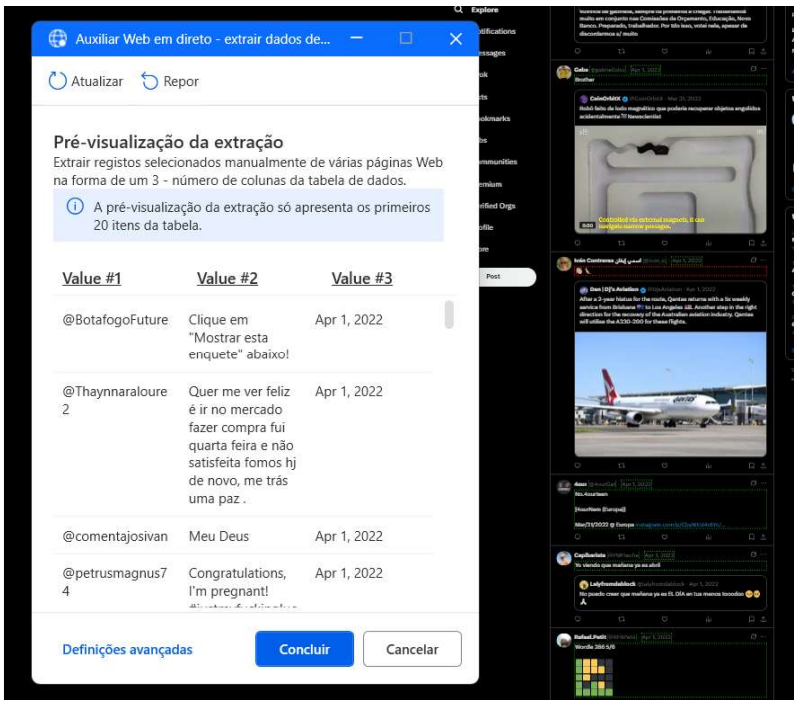

Figura 13 - Automatização da recolha de dados no power automate.

No entanto, esta ferramenta apresentou algumas limitações: apesar de recolher eficazmente os dados, incluía frequentemente *posts* duplicados e, em alguns casos, linhas

em branco. Estas últimas surgiam exclusivamente quando o conteúdo do *post* era apenas uma imagem, não havendo texto associado. O tratamento desses dados foi então realizado no Excel, onde se eliminaram os duplicados e os registos em branco, garantindo que apenas *posts* com conteúdo textual fossem considerados na análise. Além disso, a ordenação cronológica dos *posts* encontrava-se invertida em relação aos dados de qualidade do ar (ou seja, começava pelas 23:59 e terminava às 00:00), pelo que foi necessário reordenar os *posts* do dia para que iniciassem à meia-noite e terminassem às 23h59, assegurando consistência entre os diferentes conjuntos de dados utilizados.

3.2.2 Tratamento de Dados

A análise de sentimentos, ou *sentiment analysis*, constitui uma área fundamental dentro do PLN, com o objetivo de identificar, extrair e classificar emoções ou opiniões presentes em textos. No contexto da presente investigação, esta técnica revela-se essencial para inferir os impactos emocionais associados à qualidade do ar, através da observação de conteúdos partilhados nas redes sociais.

Existem diversas abordagens para realizar análise de sentimentos, sendo que as mais comuns podem ser agrupadas em três grandes categorias: métodos baseados em léxicos, em aprendizagem supervisionada, e em modelos de *deep learning*. Os métodos léxicos recorrem a dicionários previamente construídos, onde as palavras estão associadas a polaridades e intensidades emocionais. Exemplos mais conhecidos incluem o SentiWordNet [47], o AFINN [48], o LIWC [49], o textblob [50],e o VADER [51], este último especialmente desenvolvido para analisar conteúdos informais como os das redes sociais. Estas ferramentas têm a vantagem de não requererem treino prévio e são, por isso, particularmente úteis quando não se dispõe de conjuntos de dados anotados.

Por outro lado, os métodos supervisionados requerem conjuntos de dados rotulados para treinar algoritmos como Naive Bayes, Support Vector Machines ou RF [52]. Embora geralmente mais precisas do que os métodos léxicos, estas abordagens exigem maior esforço na recolha e anotação de dados. Mais recentemente, os modelos baseados em *deep learning*, como redes LSTM, GRU ou arquiteturas do tipo *transformer* (por exemplo, o BERT) [53], tornaram-se o estado da arte na análise de sentimentos, oferecendo resultados altamente precisos. No entanto, estes modelos são mais exigentes do ponto de vista computacional e mais complexos em termos de interpretação.

No presente trabalho optou-se pela utilização do algoritmo VADER incluído na biblioteca Natural Language Toolkit (NLTK) em Python. Foi feita esta escolha, por causa da natureza dos dados analisados, que são provenientes de redes sociais. O VADER foi concebido especificamente para este tipo de conteúdo, sendo capaz de reconhecer aspetos como emojis, variações de pontuação e maiúsculas, gírias ou expressões informais. Para além disso, apresenta bons níveis de precisão, mesmo sem necessidade de treino prévio, o que o torna bastante prático para qualquer tipo de análises. A sua facilidade de implementação e baixo custo computacional foram fatores adicionais que contribuíram para a sua seleção. Apesar de existirem algoritmos mais avançados e sofisticados, como os baseados em *artificial neural networks*, a escolha deste modelo representa um equilíbrio entre simplicidade e eficiência para os objetivos específicos deste estudo.

Para permitir uma análise de sentimentos mais eficaz com o modelo VADER, que opera de forma mais fiável em inglês, foi necessário traduzir todos os textos da coluna *posts*,

originalmente em português no Excel. Este processo permitiu criar um segundo ficheiro idêntico ao original, mas com os textos já traduzidos para inglês, preservando a estrutura e os nomes das colunas, por fim este ficheiro foi guardado no formato comma-separated values (CSV). Esta escolha de formato facilitou a integração com linguagens de programação como Python, que foi utilizada em conjunto com o ambiente Google Colab, uma plataforma que permite a execução de código na *cloud* sem necessidade de instalação local.

O VADER, com recurso à biblioteca NLTK, atribui um score de sentimento a cada post, classificando-o automaticamente como positivo, negativo ou neutro com base num léxico pré-treinado. O processo consistiu em aplicar o modelo a cada texto presente na coluna de interesse (*posts*), sendo guardado o valor do score e a respetiva classificação. Posteriormente, foi feita uma agregação por dia para contabilizar o número de publicações de cada tipo de sentimento (positivo, negativo e neutro). Esta contagem diária permitiu também gerar um gráfico com a distribuição de sentimentos ao longo do tempo.

Por fim, o ficheiro com os resultados foi exportado para .csv e incluído como parte do *dataset* final, já com as colunas adicionais de sentimento e score, que iram ser uteis para futuras análises.

Para verificação e validação dos dados da análise sentimental utilizou-se uma abordagem semelhante, mas usando o algoritmo textblob já referido acima, onde foram obtidos os seguintes resultados, da esquerda para a direita VADER e Textblob, **Figura 14**.

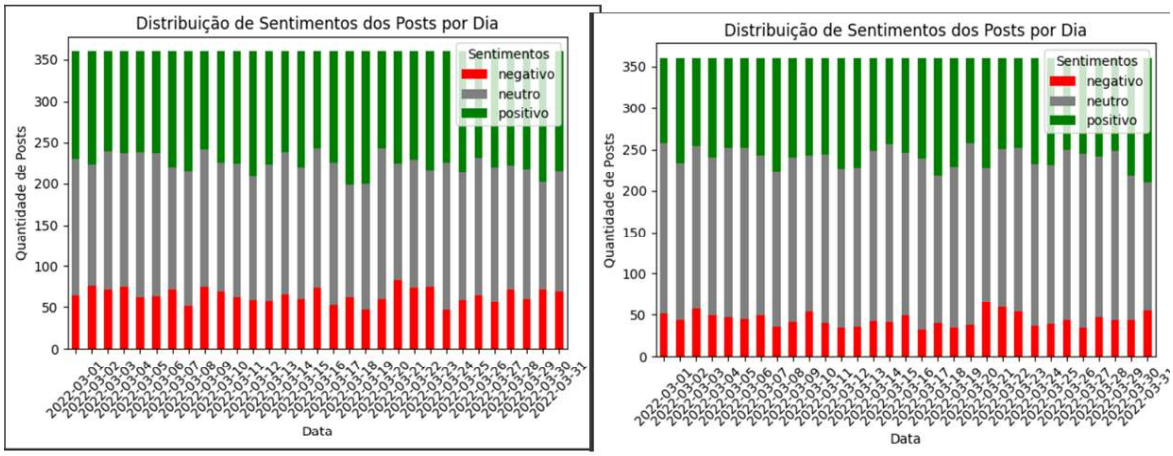

Figura 14 - Comparação de algoritmos Vader x Textblob.

Comparando os resultados dos dois algoritmos em gráficos, pode-se assim verificar a similaridade dos valores apesar deste segundo (textblob) ter mais tendência para os neutros, confirmando assim a autenticidade dos dados, procedendo-se com os dados do primeiro algoritmo, VADER.

3.2.3 Estrutura do Dataset

Resumindo **Tabela 5**, este *dataset* inclui cinco colunas. A coluna user identifica o nome de utilizador do autor da publicação, permitindo distinguir entre diferentes fontes de conteúdo. A coluna *posts* corresponde ao texto integral do *post* e constitui o principal objeto de análise do estudo.

A variável *date* representa a data associada a cada *post*, no formato YYYY-MM-DD, indicando o dia exato da publicação. Já a coluna *score* contém o valor numérico atribuído

a cada *post* pela análise de sentimentos realizada com o algoritmo VADER, refletindo o grau de polaridade emocional da publicação. Este *score* varia entre -1 (muito negativo) e 1 (muito positivo), com valores próximos de zero indicando neutralidade.

Por fim, a coluna *sentiment* classifica cada *post* em três categorias, positivo, negativo ou neutro com base no *score* gerado, permitindo uma interpretação qualitativa da opinião expressa no conteúdo textual. Importa referir que a análise de sentimentos foi aplicada exclusivamente sobre o conteúdo textual presente na coluna *posts*, sendo esta a única variável analisada do ponto de vista semântico.

Tabela 5 - Estrutura do dataset da análise de sentimentos.

Variável/Parâmetro	Descrição
User	Nome de utilizador do autor da publicação. Permite distinguir entre diferentes fontes de conteúdo.
Post	Texto integral do *post*. Principal objeto de análise do estudo, sobre o qual foi aplicada a análise de sentimentos.
Date	Data de publicação no formato YYYY-MM-DD, indicando o dia exato da publicação.
Score	Valor numérico da análise de sentimentos com o algoritmo VADER. Varia entre -1 (muito negativo) e 1 (muito positivo), sendo que valores próximos de zero indicam neutralidade.
Sentiment	Classificação qualitativa da polaridade do *post*: positivo, negativo ou neutro, com base no score atribuído.

3.2.4 Apresentação Gráfica do Dataset

Passando agora à perspetiva de análise sentimental, o gráfico da **Figura 15** apresenta o score médio diário de sentimento nas redes sociais.

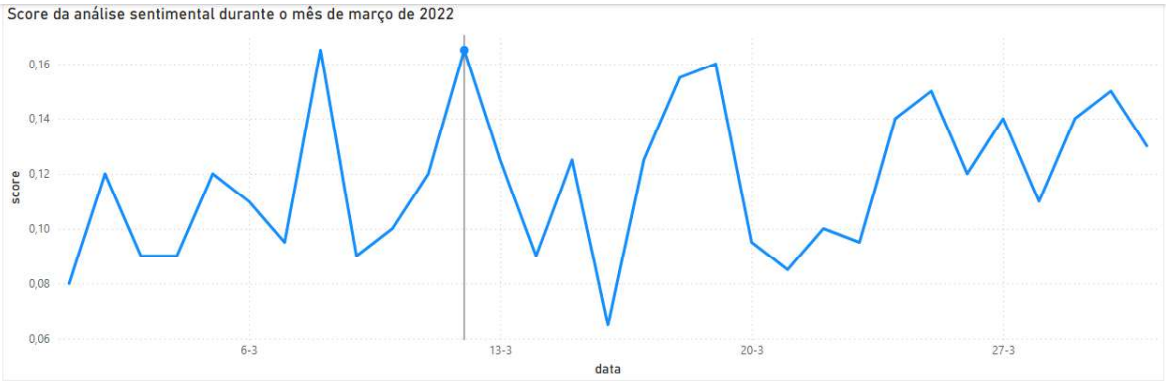

Figura 15 - Indicie de sentimento (score) x data.

O gráfico mostra uma estabilidade significativa. Durante todo o mês, os valores médios diários do *score* variam apenas entre 0.04 e 0.20, mantendo-se dentro de uma faixa levemente positiva. Mesmo nos dias mais críticos em termos de poluição atmosférica como 15, 16 e 24 de março não se verifica uma queda significativa no score. Pelo contrário, em alguns desses dias o sentimento até melhora ligeiramente.

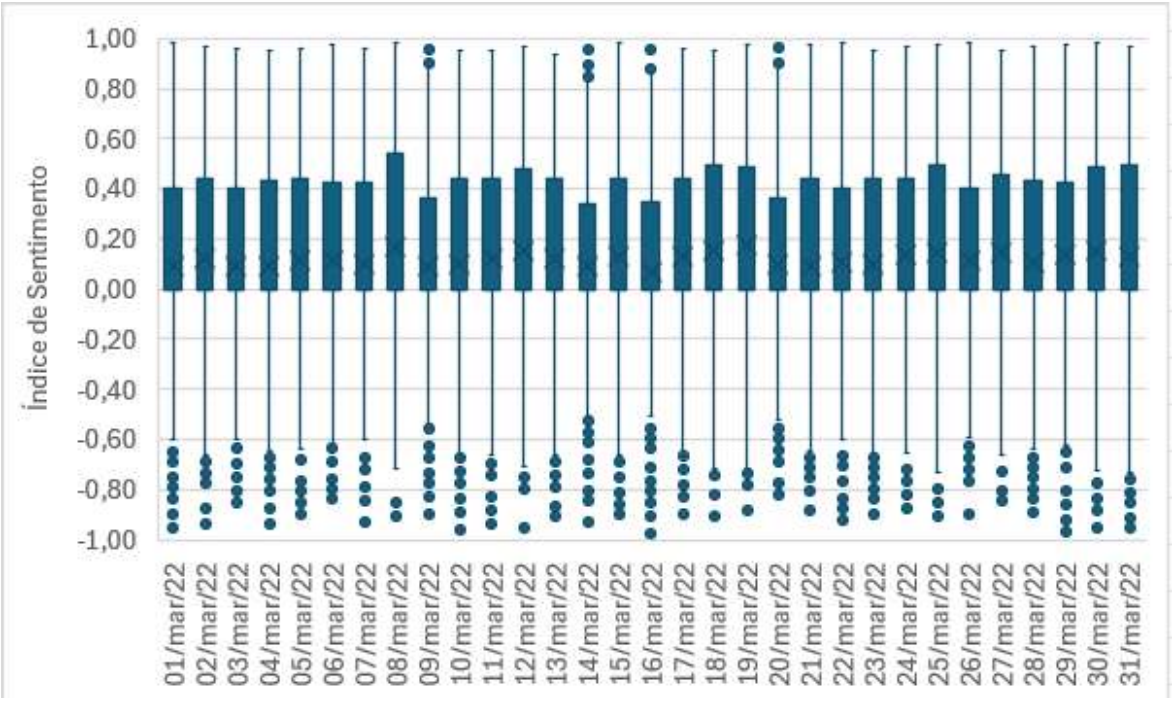

Figura 16 - Boxplot do índice de sentimento x data.

Utilizando outro tipo de gráfico como o *boxplot*, **Figura 16**, onde se podem visualizar informações diferentes, como que o primeiro quartil (Q1) é igual a zero ao longo de todo o mês de março. Este fenómeno ocorre porque uma parte significativa das publicações apresenta um score neutro, ou seja, com valor exatamente igual a zero. Como o Q1 representa o valor abaixo do qual se encontra 25% dos dados, a sua coincidência com o valor zero indica que pelo menos 25% das publicações têm sentimentos neutros. Esta característica é comum em análises de sentimentos de redes sociais, onde muitas mensagens não expressam emoções claras ou polarizadas, também são detetados alguns *outliers*, marcados pelos círculos no gráfico, o resto da informação está dentro do esperado, e será analisada no próximo capítulo.

3.3 Junção de Datasets

Durante o processo de planeamento da análise de sentimentos ainda foi discutida a pertinência de realizar uma análise temporal mais detalhada. Como resultado dessa discussão, decidiu-se dividir os dados em dois períodos distintos: antes e depois das 18h de cada dia. Esta separação foi escolhida por corresponder aproximadamente ao fim do horário laboral típico, momento em que muitas pessoas saem do trabalho e têm maior disponibilidade para aceder às redes sociais, interagir com conteúdos e expressar as suas opiniões. Além disso, é comum que notícias de maior impacto, atualizações de eventos ou medidas governamentais sejam divulgadas durante a tarde, o que pode gerar reações mais marcadas no período pós-laboral. Assim, esta segmentação permite identificar potenciais variações no sentimento das publicações consoante o momento do dia.

Como não foi possível obter automaticamente a hora exata de todos os *posts*, foi necessário identificar manualmente o *post* publicado às 18:00 de cada dia, servindo este como ponto de referência, com isto foi então criada uma coluna denominada período, com duas categorias: Antes das 18 e Depois das 18.

Após o cruzamento entre os dados ambientais e os resultados da análise de sentimentos, foi construída uma versão consolidada do *dataset* **Tabela 6**, que reflete todas as variáveis relevantes para o estudo. Esta versão final integra medições horárias de qualidade do ar e condições meteorológicas com os valores de sentimento extraídos de publicações em redes sociais.

Tabela 6 - Estrutura do dataset da junção de todos os parâmetros.

Variável/Parâmetro	Descrição
Data	Data e hora da medição no formato YYYY-MM-DD HH:MM. Representa o instante temporal de cada registo (valores horários).
Dióxido de Enxofre (µg/m³)	Concentração de dióxido de enxofre (SO2) no ar.
Partículas < 10 µm (µg/m³)	Concentração de partículas inaláveis (PM10). Variável com maior destaque devido à sua variabilidade durante o mês analisado.
Partículas < 2.5 µm (µg/m³)	Concentração de partículas finas (PM2.5), mais prejudiciais para a saúde.
Ozono (µg/m³)	Concentração de ozono troposférico (O3).
Dióxido de Azoto (µg/m³)	Presença de dióxido de azoto (NO2), associado ao tráfego urbano.
Monóxido de Carbono (mg/m³)	Concentração de monóxido de carbono (CO).
Benzeno (µg/m³)	Concentração de C6H6, composto tóxico volátil.
Temp. (ºC)	Temperatura do ar em graus Celsius.
H.Rel (%)	Humidade relativa do ar (%).
Vel. vi. (Km/h)	Velocidade do vento em km/h.
Prec. (mm)	Precipitação acumulada na hora anterior (mm).
Score	Valor de análise de sentimentos (VADER) aplicado ao conteúdo textual. Varia entre -1 (negativo) e 1 (positivo). No *dataset* final, é a única variável proveniente do Twitter.

4 Análise Score Sentimental vs Dados Ambientais

Este capítulo apresenta uma análise conjunta entre os dados de sentimento expressos em publicações online e os níveis de poluição atmosférica, já com uma divisão diária entre antes e depois das 18 horas, com foco especial na concentração de partículas inaláveis (PM10). Complementarmente, foi ainda realizado uma análise de correlação envolvendo todas as variáveis ambientais e o score sentimental, com o objetivo de avaliar possíveis associações entre esses parâmetros e os scores de sentimento.

4.1 Analise do Score Sentimental e PM10

Para facilitar a interpretação dos resultados da análise de sentimentos, foi elaborado um gráfico *boxplot* que representa visualmente a distribuição dos scores do sentimento ao longo do tempo. Este tipo de gráfico permite observar de forma clara a mediana, os quartis e possíveis variações ou dispersões nos dados, evidenciando não apenas a tendência central (mediana), mas também a amplitude dos sentimentos expressos nas publicações. O gráfico da **Figura 17** corresponde, portanto, à análise dos sentimentos (*score*) das publicações, divididas por data e período do dia (antes e depois das 18h).

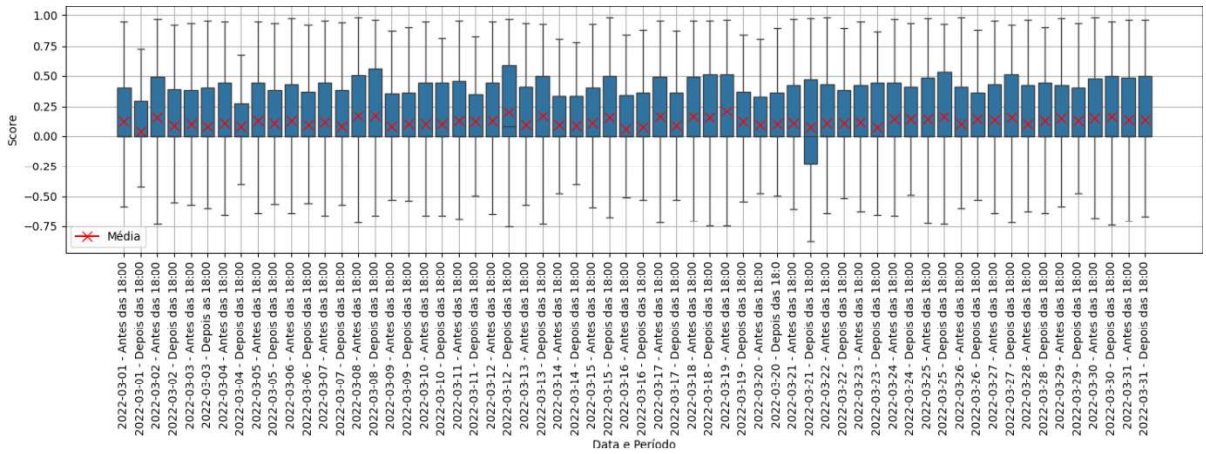

Figura 17 - Boxplot de índice de sentimento (score) x data e período.

Ao longo do mês de março de 2022, observa-se que a mediana dos *scores* de sentimento se manteve estável em 0.0 em praticamente todos os períodos. Como referido anteriormente, estes valores são expectáveis já que existem muitos valores exatamente em 0 por terem sentimento neutro, tanto antes quanto depois das 18h. Isto indica que, de forma geral, as publicações apresentavam um tom neutro, sem predominância de sentimentos positivos nem negativos. Ainda assim, há alguma variação nos valores superiores, especialmente no terceiro quartil, que em muitos dias atingiu valores entre 0.4 e 0.5, sugerindo que, embora a maioria dos conteúdos fosse neutra, havia uma parcela significativa de publicações com sentimento levemente positivo. A média acompanha esta tendência, variando maioritariamente entre 0.08 e 0.16, com picos ocasionais como acontece no dia 12 de março depois das 18h, quando atingiu 0.20 revelando momentos pontuais com maior expressão de positividade. Outro caso de interesse seria dia 21 depois das 18:00 já que é o único dia que temos um Q1 negativo, ou seja 25% dos valores estão abaixo de 0, o que pode indicar um evento que deixou esse sentimento à população. Apesar disso de modo geral, os dados apontam para uma constância emocional nas

publicações, com pequenas flutuações em direção ao positivo em determinados dias e horários.

Foi também elaborado o mesmo tipo de gráfico, **Figura 18**, para visualizar os níveis de partículas inaláveis (PM10) ao longo do tempo, com base nos valores médios, primeiros e terceiros quartis registados diariamente, segmentados por períodos do dia (antes e depois das 18h). Como visto no subcapítulo anterior as poerias do deserto do Saara que passaram pelo território nacional nesta altura tiveram uma influência significativa neste parâmetro da qualidade do ar.

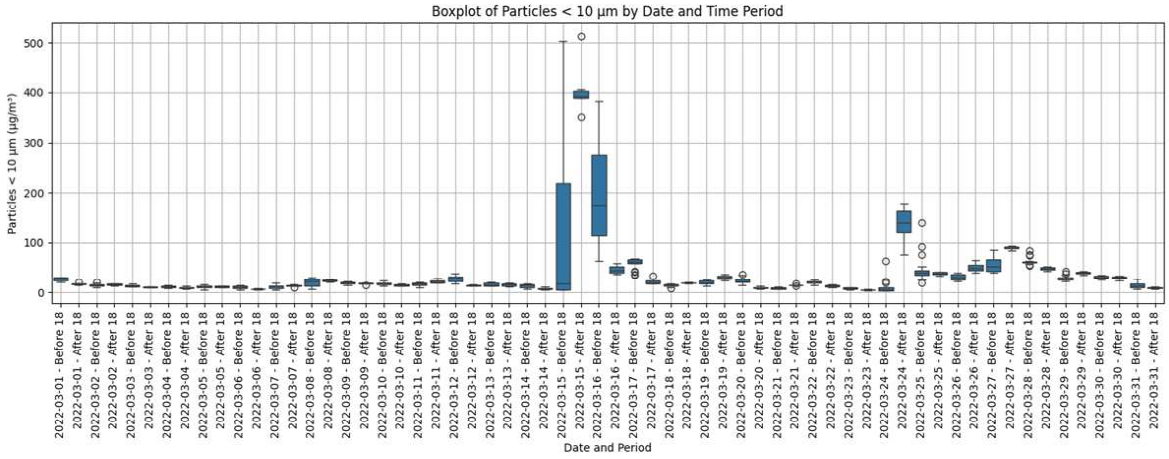

Figura 18 - Boxplot de partículas PM10 x data e período.

Em muitos dos dias com baixos níveis de poluição, especialmente na primeira quinzena de março, o *boxplot* mostrou caixas estreitas, com valores de Q1 e Q3 relativamente próximos e medianas alinhadas com a média, indicando uma distribuição simétrica e estável dos valores. Nesses casos, a poluição atmosférica manteve-se controlada, com variações pouco acentuadas entre os períodos do dia (antes e depois das 18h).

Por outro lado, nos dias em que se registaram picos elevados de PM10, como nos dias 15, 16, 24 e 27 de março, o *boxplot* revelou caixas mais largas e valores extremos (*outliers*), sinalizando grande dispersão nos dados e os episódios de poluição anómala. Nestes dias, foi comum observar que a média se afastava da mediana, o que sugere a presença de assimetrias na distribuição, muitas vezes puxada por valores extremamente altos. Por exemplo, no dia 15 de março depois das 18h, a média ultrapassou os 400 µg/m³, enquanto a mediana (não visível no dado bruto, mas interpretável no *boxplot*) estaria claramente mais baixa, evidenciando a distorção causada por valores muito elevados. Em especial no dia 15 depois das 18h em que a diferença entre quartis foi bastante estreita, evidenciando talvez a altura de maior intensidade do evento, sendo que todos os valores nessa altura estão bem acima dos 350 µg/m³. O *boxplot* foi fundamental para identificar esses padrões, ao permitir visualizar facilmente o comportamento da mediana, a amplitude interquartil (Q1–Q3) e a ocorrência de *outliers*.

4.1.1 Analise de Correlações

No gráfico da **Figura 19** cada ponto representa a média do score e o valor médio de PM10 para um dia específico e um período do dia (antes ou depois das 18h). Isso significa que cada ponto traz uma informação precisa sobre a qualidade do ar e o score sentimental naquele momento específico, permitindo analisar como essas variáveis se comportam ao longo do tempo e em diferentes momentos do dia. Para isso foram analisados

simultaneamente os dois *datasets* e criado um gráfico de correlação para avaliar se existe algum tipo de relacionamento entre as duas variáveis.

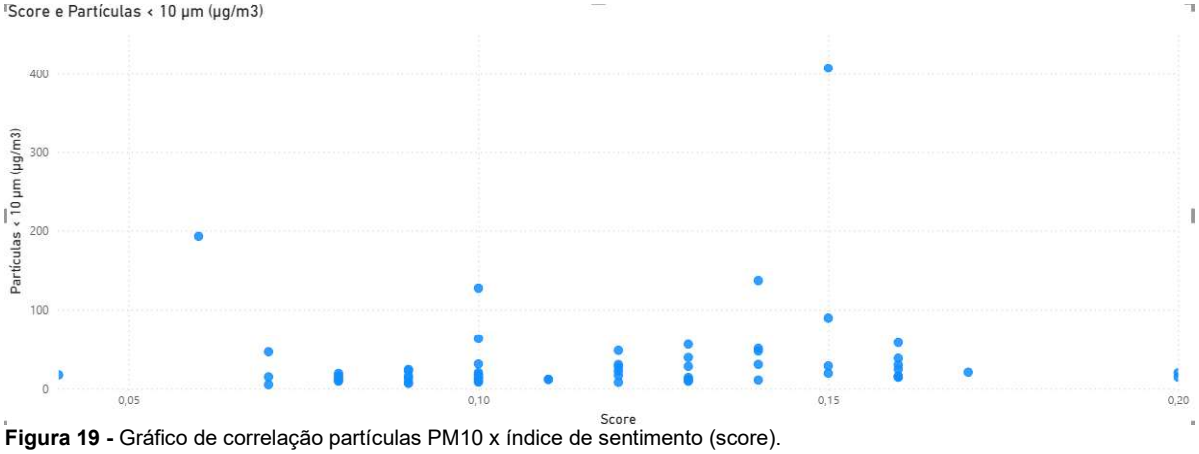

Figura 19 - Gráfico de correlação partículas PM10 x índice de sentimento (score).

A correlação entre a média do score e a concentração média de partículas PM10 representada na **Figura 19** parece ser fraca ou pouco consistente. Em muitos casos, aumentos nos níveis de PM10 não correspondem a aumentos claros no score, e vice-versa. Por exemplo, nos dias com valores muito elevados de PM10, como os dias 15 e 24 de março após as 18h (com PM10 acima de 100 µg/m³), o score não apresenta um aumento proporcional significativo, permanecendo em valores relativamente baixos ou moderados. Também há momentos em que valores de PM10 baixos ou médios aparecem com scores altos, o que indica que não há uma relação direta evidente. Portanto, com base nesses dados, não parece haver uma correlação forte e direta entre o score e os níveis de PM10.

Esta comparação mostra uma desconexão entre o que realmente se passou e a perceção, ou reação social, refletida nas redes. Embora os dados mostrem que a qualidade do ar foi severamente afetada, visto especialmente no PM10, que teve uma maior resposta à passagem das poeiras como o era esperado, o estado emocional do público nas redes sociais permaneceu praticamente inalterado.

Para investigar em concreto a relação entre o score e os níveis de partículas PM10, foi calculado o coeficiente de correlação de Pearson [54]. O coeficiente encontrado foi de aproximadamente 0.1,**Figura 20**, indicando uma correlação linear muito fraca entre as variáveis **Tabela 7**.

Além disso, o valor-p obtido foi de 0.45, muito superior ao nível de significância usual de 0.05, o que implica que não há evidências estatísticas suficientes para rejeitar a hipótese nula de ausência de correlação linear. Portanto, com base nesta análise, conclui-se que não existe uma relação linear significativa entre o score e os níveis de PM10 no período analisado.

Usando bibliotecas de Python construiu-se um mapa de correlações para se poder observar quais parâmetros que se correlacionam tanto com o score sentimental como em comparação com qualquer outro parâmetro, **Figura 20**.

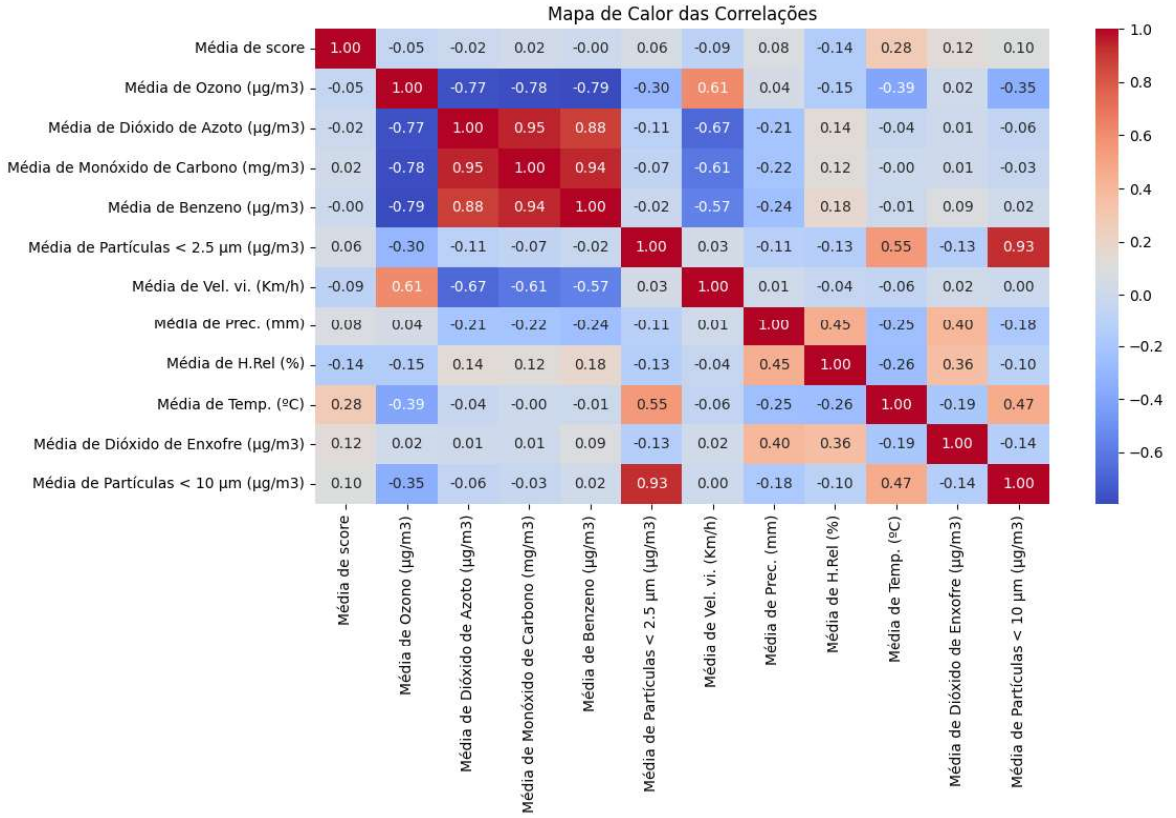

Figura 20 - Mapa de calor de correlações com todos os parâmetros.

Para ajudar a uma melhor compressão apresenta-se na **Tabela 7** a classificação dos intervalos de correlação [55].

Tabela 7 - Classificação dos intervalos de Pearson.

Intervalo (valor de r)	Tipo de correlação	Interpretação
1.0	Correlação perfeita	As variáveis variam exatamente juntas
0.9 a 1.0 ou -0.9 a -1.0	Correlação muito forte	Relação muito forte (positiva ou negativa)
0.7 a 0.9 ou -0.7 a -0.9	Correlação forte	Relação clara e consistente
0.5 a 0.7 ou -0.5 a -0.7	Correlação moderada	Relação significativa, mas não perfeita
0.3 a 0.5 ou -0.3 a -0.5	Correlação fraca	Relação fraca, pode haver outros fatores
0.0 a 0.3 ou -0.3 a 0.0	Correlação muito fraca ou nula	Pouca ou nenhuma relação linear

Consultando então o mapa da **Figura 20** e a **Tabela 7**, observa-se que a variável com a correlação mais forte com o score de sentimentos foi a temperatura, com um valor de 0.28. Isso indica praticamente uma correlação positiva fraca à medida que a temperatura aumenta, o score de sentimento tende a aumentar apesar de ligeiramente, ou seja, as pessoas parecem expressar emoções um pouco mais positivas em dias mais quentes. Este é o valor mais alto em termos de correlação com o score, sendo que os outros parâmetros têm correlação nula ou muito fraca.

Relativamente aos outros parâmetros e eventuais correlações entre si, as correlações mais fortes do conjunto de dados ocorrem entre os poluentes NO2, CO, C6H6, PM2.5 e

PM10, com coeficientes superiores a 0.88. Estas relações refletem a origem comum destes poluentes em processos de combustão, nomeadamente do tráfego rodoviário urbano. A sua elevada associação indica que tendem a variar em conjunto, reforçando o papel central dos veículos na poluição atmosférica observada. Assim, ações direcionadas à redução das emissões veiculares podem ter efeitos positivos em múltiplos parâmetros da qualidade do ar.

Por outro lado, o ozono (O3) apresenta correlações negativas relativamente fortes com C6H6 (-0.79), CO (-0.78) e NO2 (-0.77). Essa relação inversa é coerente com os mecanismos fotoquímicos envolvidos na formação do ozono troposférico: em ambientes muito poluídos, especialmente urbanos, o ozono tende a ser destruído por reações com óxidos de azoto (NO), o que explica a diminuição do ozono quando os poluentes primários estão elevados,[56]. Além disso, em zonas menos poluídas, o ozono pode formar-se mais facilmente devido à presença de radiação solar e compostos voláteis orgânicos.

Uma variável ambiental importante a considerar é a velocidade do vento, que mostra correlações negativas moderadas com NO2 (-0.67), CO (-0.61) e C6H6 (-0.57), e positiva com ozono (0.61). Isto sugere que o vento atua como um agente de dispersão dos poluentes locais, reduzindo a sua concentração ao aumentar a ventilação atmosférica. No caso do ozono, o aumento com a velocidade do vento pode dever-se ao transporte deste poluente desde áreas suburbanas ou rurais onde se forma mais facilmente, além da sua relativa estabilidade em comparação com poluentes primários.

Por fim, a correlação entre temperatura e partículas finas PM2.5 (0.55) também merece atenção. Em contextos urbanos, dias mais quentes podem promover reações químicas e a formação secundária de partículas finas, ou simplesmente coincidir com condições meteorológicas (como estabilidade atmosférica) que dificultam a dispersão dos poluentes.

5 Análise Complementar com Modelos de Machine Learning

Embora o foco principal deste estudo tenha sido a análise descritiva e a identificação de padrões através de correlações, optou-se por incluir uma análise complementar utilizando algoritmos de ML, com o objetivo de enriquecer a compreensão dos dados e explorar o seu potencial preditivo.

5.1 Seleção dos Algoritmos

Foram selecionados dois algoritmos: DT e RF. A escolha destes modelos baseia-se na sua adequação a problemas exploratórios e preditivos em dados ambientais. As DT são particularmente úteis pela interpretação direta dos critérios de decisão, permitindo identificar rapidamente quais as variáveis mais influentes no comportamento da qualidade do ar. Já o RF, que combina múltiplas Árvores de Decisão, oferece maior robustez e precisão, reduzindo o risco de *overfitting* e fornecendo uma estimativa mais estável da importância das variáveis.

A utilização destes modelos permite assim complementar as análises estatísticas anteriores, não só identificando relações lineares, mas também captando padrões não lineares e interações entre poluentes e variáveis meteorológicas, contribuindo para uma perspetiva mais rica sobre o comportamento dos parâmetros monitorizados.

5.2 Implementação e Resultados

Neste subcapítulo descreve-se a implementação prática dos algoritmos de ML selecionados, detalhando as bibliotecas utilizadas, os parâmetros de treino e os principais resultados obtidos. A análise foca-se inicialmente na DT e, posteriormente, no RF, permitindo comparar o desempenho de cada abordagem no contexto dos dados ambientais e sociais recolhidos.

5.2.1 Decision Tree

Inicialmente, realizou-se a construção de uma DT utilizando todas as variáveis disponíveis, recorrendo ao algoritmo DecisionTreeRegressor da biblioteca Scikit-learn (Python). O modelo foi treinado com os parâmetros padrão: o critério de divisão utilizado foi o Mean Squared Error (MSE), a profundidade da árvore foi ilimitada (max_depth=None), e o número mínimo de amostras para divisão de um nó foi de 2 (min_samples_split=2). Utilizou-se ainda random_state=42 para garantir a reprodutibilidade dos resultados. O modelo resultante revelou-se extenso e complexo, dificultando a interpretação visual. Por esse motivo, procedeu-se à análise da importância das variáveis, sendo identificadas como mais influentes: o monóxido de carbono (CO), a temperatura, as partículas finas PM2.5 e a humidade relativa.

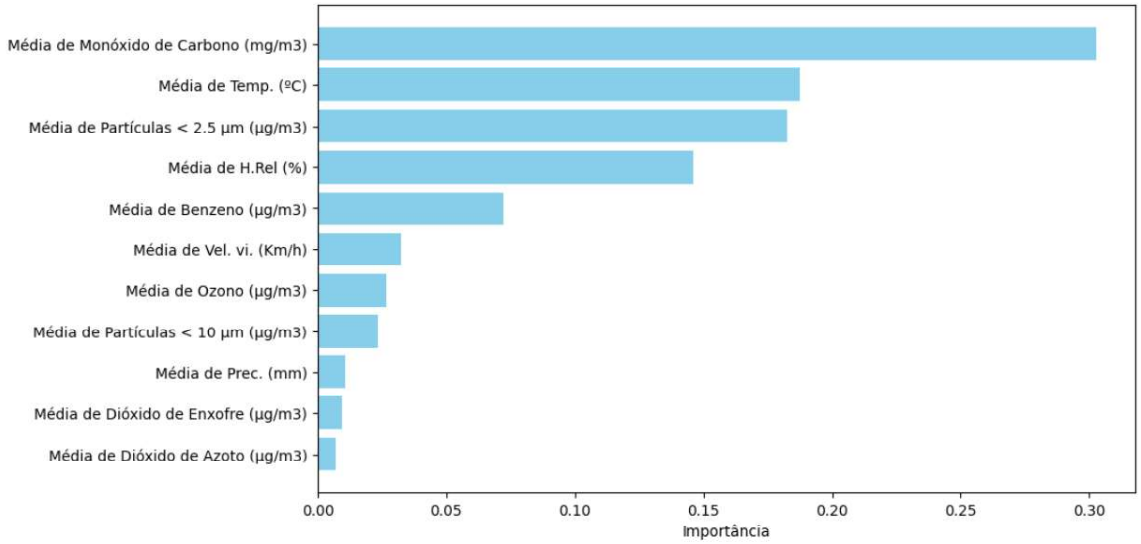

Figura 21 - Ordem de importância das variáveis no decision tree.

Com base nesta seleção, **Figura 21**, construiu-se uma DT simplificada utilizando apenas essas quatro variáveis mais relevantes. A árvore reduzida apresenta uma estrutura mais compacta e de fácil interpretação, facilitando a compreensão do impacto dos principais parâmetros sobre a predição.

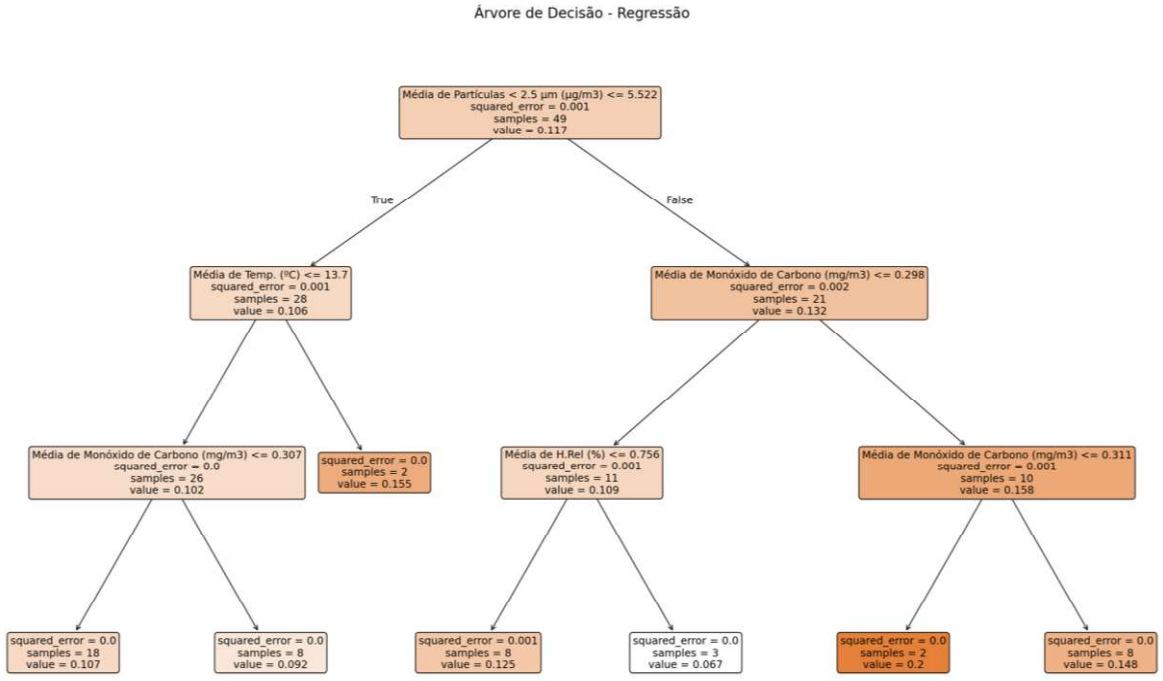

Figura 22 - Decision tree.

A DT apresentada na **Figura 22** realiza uma regressão para prever o valor de uma variável-alvo associada aos sentimentos das pessoas, com base em quatro variáveis ambientais principais: Média de Partículas Finas (PM2.5 < 2.5 µm), Temperatura Média (°C), Média de Monóxido de Carbono (CO, mg/m³) e de Humidade Relativa (%).

No nó raiz, a árvore utiliza a concentração de partículas PM2.5 com um limite de 5.522 µg/m³ para dividir os dados. Essa escolha inicial mostra que a poluição por partículas finas tem um papel central na separação inicial das amostras, o que está de acordo com

evidências científicas sobre os impactos da poluição na saúde física e emocional das pessoas. Para os casos em que o nível de PM2.5 é menor ou igual a 5.522, a árvore avalia a temperatura média, com ponto de corte em 13.7°C. Essa divisão sugere que condições térmicas mais baixas podem afetar a dispersão de poluentes e, por consequência, o impacto sobre o bem-estar das pessoas. A partir daí, a árvore passa a usar o CO para refinar a predição, mostrando a sua importância direta na modelação dos sentimentos. Nos grupos com baixa poluição e temperatura mais baixa, a previsão da variável-alvo é de 0.102. Considerando que os sentimentos são medidos numa escala de -1 (muito negativo) a 1 (muito positivo), esse valor indica um sentimento levemente positivo ou próximo do neutro.

Do outro lado da árvore, nos casos em que PM2.5 é maior que 5.522, o modelo continua a usar o CO como principal critério de divisão (limite de 0.298 mg/m³), e mesmo com um valor máximo de 0.2 ainda relativamente baixo dentro da escala, indica uma tendência para sentimentos um pouco mais positivos ou neutros, mas longe dos extremos máximos da escala. Em seguida, dependendo do caminho, a humidade relativa também é usada para dividir os dados. Isso mostra que, sob condições de poluição mais intensa, a interação entre gases tóxicos e variáveis climáticas torna-se mais relevante.

Apesar de a estrutura da árvore ser coerente, as métricas globais do modelo indicam um desempenho insatisfatório para fins preditivos: o MSE é 0.00295, indicando um erro médio pequeno, mas o Coeficiente de Determinação (R^2) é negativo (-2.10), o que revela que o modelo explica menos a variabilidade dos dados do que uma simples média. Ou seja, embora a árvore consiga separar os dados em grupos com comportamentos distintos e consistentes, ela não generaliza bem para o conjunto completo. Como temos apenas 62 dados para treinar, é comum que a árvore se tenha ajustado demais a essas informações específicas, capturando detalhes que talvez não apareçam em novos dados. Isso faz com que o modelo funcione bem com os dados usados no treino, mas tenha dificuldade para prever corretamente em situações diferentes.

Apesar das limitações do modelo para previsões precisas, ele cumpre bem o papel de ferramenta exploratória, ajudando a identificar e entender as possíveis relações entre a poluição do ar e o estado emocional das pessoas. Esse estudo também abre espaço para o desenvolvimento e teste de novos algoritmos que possam melhorar a capacidade preditiva no futuro.

5.2.2 Random Forest

Para aprofundar a análise realizada com a DT, foi utilizado o modelo RF. Esta técnica combina múltiplas árvores de decisão para melhorar a robustez e a precisão das previsões, além de fornecer uma visão mais consistente sobre a importância das variáveis no problema em estudo.

Aplicamos o modelo RF para prever os sentimentos das pessoas com base nas variáveis ambientais estudadas. Na **Figura 23** são apresentados por ordem a importância das variáveis no modelo RF.

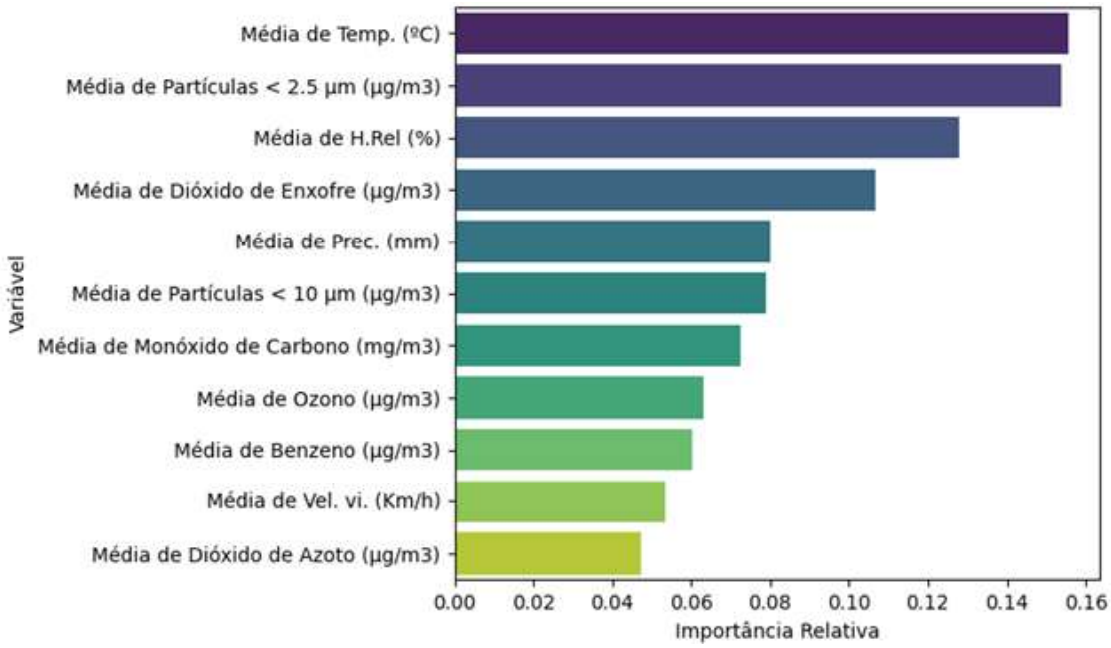

Figura 23 - Ordem de importância das variáveis no random forest.

Comparando com a DT, podemos ver que as quatro variáveis mais importantes para o modelo agora foram Temperatura Média (°C), Média de Partículas Finas (PM2.5 < 2.5 µm), Humidade Relativa Média (%) e a Média de Dióxido de Enxofre (SO_2, µg/m³), embora de ordem trocada, as variáveis apresentam ordens de importância parecidas, com a exceção do monóxido de carbono que foi a variável de maior importância na DT, e aqui no caso RF foi apenas a sétima mais importante. Esta diferença pode ser explicada pelo funcionamento distinto dos dois algoritmos, a DT baseia-se numa única árvore e é altamente sensível às divisões iniciais dos dados, o que pode destacar exageradamente uma variável que, por acaso, cria uma boa separação no conjunto de treino. Já o RF constrói várias árvores com subconjuntos aleatórios dos dados e variáveis, calculando a importância média de cada uma. Assim, variáveis que são consistentes em várias árvores, como a temperatura ou PM2.5, tendem a ganhar mais importância. Se o CO for altamente correlacionado com outras variáveis ou só for útil em contextos muito específicos, a sua importância média no RF será reduzida, o que pode justificar a sua menor relevância nesse modelo mais robusto.

Quanto ao desempenho, o modelo RF apresentou um MSE de aproximadamente 0.00109, inferior ao da DT, indicando melhor ajuste aos dados. No entanto, o Coeficiente de Determinação (R^2) permaneceu negativo, em torno de -0.14, o que é uma melhoria em comparação como a DT, mas ainda significa que o modelo ainda não explica a variabilidade dos sentimentos melhor do que a média simples.

Uma vantagem importante do RF é a capacidade de gerar previsões individuais para cada amostra, facilitando a comparação direta entre valores reais e previstos. Essa granularidade ajuda a compreender melhor o comportamento do modelo e a identificar pontos fortes e limitações em suas previsões.

A análise dos valores reais e previstos, **Tabela 8**, pelo modelo RF mostra que embora em várias amostras as previsões estejam próximas dos valores observados, em alguns casos as diferenças são mais expressivas, como para um valor real de 0.16 existe uma previsão de 0.11. Isso indica que o modelo consegue captar as tendências gerais dos sentimentos, mas ainda apresenta limitações para estimar com precisão todas as variações

individuais. Esses resultados reforçam que o modelo oferece estimativas razoáveis, ainda que não seja um modelo confiável em todos os casos.

Tabela 8 - Real x previsto no modelo random forest.

Score Real	Score Previsto
0.10	0.1318
0.15	0.1078
0.12	0.1406
0.13	0.1083
0.08	0.1054
0.14	0.1120
0.08	0.1068
0.11	0.1293
0.16	0.1074
0.16	0.1261
0.06	0.1192
0.11	0.0982
0.14	0.1214

Por fim, como se pode ver pelo MSE e R^2 apesar da melhora no ajuste, o modelo RF ainda indica que as variáveis ambientais analisadas explicam de forma limitada as variações nos sentimentos. Ainda assim pode-se considerar que o modelo cumpre bem seu papel exploratório, identificando e validando a importância relativa dos fatores ambientais na modelagem dos sentimentos, e serve como base para futuras pesquisas que busquem aprofundar essa relação e desenvolver algoritmos com maior desempenho.

5.2.3 Teste com Algoritmo de Classificação

Além da análise de regressão, também foi explorada uma abordagem de classificação utilizando a DT para classificar os sentimentos em classes “Baixo”, “Médio” e “Alto”. Para isso, a variável contínua dos sentimentos, originalmente com valores entre -1 e 1, foi separada em três categorias, considerando que os valores observados estavam concentrados entre aproximadamente 0.04 e 0.2. Assim, definiu-se que scores menores que 0.09 pertencem à classe “Baixo” (0), scores entre 0.09 e 0.14 à classe “Médio” (1), e scores iguais ou superiores a 0.14 à classe “Alto” (2). Esta categorização permite uma interpretação mais clara dos resultados e possibilita o uso de métodos de classificação para analisar a relação entre qualidade do ar e estado emocional.

O relatório de classificação, **Tabela 9**, indica que o modelo tem desempenho moderado, com precisão, *recall* e f1-score maiores para a classe “Alto” (f1-score 0.67), enquanto apresenta dificuldades em prever corretamente a classe “Baixo” (f1-score 0.00). A exatidão geral ficou em 46%, mostrando que, apesar das limitações, o modelo captura algumas diferenças entre as classes.

Tabela 9 - Resultados do relatório de classificação.

Classe	Precisão (Precision)	Recall	F1-score
Baixo	0.00	0.00	0.00
Médio	0.33	0.40	0.36
Alto	0.57	0.80	0.67
Exatidão (Acuracy)			0.46

Os dados também foram representados numa matriz de confusão **Figura 24**.

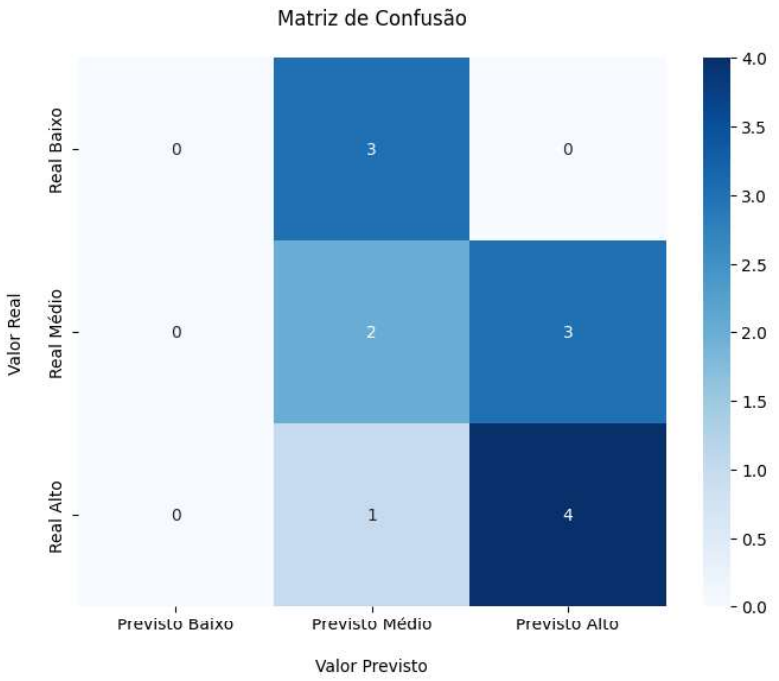

Figura 24 - Matriz de confusão.

Cada linha da matriz de confusão representa as classes reais, enquanto cada coluna representa as classes previstas pelo modelo, considerando as classes “Baixo” (linha 1), “Médio” (linha 2) e “Alto” (linha 3). Na linha 1, referente à classe real “Baixo”, nenhuma amostra foi corretamente classificada como “Baixo” (verdadeiros positivos), sendo que 3 amostras foram classificadas incorretamente como “Médio” (falsos negativos para “Baixo”) e nenhuma foi prevista como “Alto”. Na linha 2, referente à classe real “Médio”, nenhuma amostra foi classificada como “Baixo”, duas foram corretamente classificadas como “Médio” e 3 foram classificadas incorretamente como “Alto”. Já na linha 3, que corresponde à classe real “Alto”, nenhuma amostra foi classificada como “Baixo”, uma amostra foi incorretamente classificada como “Médio” e 4 amostras foram corretamente classificadas como “Alto”.

Esses resultados mostram que o modelo não conseguiu identificar nenhum exemplo da classe “Baixo” corretamente, classificando todos esses exemplos como “Médio”. A classe “Médio” foi parcialmente identificada, com 2 acertos, mas 3 exemplos foram confundidos com a classe “Alto”. A classe “Alto” apresentou melhor desempenho, com 4 de 5 exemplos classificados corretamente e 1 confundido com “Médio”. De modo geral, o modelo tende a confundir principalmente classes adjacentes, como “Baixo” com “Médio” e “Médio” com “Alto”, indicando uma dificuldade maior para distinguir as classes mais baixas. Essa matriz evidencia que o modelo tem desempenho razoável para a classe “Alto”, mas necessita aprimorar a identificação das classes “Baixo” e “Médio”.

Como vimos, a análise de classificação utilizando a DT apresentou limitações, especialmente na distinção das classes “Baixo” e “Médio”, com baixa precisão e *recall* para essas categorias. Considerando isso, optou-se por não aplicar a abordagem de classificação ao modelo RF. Isso porque os dados de sentimentos são originalmente contínuos, variando entre -1 e 1, e no conjunto analisado encontram-se concentrados numa faixa estreita (entre 0.04 e 0.2), o que dificulta a segmentação clara e significativa em classes discretas. Além disso, o RF, por sua natureza, geralmente requer um volume maior de dados para uma generalização eficiente, e a fragmentação em poucas classes pode

levar a um desempenho ainda mais limitado. Portanto, a regressão permanece como a abordagem mais adequada para este tipo de dados, permitindo prever diretamente os valores contínuos dos sentimentos e capturar melhor suas variações subtis.

6 Discussão

A análise desenvolvida neste trabalho permitiu compreender a relação entre um evento específico de degradação da qualidade do ar, no período da passagem de poeiras do Saara sobre Portugal em março de 2022, e a perceção pública expressa nas redes sociais. A metodologia aplicada integrou dados ambientais e meteorológicos, recolhidos através das plataformas Qualar e Meteomanz, com dados sociais provenientes do Twitter, processados via Power Automate. Esta integração possibilitou cruzar fenómenos físicos mensuráveis (como concentrações de partículas e gases) com manifestações sociais (sentimentos expressos em texto).

Os resultados confirmam que o episódio teve um impacto ambiental evidente. Os picos de PM10 atingiram valores muito elevados (superiores a 400 µg/m³ no dia 15 de março, depois das 18h), largamente acima dos limites habituais e claramente identificados como episódios de má qualidade do ar. Estes valores alinham-se com o que outros estudos sobre poeiras do Saara em Portugal e na Península Ibérica têm documentado, reforçando que o fenómeno tem consequências diretas e mensuráveis na atmosfera.

No entanto, ao contrário do que alguns trabalhos anteriores observaram em contextos de poluição urbana prolongada, neste caso não se verificou uma correlação forte entre o agravamento da qualidade do ar e a negatividade expressa nas redes sociais. O coeficiente de correlação de Pearson entre as duas variáveis (score sentimental e PM10) foi de aproximadamente 0.1, indicando uma relação fraca e estatisticamente não significativa. Ainda assim, observaram-se situações pontuais interessantes, como no dia 15 de março depois das 18h: os valores de PM10 estavam extremamente elevados, mas os sentimentos expressos online mantiveram-se praticamente inalterados. Este contraste sugere que os efeitos físicos do evento foram muito mais evidentes nos dados ambientais do que nos sociais, levantando a questão sobre até que ponto as redes sociais refletem a perceção pública em eventos ambientais extraordinários.

Comparando com os estudos relacionados, verificam-se diferenças relevantes. Trabalhos internacionais que exploraram a relação entre poluição atmosférica e redes sociais encontraram, em alguns casos, correlações moderadas ou fortes, especialmente em países onde a população está mais sensibilizada para questões ambientais, como a China. Nestes contextos, aumentos da poluição estiveram frequentemente associados a um maior número de publicações negativas, muitas vezes acompanhadas de hashtags relacionadas com saúde ou meio ambiente. Em contraste, no caso português analisado, a resposta online foi mais ténue, sugerindo que fatores culturais, sociais ou mesmo o baixo volume de utilização do Twitter em Portugal podem ter influenciado os resultados. Este contributo mostra que metodologias semelhantes aplicadas em contextos distintos podem gerar padrões muito diferentes, exigindo adaptações.

Por fim, importa destacar algumas limitações do presente estudo. Em primeiro lugar, o *dataset* de *posts* foi relativamente restrito, centrado apenas no mês de março de 2022 e numa área geográfica delimitada. Estudos que utilizaram séries temporais mais longas conseguiram captar variações sazonais e tendências de maior alcance. Em segundo lugar, a forma como os utilizadores portugueses do Twitter se expressam pode não refletir diretamente preocupações ambientais: o fenómeno foi visível no quotidiano (céu alaranjado, pó acumulado), mas não necessariamente associado a impactos de saúde ou poluição nas publicações online. Além disso, ferramentas mais avançadas, como APIs

dedicadas ou modelos de mineração de texto em larga escala, poderiam permitir recolhas mais amplas e diversificadas.

Em síntese, a discussão mostra que, enquanto os dados ambientais confirmam inequivocamente o impacto das poeiras do Saara na qualidade do ar, os dados sociais não revelam uma correlação clara com esse fenómeno, respondendo assim parcialmente às Questões de Investigação introduzidas no capítulo da análise dos trabalhos relacionados. Para QI1 é observado que as variações da qualidade do ar nem sempre se refletem nas emoções expressas nas redes sociais, como demonstrado pela fraca correlação encontrada. Para QI2, por sua vez, é evidenciado que as emoções predominantes durante o evento foram relativamente neutras, diferindo dos padrões negativos mais comuns em outros contextos. Finalmente, para QI3 é destacado as limitações do uso da análise emocional das redes sociais como indicador indireto da qualidade do ar, reforçadas pelas especificidades culturais e pelo volume reduzido de dados disponíveis. Apesar dessas limitações, este trabalho demonstra a viabilidade da integração entre dados ambientais e sociais, abrindo caminho para futuras investigações com maior cobertura temporal, múltiplas cidades e metodologias de análise de texto mais robustas.

7 Conclusão

Este estudo atingiu o seu principal objetivo: integrar dados ambientais e sociais para analisar a qualidade do ar em contexto urbano, tendo como foco o episódio concreto da chegada de poeiras do deserto do Saara que afetou Lisboa em março de 2022. A investigação confirmou de forma clara o impacto atmosférico do fenómeno, com níveis de PM10 muito acima dos valores de referência e efeitos visíveis no quotidiano da cidade. Ao cruzar esses dados com as publicações do Twitter, verificou-se que os sentimentos expressos online permaneceram, em grande parte, neutros ou ligeiramente positivos, mesmo nos momentos de maior concentração de partículas. Esta ausência de correlação significativa mostra que fenómenos ambientais intensos não se traduzem necessariamente em reações nas redes sociais, sobretudo no contexto português, onde fatores culturais, sociais podem ter condicionado a expressão pública. Além disso, os resultados sugerem que outros temas de natureza política, social ou até desportiva podem exercer maior influência sobre os sentimentos manifestados online do que questões ambientais, pondo estas para segundo plano na atenção digital.

Uma das lições que se pode retirar deste trabalho é a evidência de uma desconexão entre a realidade física e a realidade digital na nossa cultura. Enquanto os sensores registaram um impacto ambiental severo, a comunidade online mostrou-se pouco responsiva. Esta constatação reforça a necessidade de melhorar a metodologia adotada, mas mostra o seu potencial e como pode ser expandida em contextos mais amplos de tempo, espaço e diversidade de plataformas.

O contributo desta dissertação reside, assim, em trazer evidência experimental nacional a uma área ainda pouco explorada em Portugal, demonstrando a viabilidade de integrar dados ambientais, meteorológicos e sociais. Apesar de não ter havido respostas conclusivas, o estudo abre espaço para novas perguntas: como evolui esta relação ao longo de períodos mais extensos? Que diferenças se observam entre cidades ou plataformas digitais distintas? E de que forma metodologias de análise de linguagem mais avançadas podem captar detalhes ou variações que escaparam nesta abordagem?

Em última análise, este trabalho mostra que compreender a relação entre ambiente e perceção pública é um desafio complexo, mas também uma oportunidade para construir olhares mais completos e cruzando abordagens de várias áreas de estudo sobre problemas ambientais, reforçando assim a utilidade da metodologia e o potencial para expandir a análise.

8 Bibliografia

[1] D. Moher *et al.*, "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement," *PLoS Med*, vol. 6, no. 7, p. e1000097, Jul. 2009, doi: 10.1371/JOURNAL.PMED.1000097.

[2] Y. Tao, F. Zhang, C. Shi, and Y. Chen, "Social media data-based sentiment analysis of tourists' air quality perceptions," *Sustainability (Switzerland)*, vol. 11, no. 18, p. 5070, Sep. 2019, doi: 10.3390/su11185070.

[3] S. Shan, X. Ju, Y. Wei, and Z. Wang, "Effects of pm2.5 on people's emotion: A case study of weibo (chinese twitter) in beijing," *Int J Environ Res Public Health*, vol. 18, no. 10, May 2021, doi: 10.3390/IJERPH18105422.

[4] Y. E. García *et al.*, "Wildfires and social media discourse: exploring mental health and emotional wellbeing through Twitter," *Front Public Health*, vol. 12, 2024, doi: 10.3389/FPUBH.2024.1349609.

[5] B. Ye, P. Krishnan, and S. Jia, "Public Concern about Air Pollution and Related Health Outcomes on Social Media in China: An Analysis of Data from Sina Weibo (Chinese Twitter) and Air Monitoring Stations," *Int J Environ Res Public Health*, vol. 19, no. 23, Dec. 2022, doi: 10.3390/IJERPH192316115.

[6] B. Wang, N. Wang, and Z. Chen, "Research on air quality forecast based on web text sentiment analysis," *Ecol Inform*, vol. 64, Sep. 2021, doi: 10.1016/J.ECOINF.2021.101354.

[7] W. Zhai and C. Cheng, "A long short-term memory approach to predicting air quality based on social media data," *Atmos Environ*, vol. 237, Sep. 2020, doi: 10.1016/J.ATMOSENV.2020.117411.

[8] J. Min, H. Yunxiu, S. Yong, J. Fengxiang, and L. Ting, "Research on Analysis of Evaluation Influence Factors of Air Quality Opinions Sentiment Value and Quantification Method," *Water Air Soil Pollut*, vol. 232, no. 7, p. 256, Jul. 2021, doi: 10.1007/s11270-021-05197-x.

[9] H. Ji *et al.*, "Research on adaption to air pollution in Chinese cities: Evidence from social media-based health sensing," *Environ Res*, vol. 210, p. 112762, Jul. 2022, doi: 10.1016/j.envres.2022.112762.

[10] S. Zheng, J. Wang, C. Sun, X. Zhang, and M. E. Kahn, "Air pollution lowers Chinese urbanites' expressed happiness on social media," *Nat Hum Behav*, vol. 3, no. 3, pp. 237–243, Mar. 2019, doi: 10.1038/s41562-018-0521-2.

[11] J. M. Postma *et al.*, "Assessing community response to wildfire smoke: A multimethod study using social media," *Public Health Nurs*, vol. 40, no. 1, pp. 153–162, Feb. 2023, doi: 10.1111/phn.13140.

[12] G. Yang, Y. Ju, and W. Ni, "Does the air pollution level information matter in public perception? Insights from China," *J Environ Manage*, vol. 349, p. 119582, Jan. 2024, doi: 10.1016/J.JENVMAN.2023.119582.

[13] D. Amangeldi, A. Usmanova, and P. Shamoi, "Understanding Environmental Posts: Sentiment and Emotion Analysis of Social Media Data," *IEEE Access*, vol. 12, pp. 33504–33523, Jan. 2024, doi: 10.1109/ACCESS.2024.3371585.

[14] C. E. Slavik, D. A. Chapman, A. S. Cohen, N. Bendefaa, and E. Peters, “Clearing the air: evaluating institutions’ social media health messaging on wildfire and smoke risks in the US Pacific Northwest,” *BMC Public Health*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12889-024-17907-1.

[15] G. Marques, I. M. Pires, N. Miranda, and R. Pitarma, “Air quality monitoring using assistive robots for ambient assisted living and enhanced living environments through internet of things,” *Electronics (Switzerland)*, vol. 8, no. 12, p. 1375, Dec. 2019, doi: 10.3390/electronics8121375.

[16] M. T. Marfori *et al.*, “Public Health Messaging During Extreme Smoke Events: Are We Hitting the Mark?,” *Front Public Health*, vol. 8, Sep. 2020, doi: 10.3389/FPUBH.2020.00465.

[17] W. Hong, Y. Wei, and S. Wang, “Left behind in perception of air pollution? A hidden form of spatial injustice in China,” *Environment and Planning C: Politics and Space*, vol. 40, no. 3, pp. 666–684, May 2022, doi: 10.1177/23996544211036145/ASSET/FDB83BA2-7C31-4935-88EF-99653086DD9C/ASSETS/IMAGES/LARGE/10.1177_23996544211036145-FIG1.JPG.

[18] L. Yan, F. Duarte, De Wang, S. Zheng, and C. Ratti, “Exploring the effect of air pollution on social activity in China using geotagged social media check-in data,” *Cities*, vol. 91, pp. 116–125, Aug. 2019, doi: 10.1016/J.CITIES.2018.11.011.

[19] C. Zhang and G. Zhang, “How Does Air Pollution Impact Residence Intention of Rural Migrants? Empirical Evidence from the CMDS,” *Sustainability 2024, Vol. 16, Page 5784*, vol. 16, no. 13, p. 5784, Jul. 2024, doi: 10.3390/SU16135784.

[20] P. Brimblecombe and Y. Lai, “Effect of fireworks, Chinese new year and the COVID-19 lockdown on air pollution and public attitudes,” *Aerosol Air Qual Res*, vol. 20, no. 11, pp. 2318–2331, Nov. 2020, doi: 10.4209/AAQR.2020.06.0299.

[21] M. Burke *et al.*, “Exposures and behavioural responses to wildfire smoke,” *Nat Hum Behav*, vol. 6, no. 10, pp. 1351–1361, Oct. 2022, doi: 10.1038/S41562-022-01396-6.

[22] H. O’Leary, S. Parr, and M. M. H. El-Sayed, “The breathing human infrastructure: Integrating air quality, traffic, and social media indicators,” *Science of the Total Environment*, vol. 827, Jun. 2022, doi: 10.1016/J.SCITOTENV.2022.154209.

[23] Y. Sun, F. Jin, Y. Zheng, M. Ji, and H. Wang, “A new indicator to assess public perception of air pollution based on complaint data,” *Applied Sciences (Switzerland)*, vol. 11, no. 4, pp. 1–18, Feb. 2021, doi: 10.3390/APP11041894.

[24] M. Ashayeri, “Decoding Global Indoor Health Perception on Social Media Through NLP and Transformer Deep Learning,” in *Artificial Intelligence in Performance-Driven Design: Theories, Methods, and Tools*, wiley, 2024, pp. 159–185. doi: 10.1002/9781394172092.ch8.

[25] T. Carpi, A. Hino, S. M. Iacus, and G. Porro, “The Impact of COVID-19 on Subjective Well-Being: Evidence from Twitter Data,” *Journal of Data Science*, vol. 21, no. 4, pp. 761–780, Oct. 2023, doi: 10.6339/22-JDS1066.

[26] A. Madjar, I. Gjorshoska, J. Prodanova, A. Dedinec, and L. Kocarev, “Western Balkan societies’ awareness of air pollution. Estimations using natural language processing

techniques," *Ecol Inform*, vol. 75, p. 102097, Jul. 2023, doi: 10.1016/j.ecoinf.2023.102097.

[27] R. Yu, C. Zeng, M. Chang, C. Bao, M. Tang, and F. Xiong, "Effects of Urban Vibrancy on an Urban Eco-Environment: Case Study on Wuhan City," *Int J Environ Res Public Health*, vol. 19, no. 6, p. 3200, Mar. 2022, doi: 10.3390/ijerph19063200.

[28] S. Gurajala, S. Dhaniyala, and J. N. Matthews, "Understanding Public Response to Air Quality Using Tweet Analysis," *Social Media and Society*, vol. 5, no. 3, Jul. 2019, doi: 10.1177/2056305119867656.

[29] Y. Yang, K. Y. Goh, H. H. Teo, and S. S. L. Tan, "The Impact of Air Pollution Information on Individuals' Exercise Behavior: Empirical Study Using Wearable and Mobile Devices Data," *JMIR Mhealth Uhealth*, vol. 12, p. e55207, Jan. 2024, doi: 10.2196/55207.

[30] H. O'Leary, D. Smiles, S. Parr, and M. M. H. El-Sayed, "'I Can't Breathe:' The Invisible Slow Violence of Breathing Politics in Minneapolis," *Soc Nat Resour*, vol. 36, no. 9, pp. 1098–1118, 2023, doi: 10.1080/08941920.2023.2194068.

[31] Y. Tao, W. Liu, Z. Huang, and C. Shi, "Thematic analysis of reviews on the air quality of tourist destinations from a sentiment analysis perspective," *Tour Manag Perspect*, vol. 42, Apr. 2022, doi: 10.1016/J.TMP.2022.100969.

[32] H. K. Al-Shidi, A. K. Ambusaidi, and H. Sulaiman, "Public awareness, perceptions and attitudes on air pollution and its health effects in Muscat, Oman," *J Air Waste Manage Assoc*, vol. 71, no. 9, pp. 1159–1174, Jan. 2021, doi: 10.1080/10962247.2021.1930287.

[33] M. von Szombathely, B. Bechtel, B. Lemke, J. Oßenbrügge, T. Pohl, and M. Pott, "Empirical evidences for urban influences on public health in Hamburg," *Applied Sciences (Switzerland)*, vol. 9, no. 11, p. 2303, Jun. 2019, doi: 10.3390/app9112303.

[34] Z. H. Wang *et al.*, "Environmentally vulnerable or sensitive groups exhibiting varying concerns toward air pollution can drive government response to improve air quality," *iScience*, vol. 25, no. 6, p. 104460, Jun. 2022, doi: 10.1016/j.isci.2022.104460.

[35] K. Wadhwa, D. Mehra, H. Gosain, and A. K. Haritash, "Sentiment Analysis of Air Quality Perception in Major Metro Cities of India," *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*, p. Delhi, doi: 10.1109/ICCCNT56998.2023.10307448.

[36] M. L. Loureiro, M. Alló, and P. Coello, "Hot in Twitter: Assessing the emotional impacts of wildfires with sentiment analysis," *Ecological Economics*, vol. 200, p. 107502, Oct. 2022, doi: 10.1016/j.ecolecon.2022.107502.

[37] Y. Zhu *et al.*, "Quantifying Spatiotemporal Heterogeneities in PM2.5-Related Health and Associated Determinants Using Geospatial Big Data: A Case Study in Beijing," *Remote Sens (Basel)*, vol. 14, no. 16, p. 4012, Aug. 2022, doi: 10.3390/rs14164012.

[38] "QualAr." Accessed: Jun. 18, 2025. [Online]. Available: https://qualar.apambiente.pt/

[39] "Datos meteorológicos de SYNOPS/BUFR - Predicciones GFS/ECMWF - Meteomanz.com." Accessed: Jun. 18, 2025. [Online]. Available: http://www.meteomanz.com/

[40] "Official Microsoft Power Automate documentation - Power Automate | Microsoft Learn." Accessed: Jul. 02, 2025. [Online]. Available: https://learn.microsoft.com/en-us/power-automate/

[41] "What is LoRaWAN® - LoRa Alliance®." Accessed: Jun. 26, 2025. [Online]. Available: https://lora-alliance.org/resource_hub/what-is-lorawan/?utm_source=chatgpt.com

[42] "Quick Start | The Things Network." Accessed: Jun. 26, 2025. [Online]. Available: https://www.thethingsnetwork.org/docs/applications/nodered/quick-start/?utm_source=chatgpt.com

[43] "Documentation : Node-RED." Accessed: Jun. 26, 2025. [Online]. Available: https://nodered.org/docs/

[44] "Eclipse Mosquitto." Accessed: Jun. 26, 2025. [Online]. Available: https://www.mosquitto.org/?utm_source=chatgpt.com

[45] R. R. Buchholz *et al.*, "Air pollution trends measured from Terra: CO and AOD over industrial, fire-prone, and background regions," *Remote Sens Environ*, vol. 256, Apr. 2021, doi: 10.1016/j.rse.2020.112275.

[46] "Proposed Residential Indoor Air Quality Guidelines for Benzene - Canada.ca." Accessed: Jun. 25, 2025. [Online]. Available: https://www.canada.ca/en/health-canada/programs/consultation-proposed-residential-indoor-air-quality-guidelines-benzene/document.html?utm_source=chatgpt.com

[47] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," 2010. Accessed: Jun. 18, 2025. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf

[48] F. Å. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," *CEUR Workshop Proc*, vol. 718, pp. 93–98, Mar. 2011, Accessed: Jun. 18, 2025. [Online]. Available: https://arxiv.org/pdf/1103.2903

[49] "The Case for Using the General Linear Model as a Unifying Conceptual Framework for Teaching Statistics and Psychometric Theory," *The Case for Using the General Linear Model as a Unifying Conceptual Framework for Teaching Statistics and Psychometric Theory*, 2010, doi: 10.2458/AZU_JMMSS_V3I2_MEHI.

[50] "TextBlob: Simplified Text Processing — TextBlob 0.19.0 documentation." Accessed: Aug. 12, 2025. [Online]. Available: https://textblob.readthedocs.io/en/dev/

[51] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, May 2014, doi: 10.1609/ICWSM.V8I1.14550.

[52] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008, doi: 10.1561/1500000011.

[53] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: Jun. 18, 2025. [Online]. Available: https://arxiv.org/pdf/1810.04805

[54] J. J. Berman, "Understanding Your Data," *Data Simplification*, pp. 135–187, 2016, doi: 10.1016/B978-0-12-803781-2.00004-7.

[55] M. M. Mukaka, "A guide to appropriate use of Correlation coefficient in medical research," *Malawi Med J*, vol. 24, no. 3, p. 69, 2012, Accessed: Sep. 16, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC3576830/

[56] Y. Y. Wang, W. Yang, X. Y. Wang, S. Wang, J. F. Bai, and Y. Cheng, "[Characteristics of Ozone Pollution and Influencing Factors in Urban and Suburban Areas in Zibo]," *Huan Jing Ke Xue*, vol. 43, no. 1, pp. 170–179, Jan. 2022, doi: 10.13227/J.HJKX.202105009.