Classificação de Denúncias Ambientais no X: Criação de Dataset e Avaliação de Features

Guilherme Lopes[1], **Sérgio L. S. Mergen**[2]

[1]Centro de Tecnologia – Universidade Federal de Santa Maria (UFSM)
97105-900 -– Santa Maria -– RS — Brasil

[2]Departamento de Linguagens e Sistemas de Computação
Universidade Federal de Santa Maria (UFSM)

guilherme.lopes@acad.ufsm.br, mergen@inf.ufsm.br

Abstract. *This article presents the creation of a dataset aimed at classifying tweets related to environmental complaints. Data collection was carried out through the X API, using carefully selected keywords to compose three distinct sets that reflect different levels of task complexity. In addition, semantic and contextual features were defined to complement the raw text of the tweets. Experiments with Support Vector Classification (LinearSVC) showed that the combination of text and features improves classifier performance compared to their isolated use. Even though the dataset is relatively small, the results indicate its usefulness and open avenues for future work, such as expanding the corpus, automating feature extraction, and conducting feature selection analyses.*

Resumo. *Este artigo apresenta a criação de um dataset voltado à classificação de tuítes relacionados a denúncias ambientais. A coleta foi realizada por meio da API do X, utilizando keywords selecionadas criteriosamente para compor três conjuntos distintos, que refletem diferentes níveis de complexidade da tarefa. Além disso, foram definidas features semânticas e contextuais que complementam o texto bruto dos tuítes. Os experimentos com Support Vector Classification (LinearSVC) mostraram que a combinação de texto e features melhora o desempenho dos classificadores em comparação ao uso isolado. Mesmo sendo um dataset pequeno, os resultados indicam sua utilidade e abrem caminho para trabalhos futuros, como a ampliação do corpus, a automatização da extração de features e análises de seleção de atributos.*

1. Introdução

A emergência climática tem se consolidado como um dos maiores desafios contemporâneos, trazendo consigo impactos diretos na vida urbana e rural. Eventos extremos como enchentes, secas prolongadas, queimadas e deslizamentos de terra tornaram-se cada vez mais frequentes, exigindo respostas rápidas e eficazes por parte das autoridades. Esses problemas não apenas comprometem a infraestrutura das cidades, mas também afetam a saúde pública, a segurança da população e a preservação dos ecossistemas.

Nesse contexto, as redes sociais emergem como espaços centrais para a circulação de informações relacionadas a crises ambientais. Plataformas como o X (antigo Twitter) permitem que cidadãos relatem, em tempo real, situações de risco ou degradação ambiental. Esses relatos espontâneos podem servir como fonte relevante para pesquisadores,

jornalistas e órgãos de fiscalização, oferecendo dados valiosos para o monitoramento e a tomada de decisão. No entanto, transformar esse fluxo de mensagens informais em informação estruturada é uma tarefa complexa.

A identificação automática de tuítes de denúncia enfrenta obstáculos significativos. O texto curto, a informalidade da linguagem, o uso de regionalismos e a presença de ambiguidades dificultam a tarefa de classificação. Além disso, muitas mensagens compartilham vocabulário semelhante ao das denúncias sem configurarem um relato ambiental, o que aumenta a complexidade da triagem. Nesse cenário, técnicas de Processamento de Linguagem Natural (PLN) e aprendizado de máquina tornam-se fundamentais para estruturar dados não estruturados e apoiar sistemas de monitoramento ambiental.

Este trabalho apresenta duas contribuições principais. A primeira consiste na criação de um *dataset* cuidadosamente estruturado, dividido em três conjuntos: (i) tuítes de denúncia ambiental, (ii) tuítes que compartilham palavras-chave semelhantes, mas não configuram denúncias, e (iii) tuítes de assuntos aleatórios. Essa diversidade foi planejada para introduzir diferentes níveis de complexidade no processo de classificação, permitindo que pesquisadores testem a robustez de seus modelos em cenários mais desafiadores. A coleta foi realizada por meio da API do Twitter, utilizando *keywords* selecionadas criteriosamente para garantir representatividade e diversidade nos dados.

A segunda contribuição está relacionada à engenharia de *features*. Para cada tuíte do *dataset*, foram extraídas características adicionais que complementam o texto original e podem auxiliar na tarefa de classificação. Essas *features* incluem dimensões semânticas e contextuais, como geolocalização, tipificação do dano e conectividade institucional. A análise experimental demonstra que o uso combinado de texto e *features* resulta em melhor desempenho, evidenciando o valor da abordagem proposta.

Ao unir a disponibilização de um *dataset* diversificado com a exploração de *features* específicas, este trabalho busca avançar na compreensão dos desafios envolvidos na classificação de denúncias ambientais em redes sociais. A proposta oferece recursos e *insights* que podem ser aproveitados em pesquisas futuras, contribuindo para o desenvolvimento de sistemas mais eficazes de monitoramento ambiental e apoio à tomada de decisão.

Este trabalho está dividido da seguinte forma: a seção 2 apresenta os trabalhos relacionados. A seção 3 descreve o processo de criação do *dataset*. Já a seção 4 apresenta o processo de engenharia de *features*, que culminou com a definição de características representativas de denúncias ambientais. A seção 5 traz experimentos de classificação usando o dataset desenvolvido e as features concebidas. Por fim, a seção 6 traz as considerações finais.

2. Trabalhos Relacionados

A presente pesquisa insere-se em um conjunto de estudos que exploram o uso de dados provenientes de denúncias ambientais para apoiar processos de monitoramento e gestão. Nesse contexto, dialoga com a iniciativa de Dalci [Dalci 2025], que analisa registros da FEPAM no Rio Grande do Sul e evidencia a elevada demanda por fiscalização, com mais de 1.500 denúncias anuais. Enquanto o referido trabalho enfatiza aspectos administrativos e institucionais da gestão pública, este artigo concentra-se na dimensão tecnológica,

propondo a criação de um *dataset* especializado e avaliando mecanismos de automação para a classificação de denúncias.

Há também proximidade com pesquisas que aplicam técnicas de PLN em mensagens informais de cidadãos, como as de [Gusmão et al. 2021] e Peixoto [Peixoto 2021]. Esses estudos demonstram como algoritmos podem lidar com variações linguísticas e transformar relatos espontâneos em dados estruturados. A presente investigação expande essa lógica ao aplicá-la especificamente ao domínio ambiental, com foco na identificação e categorização de eventos críticos.

No que se refere à preparação dos dados, o projeto adota procedimentos metodológicos semelhantes aos descritos em [Paiva and Ebecken 2022], Almeida [Almeida 2018] e Preotiuc-Pietro et al. [Preoţiuc-Pietro et al. 2019], que tratam da construção de *datasets* a partir de redes sociais. Assim como nesses trabalhos, a proposta aqui apresentada busca capturar denúncias reais e informais, submetendo os registros a um processo de organização e limpeza, de modo a torná-los adequados para uso em modelos de aprendizado de máquina.

Em síntese, a literatura existente demonstra a relevância de integrar dados não estruturados à metodologias de classificação. O presente artigo contribui para esse campo ao propor um conjunto de atributos específicos e avaliar seu impacto na tarefa de classificação de denúncias ambientais. Ao estruturar relatos informais em dimensões técnicas, o sistema avança em direção a soluções mais eficazes para o monitoramento preventivo e para a gestão de eventos ambientais.

3. Criação do Dataset

A construção do dataset foi guiada pela necessidade de capturar diferentes níveis de complexidade na tarefa de classificação de tuítes relacionados a denúncias ambientais. Para isso, utilizamos a API do X como ferramenta de coleta, inserindo *keywords* de busca previamente selecionadas de forma criteriosa. Esse processo garantiu que os dados refletissem tanto situações de denúncia quanto mensagens não relacionadas, compondo um corpus diversificado e adequado para os experimentos.

Os registros foram organizados em três conjuntos distintos:

- **Conjunto A (Denúncias)**: composto por tuítes que efetivamente configuram denúncias ambientais.
- **Conjunto B (Não-denúncias com *keywords* de denúncia)**: formado por tuítes que compartilham palavras-chave típicas de denúncias, mas cujo conteúdo não corresponde a uma denúncia.
- **Conjunto C (Assuntos aleatórios)**: inclui tuítes de temas diversos, coletados a partir de *keywords* não relacionadas ao contexto ambiental.

A diferença fundamental entre os conjuntos está na forma como as *keywords* foram utilizadas. Para os conjuntos A e B, empregamos palavras-chave associadas a denúncias ambientais, como termos relacionados a poluição, desmatamento ou crimes ambientais. Já para o conjunto C, utilizamos *keywords* aleatórias, sem relação com o tema ambiental, de modo a garantir diversidade e introduzir um nível adicional de contraste. A Tabela 1 apresenta as *keywords* usadas em cada caso.

Tabela 1. Keywords utilizadas na criação dos conjuntos A, B e C

Conjunto	Keywords
A e B	Desmatamento, Deslizamento, Lixo, Queimadas, Fogo, Mata/-Mata ciliar, Alagamento,Cheias, Entulho, Descarte irregular, água, poluição,Rios,Aterros, Crime ambiental, Chuvas
C	**sem keywords específicas**

O processo de construção do dataset seguiu um fluxo sistemático, ilustrado na Figura 1. Inicialmente, foram extraídos tuítes a partir das *keywords* de busca. Em seguida, cada tuíte passou por uma etapa de análise manual, na qual se inferiu se o conteúdo correspondia a uma denúncia ambiental ou não.

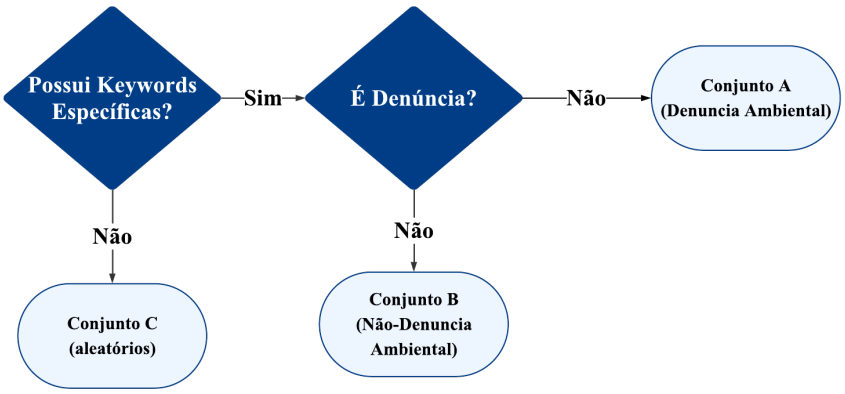

Figura 1. Fluxograma de classificação dos tuítes

Caso o tuíte fosse identificado como denúncia, era alocado no **Conjunto A**. Caso não fosse denúncia, mas tivesse sido coletado por meio de *keywords* de denúncia, era direcionado ao **Conjunto B**. Por fim, tuítes coletados a partir de *keywords* aleatórias eram incluídos no **Conjunto C**.

Esse processo assegura que o *dataset* não apenas contenha exemplos claros de denúncias, mas também casos ambíguos e irrelevantes, criando um recurso valioso para avaliar a robustez de diferentes classificadores e explorar o papel das *features* no enriquecimento da análise.

Os tuítes classificados como denúncias passaram por uma etapa adicional de **enriquecimento semântico**, na qual foi atribuída manualmente a natureza do evento relatado. Essa categorização permite estruturar melhor os dados e direcionar a denúncia para o órgão competente. Os tipos utilizados são *Saneamento Básico, queimadas*, *corpos d'água*, *acúmulo material,escoamento, deslizamento, poluição do ar* e *desmatamento*.

Por fim, os registros foram organizados em um formato padronizado, no qual cada linha representa uma unidade de análise (o *tweet*) composta por quatro dimensões técnicas essenciais:

- o **ID do *tweet***, que garante rastreabilidade e integridade da fonte;
- o **conteúdo**, preservando a mensagem bruta com seus regionalismos e variações gramaticais;
- a **variável de relevância**, que indica se o tuíte é denúncia ou não;
- a **natureza do evento**, composta pelas *tags* extraídas pela etapa de enriquecimento semântico (para tuítes que não sejam denúncias ambientais, essa dimensão é deixada sem preenchimento).

A adoção desse formato estruturado é fundamental por três razões principais. Primeiro, garante **consistência** na representação dos dados, permitindo que diferentes pesquisadores utilizem o corpus sem ambiguidades. Segundo, facilita a **reprodutibilidade** dos experimentos, já que cada tuíte pode ser rastreado e reavaliado em análises futuras. Terceiro, possibilita a **integração** com técnicas de aprendizado de máquina, que dependem de entradas organizadas em vetores de atributos. Dessa forma, mesmo sendo um *dataset* de tamanho reduzido, sua estruturação assegura maior valor científico e abre caminho para expansões futuras, nas quais novos registros poderão ser incorporados sem comprometer a coerência metodológica estabelecida nesta versão inicial.

Ao final do processo de coleta e organização, cada conjunto foi composto por 70 tuítes, totalizando 210 registros distribuídos entre denúncias ambientais (Conjunto A), não-denúncias com vocabulário semelhante (Conjunto B) e mensagens aleatórias (Conjunto C). O conjunto completo está disponível em `https://docs.google.com/spreadsheets/d/1yWmcLxQrcoIaYHJpgVOEsSWkwzMBSsF__KsiDjCX-24/edit?usp=sharing`.

O número reduzido de exemplos decorre da opção metodológica por uma análise completamente manual, na qual cada tuíte foi inspecionado individualmente para garantir sua correta classificação. Embora essa escolha limite naturalmente o tamanho do corpus, ela apresenta uma vantagem significativa: a intervenção humana na etapa de composição assegura maior riqueza semântica e precisão na definição das classes. Dessa forma, o dataset resultante não apenas oferece material para experimentos iniciais de classificação, mas também constitui uma base confiável e qualitativamente robusta para estudos futuros, nos quais a ampliação do volume de dados poderá ser realizada sem comprometer a consistência conceitual estabelecida nesta primeira versão.

4. Engenharia de Features

Além da disponibilização do *dataset*, este trabalho contribui com a definição de um conjunto de *features* semânticas e contextuais que complementam o texto dos tuítes. A construção desses atributos foi orientada por uma análise exploratória do corpus de 70 tuítes de denúncia coletados, na qual se identificaram os termos, referências e entidades mais recorrentes.

Os elementos linguísticos e semânticos mais frequentes identificados são:

- **Localidade**: menções a endereços específicos, como ruas, bairros ou pontos de referência.
- **Nome local**: referências relativas a locais sem especificidade, como “aqui” ou “onde eu moro”.
- **Instituição**: entidades nomeadas, como órgãos públicos, empresas ou organizações.

- **Descarte irregular, saneamento básico, queimadas, acúmulo de material, corpos d'água**: situações ambientais concretas relatadas pelos usuários.
- **Relato**: formato textual caracterizado por narrativas pessoais ou descrições em primeira pessoa.

A Tabela 2 apresenta a frequência de cada termo identificado. Os dados evidenciam que os termos mais recorrentes nos tuítes analisados estão relacionados à localidade(54 ocorrências) e ao **nome local** (45 ocorrências). Esse resultado sugere que os usuários tendem a contextualizar suas denúncias em relação ao espaço físico, seja por meio de endereços específicos ou de referências mais subjetivas, como "aqui" ou "onde eu moro". Tal padrão reforça a importância da dimensão espacial para a interpretação das denúncias, uma vez que a localização é um elemento essencial para a atuação de órgãos de fiscalização e resposta.

Tabela 2. Termos mais frequentes identificados nos tuítes de denúncia

Termo	Ocorrências
Localidade	54
Nome local	45
Descarte irregular	19
Instituição	18
Saneamento básico	10
Consequências	10
Queimadas	9
Acúmulo de material	8
Relato	8
Corpos d'água	6

Em seguida, aparecem termos ligados a situações ambientais concretas, como **descarte irregular** (19), **Saneamento Básico** (10), **queimadas** (9), **acúmulo de material** (8) e **corpos d'água** (6). A presença desses elementos indica que os cidadãos não apenas relatam problemas, mas também os tipificam, atribuindo categorias específicas ao dano ambiental. Esse comportamento sugere uma consciência ambiental crescente e fornece subsídios para a criação de atributos categóricos que diferenciem tipos de impacto.

A categoria **instituição** (18 ocorrências) mostra que parte dos tuítes faz referência a entidades nomeadas, como órgãos públicos ou empresas. Esse dado é relevante porque aponta para a atribuição de responsabilidade, seja na forma de cobrança de ação do poder público ou de denúncia de agentes privados. Já o **relato** (8 ocorrências) evidencia a dimensão narrativa, em que o usuário descreve a situação em primeira pessoa, reforçando a credibilidade e a urgência da denúncia.

O termo **consequências** (10 ocorrências) merece destaque adicional, pois indica que alguns usuários não apenas descrevem o problema, mas também mencionam seus efeitos diretos, como prejuízos materiais, riscos à saúde ou impactos sociais. Essa dimensão amplia o valor informativo das denúncias, permitindo que os modelos de classificação considerem não apenas a ocorrência do dano, mas também sua gravidade percebida.

Por fim, a baixa frequência de termos como **saneamento básico** e **corpos d'água** sugere que, embora presentes, esses tópicos aparecem de forma mais pontual. Isso pode estar relacionado à especificidade dos problemas ou à menor visibilidade desses temas no cotidiano urbano.

A partir desses elementos, foram definidas as seguintes *features*:

- **Geolocalização Contextual (Localidade)**: verifica a presença de marcadores geográficos, como ruas, bairros, pontos de referência ou coordenadas. É tratada como uma variável booleana.
- **Natureza Narrativa**: diferencia se o usuário está vivenciando o fato (relato em primeira pessoa), compartilhando uma notícia de terceiros ou apenas emitindo uma opinião. É tratada como uma variável categórica, com três valores possíveis: Relato, Notícia ou Opinião.
- **Tipificação do Dano**: classifica a categoria do impacto ambiental descrito, permitindo o encaminhamento automático para órgãos competentes. É tratada como uma variável categórica, com valores como descarte irregular, obstrução de rede pluvial e poluição hídrica.
- **Atribuição de Responsabilidade (Agente)**: identifica se o tuíte aponta o causador do dano, como uma empresa específica ou um descarte clandestino. Esse atributo é essencial para processos de fiscalização e autuação. É tratada como uma variável booleana.
- **Conectividade Institucional**: mapeia o uso de marcações de perfis oficiais (@prefpoa, @fepam, @defesacivilrs) ou hashtags de mobilização. Esse indicador reflete a expectativa do cidadão por uma resposta imediata do poder público. É tratada como uma variável booleana.

Por fim, é importante destacar que o processo de extração dessas *features* foi realizado de forma inteiramente manual. Nosso objetivo neste momento não é propor um sistema automatizado de extração, mas sim avaliar o impacto que essas características podem ter na tarefa de classificação. Dessa forma, buscamos verificar se compensa investir esforços futuros em mecanismos de extração automática dessas informações.

5. Experimentos

Para avaliar o impacto do *dataset* e das *features* propostas, realizamos experimentos utilizando o algoritmo Support Vector Classification (LinearSVC) [Vapnik 1997], implementada na biblioteca *scikit-learn* [Pedregosa et al. 2011]. Trata-se de um algoritmo supervisionado de aprendizado de máquina, no qual cada objeto (tuíte, no nosso caso) é representado por um vetor de atributos — sejam eles derivados do texto ou das *features* semânticas e contextuais — e classificado em categorias previamente definidas(denúncia ou não denúncia). O modelo busca encontrar um hiperplano que maximize a separação entre as classes, garantindo maior capacidade de generalização.

Os experimentos foram conduzidos em cenários distintos, explorando diferentes combinações de conjuntos de dados e estratégias de representação. Para assegurar a confiabilidade dos resultados, os dados foram balanceados de modo que cada classe tivesse o mesmo número de exemplos, evitando vieses decorrentes de distribuições desiguais. Além disso, os testes seguiram uma lógica de validação cruzada, permitindo avaliar o

desempenho do modelo em subconjuntos diferentes do corpus e reduzir o risco de sobreajuste.

A avaliação dos classificadores foi realizada por meio de métricas amplamente utilizadas em tarefas de classificação: **acurácia**, que mede a proporção de previsões corretas; **precisão**, que indica a proporção de denúncias corretamente identificadas entre todas as classificadas como denúncia; **recall**, que avalia a capacidade de recuperar todas as denúncias presentes no conjunto; e **F1-score**, que combina precisão e *recall* em uma média harmônica, oferecendo uma visão equilibrada do desempenho. Essas métricas são fundamentais para compreender não apenas a eficácia geral do modelo, mas também sua capacidade de lidar com casos ambíguos e evitar falsos negativos, que em cenários ambientais podem representar riscos relevantes.

No primeiro experimento, utilizamos o **Conjunto A** (denúncias ambientais) combinado com o **Conjunto C** (tuítes aleatórios). O objetivo foi avaliar o desempenho em um cenário menos complexo, onde os exemplos de não-denúncia são mais distintos dos de denúncia. Três estratégias foram comparadas:

1. Classificação baseada apenas no texto dos tuítes.
2. Classificação baseada apenas nas *features* extraídas manualmente.
3. Classificação baseada na combinação de texto e *features*.

Os resultados obtidos estão apresentados na Tabela 3.

Tabela 3. Resultados do Experimento 1 (Conjunto A + Conjunto C)

Estratégia	Acurácia	Precisão	Recall	F1-score
Texto apenas	0.92	1.0	0.83	0.91
Features apenas	1.0	1.0	1.0	1.0
Texto + Features	**1.0**	**1.0**	**1.0**	**1.0**

No segundo experimento, utilizamos o **Conjunto A** (denúncias ambientais) combinado com o **Conjunto B** (tuítes não-denúncia, mas com *keywords* de denúncia). Este cenário é mais desafiador, pois os exemplos de não-denúncia compartilham vocabulário semelhante ao das denúncias, aumentando a complexidade da tarefa de classificação. Novamente, foram comparadas as três estratégias:

1. Classificação baseada apenas no texto dos tuítes.
2. Classificação baseada apenas nas *features*.
3. Classificação baseada na combinação de texto e *features*.

Os resultados obtidos estão apresentados na Tabela 4.

Tabela 4. Resultados do Experimento 2 (Conjunto A + Conjunto B)

Estratégia	Acurácia	Precisão	Recall	F1-score
Texto apenas	0.83	0.80	0.89	0.84
Features apenas	0.83	0.80	0.89	0.84
Texto + Features	**0.86**	**0.84**	**0.89**	**0.86**

Os experimentos indicam dois pontos principais. Primeiro, o uso do **Conjunto B** leva a resultados consistentemente inferiores em comparação ao **Conjunto C**, confirmando que se trata de um subconjunto mais complexo e desafiador para os classificadores, devido à sobreposição de vocabulário entre denúncias e não-denúncias. Segundo, a estratégia de combinar **texto + features** apresenta desempenho superior em ambos os cenários, evidenciando que as *features* propostas agregam valor ao processo de classificação.

Em síntese, os resultados mostram que, mesmo em um *dataset* pequeno, a engenharia de *features* contribui para aumentar a robustez dos modelos, especialmente em contextos de maior ambiguidade linguística. Isso reforça a importância de investir na definição e automatização desses atributos em trabalhos futuros.

6. Considerações Finais

Este trabalho apresentou duas contribuições principais para a tarefa de classificação de tuítes relacionados a denúncias ambientais. A primeira foi a criação de um *dataset* estruturado em três conjuntos distintos, que introduzem diferentes níveis de complexidade e permitem avaliar a robustez dos classificadores. Embora o *dataset* seja relativamente pequeno, já se mostrou útil para experimentos iniciais e pode servir como recurso para outros pesquisadores interessados em explorar o tema. A segunda contribuição foi a definição e utilização de um conjunto de *features* semânticas e contextuais, extraídas manualmente, que complementam o texto dos tuítes. Os resultados experimentais demonstraram que a combinação de texto e *features* melhora significativamente o desempenho dos classificadores, evidenciando o valor dessas características no processo de classificação.

Mesmo em sua versão inicial, o *dataset* e as *features* propostas já demonstraram potencial para avançar a pesquisa na área. Como trabalhos futuros, pode-se ampliar o corpus de tuítes, de modo a enriquecer a diversidade e a representatividade dos dados. Também é possível automatizar o processo de extração das *features*, já que se observou que elas são importantes para a classificação. Além disso, análises voltadas à seleção de atributos podem relativizar a importância de cada *feature*, indicando quais são mais críticas para o desempenho dos modelos. Outro caminho promissor é utilizar as próprias *features* como filtros preliminares na criação do *dataset*, funcionando como um crivo antes da etapa manual em que um especialista decide se o tuíte é ou não uma denúncia.

Em síntese, este trabalho abre espaço para novas investigações, mostrando que mesmo um *dataset* pequeno, aliado a um conjunto de *features* cuidadosamente definidas, pode oferecer resultados relevantes e servir de base para o desenvolvimento de soluções mais eficazes no monitoramento de denúncias ambientais em redes sociais.

Referências

Almeida, T. L. M. d. (2018). Estudo sobre aplicação de aprendizado de máquina para identificação de assaltos através de informações do twitter.

Calheiros, A. J. P. and Pizzigatti, P. L. (2025). Ferramenta de descoberta e exploração de dados ambientais com respostas a prompts utilizando modelos de linguagens avançados.

Dalci, A. A. L. (2025). Denambrs: Sistema de categorização fmea de denúncias para identificação de fragilidades ambientais no rs. Master's thesis, Universidade Estadual do Rio Grande do Sul.

Gusmão, C., Figueiredo, K., and Brito, W. (2021). Técnicas de processamento de linguagem natural em denúncias criminais: Automatização e classificação de texto em português coloquial. In *Anais do XLVIII Seminário Integrado de Software e Hardware*, pages 172–182, Porto Alegre, RS, Brasil. SBC.

Khurana, U., Turaga, D., Samulowitz, H., and Parthasrathy, S. (2016). Cognito: Automated feature engineering for supervised learning. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 1304–1307.

Manzoor, S. I., Singla, J., and Nikita (2019). Fake news detection using machine learning approaches: A systematic review. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 230–234.

Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E., and Turaga, S. D. (2017). Learning feature engineering for classification. pages 2529–2535.

Paiva, E. and Ebecken, N. (2022). Ferramenta para classificação de denúncias: Uma abordagem baseada em textos e dados estruturados. In *Anais Estendidos do XVIII Simpósio Brasileiro de Sistemas de Informação*, pages 83–90, Porto Alegre, RS, Brasil. SBC.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Peixoto, L. H. R. (2021). *Aprendizado de Máquina Aplicado no Atendimento de Reclamações de Clientes*. PhD thesis, Universidade de São Paulo.

Preoţiuc-Pietro, D., Gaman, M., and Aletras, N. (2019). Automatically identifying complaints in social media. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5019, Florence, Italy. Association for Computational Linguistics.

Rizzotto, M. L. F., Costa, A. M., and Lobato, L. d. V. d. C. (2024). Crise climática e os novos desafios para os sistemas de saúde: o caso das enchentes no rio grande do sul/brasil. *Saúde em Debate*, 48(141):e141ED.

Vapnik, V. N. (1997). The support vector method. In *International conference on artificial neural networks*, pages 261–271. Springer.