Automating the Classification of Crime Reports in Altamira, Pará using BERT-based Architectures

Gabryel Silva[1], **Hugo Kuribayashi**[1,2], **Reginaldo Santos**[3], **Adam Santos**[1,2]

[1]Faculdade de Sistemas de Informação
Universidade Federal do Sul e Sudeste do Pará (UNIFESSPA)
Marabá - PA - Brasil

[2]Programa de Pós-Graduação em Ciências Forenses
Universidade Federal do Sul e Sudeste do Pará (UNIFESSPA)
Marabá - PA - Brasil

[3]Faculdade de Computação - Universidade Federal do Pará (UFPA)
Belém - PA - Brasil

{gabryelmarcos, hugo, adamdreyton}@unifesspa.edu.br, regicsf@ufpa.br

Abstract. *With the increasing volume of data generated in public safety, there is an urgent need to automate internal procedures to ensure a faster and more effective response. This paper presents the application of algorithms based on the fine-tuned Bidirectional Encoder Representations from Transformers (BERT) architecture to classify crimes in police report narratives from Altamira City, Pará. The results showed that BERTimbau achieved between 88% and 90% accuracy in the tests conducted, indicating significant potential to optimize the consolidation of police reports and make a substantial contribution to public safety.*

Resumo. *Com o crescente volume de dados gerados na segurança pública, surge uma necessidade urgente de automatizar os procedimentos internos para garantir uma resposta mais rápida e eficaz. Este artigo apresenta a aplicação de algoritmos baseados no ajuste fino da arquitetura Bidirectional Encoder Representations from Transformers (BERT) para classificar crimes em relatos de boletins de ocorrência da Cidade de Altamira, Pará. Os resultados obtidos demonstraram que o BERTimbau alcançou acurácias entre 88% e 90% nos testes realizados, indicando um grande potencial para otimizar a consolidação dos boletins de ocorrência e contribuir significativamente para a segurança pública.*

1. Introdução

A segurança pública no Brasil é uma responsabilidade governamental, garantida pela Constituição Federal de 1988. Para atender a esse direito, foi criado o Sistema Nacional de Informações de Segurança Pública, Prisionais, de Rastreabilidade de Armas e Munições, de Material Genético, de Digitais e de Drogas (SINESP) [Brasil 2019], o qual representa a principal referência sobre informações criminais fornecidas e padronizadas para auxiliar na elaboração de políticas públicas. Os esforços para manter a ordem pública possuem um custo elevado; de acordo com [Koegl and Day 2019], o crime é considerado caro, tanto para o custo financeiro quanto para o custo individual e social.

Em tempos modernos, o uso do aprendizado de máquina é uma realidade na administração pública de vários países, utilizando-se da coleta de dados públicos por meio de órgãos governamentais para contribuir com a sociedade. O Brasil não fica para trás, com iniciativas como o da Advocacia-Geral da União implementando um assistente virtual para auxiliar na analise de documentos e na elaboração de textos jurídicos [Brasil 2023]. Com isso, a tecnologia da informação e comunicação (TIC) tem se mostrado uma forte aliada aos serviços públicos, destacando-se o uso de modelos de aprendizado de máquina como uma das áreas mais promissoras.

Este estudo tem como objetivo explorar o uso de modelos baseados no ajuste fino da arquitetura *Bidirectional Encoder Representations from Transformers* (BERT) para agilizar o processo de consolidação de relatos policiais, utilizando dados coletados junto à Secretaria Adjunta Pública e Defesa Social (SIAC), vinculada à Secretaria de Segurança Pública e Defesa Social (SEGUP), no Pará, Brasil. O conjunto de dados utilizado para treino, validação e teste consiste nas dez categorias criminais com maior incidência em boletins de ocorrência de uma região escolhida para a realização do estudo.

De acordo com o Anuário Brasileiro, disponibilizado pelo Fórum Brasileiro de Segurança Pública (FBSP), o município de Altamira/PA está entre as maiores taxas do país quando se trata de crimes que envolvem mortes violentas intencionais e estupro [Brasil 2024]. Devido a esses altos índices, Altamira foi a cidade escolhida para a seleção dos dados criminais e para realizar uma consequente comparação entre os modelos utilizados para classificação.

As seções restantes deste trabalho estão estruturadas da seguinte forma: a Seção 2 aborda a revisão da literatura relevante sobre o tema; a Seção 3 descreve em detalhes a metodologia empregada para pré-processamento dos dados e classificação de crimes; a análise comparativa dos modelos implementados é apresentada na Seção 4; por fim, a Seção 5 discute os resultados obtidos e sugere direções para pesquisas futuras.

2. Trabalhos Relacionados

Os fatores relacionados a criminalidade têm sido objetos de estudos por muitos anos, utilizando uma variedade de abordagens para compreender as dinâmicas por trás desses fenômenos. Nos últimos anos, surgiram diversos estudos propondo formas de empregar a mineração de texto e a ciência de dados para extrair conteúdo de valor em conjunto de dados de segurança publica.

Com o propósito de estudar a aplicação de modelos de aprendizado de máquina em boletins de ocorrência no Estado de São Paulo. Assad *et al.* [Assad and Chagas 2019] selecionaram os algoritmos Árvore de Decisão e Regressão Logística para a identificação de padrões presentes nos registros criminais, como veículos, marcas de celulares e locais mais frequente para a ocorrência do crime. Por fim, foi elaborado um mapeamento geográfico da cidade, que ilustra a distribuição das ocorrências criminais predominantes em cada região. Além disso, foi realizada uma análise comparativa da eficiência dos modelos utilizados, permitindo avaliar como cada abordagem se destacou na identificação e compreensão dos padrões criminais.

Um estudo realizado na região metropolitana de Belém/PA propôs verificar a influência da vulnerabilidade social sobre a taxa de homicídios entre jovens de 15 a 19

anos [Trindade 2019]. Através de análises quantitativas e qualitativas, o estudo revelou tendências e fenômenos presentes nos dados. Os resultados indicaram que jovens do sexo masculino que abandonaram o ambiente escolar, que trabalhavam ou estavam sem ocupação, possuíam uma maior fator de risco para crimes de caráter homicida devido à falta de políticas públicas eficientes.

Ao analisar um conjunto de dados pertencente à Cidade de Denver, EUA, abrangendo o período de janeiro 2014 a março de 2019 [Ratul 2020], foram realizados esforços para prever ocorrências criminais por meio de algoritmos de classificação, e assim apoiar as entidades legislativas a se precaver em relação a taxas de ocorrência registradas. Foram realizados testes a partir de uma seleção de algoritmos de classificação, como *Random Forest*, Árvore de Decisão, *K-Neighbors Classifier* (KNN), *Linear Discriminant Analysis* (LDA), AdaBoost, ExtraTrees e mais quatro modelos diferentes de *Ensemble* para classificar quinze diferentes classes criminais. Com exceção do AdaBoost, todos os algoritmos obtiveram desempenhos satisfatórios, alcançando resultados acima de 90% de acurácia, com um destaque especial para um dos métodos de *Ensemble* que manteve a acurácia citada para todas as quinze classes definidas.

Para auxiliar as autoridades de segurança pública, foi realizado uma análise exploratória e a extração de regras associativas de uma base de dados de boletins de ocorrência nos anos de 2019 a 2021, de forma a obter conhecimentos de registros criminais em Belém/PA [Souza 2022]. Utilizando a metodologia *Cross-industry Standard Process for Data Mining* (CRISP-DM) para guiar o seus estudos, foi avaliado que adultos do sexo masculino da faixa etária de 35 a 64 anos são as principais vítimas de crimes. Além disso, houve a utilização de dois algoritmos diferentes para o estudo da base de dados, com o uso do modelo *Apriori* para a extração de relacionamentos frequentes entre os itens da base de dados, e a implementação do modelo *Convolutional Neural Networks (CNN)* para classificação em categorias criminais, onde o modelo alcançou uma acurácia geral de aproximadamente 78% no conjunto de teste.

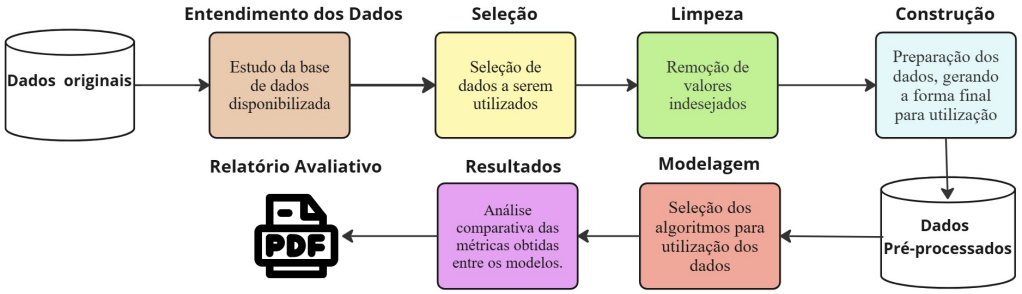
Com o objetivo de aplicar modelos de Processamento de Linguagem Natural (PLN) para automatizar a classificação de boletins de ocorrência, Alves *et al.* [Alves et al. 2024] conduziu um estudo utilizando os algoritmos BERT, RoBERTa e ALBERT para analisar os registros criminais mais recorrentes no município de Marabá/PA, entre os anos de 2019 e 2021. Inicialmente, foi necessário submeter os dados a uma etapa de pré-processamento, a fim de remover informações irrelevantes e garantir a qualidade dos dados a serem classificados. Os resultados apresentados no estudo destacaram os modelos BERT e RoBERTa, que demonstraram um desempenho superior, alcançando uma acurácia entre 89% e 90% nos testes realizados.

3. Metodologia

CRISP-DM é uma metodologia que se popularizou por propor uma modelo de processo abrangente para a realização de projetos que envolvem mineração de dados. Suas fases englobam: Entendimento do negócio, entendimento de dados, preparação dos dados, modelagem, avaliação e implementação. De acordo com [Chapman et al. 2000], apesar da metodologia possuir etapas bem definidas, o modelo foi desenvolvido para permitir a flexibilidade conforme as necessidades do projeto. Com isso, este estudo realizou adaptações para que fossem atendidas as necessidades do desenvolvimento do classifica-

dor, onde as adaptações realizadas são demonstradas na Figura 1 e discutidas a seguir.

Figura 1. Fases adaptadas do modelo CRISP-DM.



3.1. Entendimento dos Dados

O conjunto de dados disponibilizados pela SIAC (Estado do Pará), possui os registros policiais coletados do período de 2019 a 2023, com grande parte dos seus atributos sendo preenchidos pelo relator (escrivão da Polícia Civil nas delegacias ou a própria vítima através das delegacias virtuais). Um campo a ser destacado é o *relato*, uma descrição textual que narra o evento ocorrido a partir das informações coletadas nesse campo, como vítimas, autores, localização, *modus operandi*, etc. É importante destacar que os campos a serem preenchidos no sistema de registro aceitam qualquer tipo de entrada, e não possuem mecanismos de seleção ou correção automática. Essa ausência de controle na inserção dos dados pode levar à inclusão de registros imprecisos, especialmente quando os relatos são fornecidos em circunstâncias emergenciais.

A padronização dos dados é feita por meio da leitura dos relatos e da identificação do crime ocorrido. Para isso, é designado um grupo de quinze analistas do departamento de estatística da SIAC, selecionados com base em seus conhecimentos da legislação brasileira. Em casos mais complexos de identificação do crime com base no relato, um julgamento coletivo dos analistas é necessário. Após a identificação do crime ocorrido, a informação será direcionado para o atributo *consolidado*, o qual define uma classe final para que possam ser gerados relatórios estatísticos. Nessa análise geralmente é investido esforço para categorizar crimes violentos (homicídio, estupro, feminicídio, etc.) devido à sensibilidade dos relatos registrados, necessitando uma análise meticulosa. Diante do vasto volume de dados coletados diariamente nas delegacias, o atributo "consolidado" torna-se a coluna ideal para uma ferramenta que auxilie na confirmação das classes criminais, alinhando-se aos padrões da SIAC para a geração de relatórios estatísticos de segurança pública.

3.2. Seleção

Em sua forma original, a base de dados possui um total de 2.400.792 amostras e 72 atributos (*e.g.*, data, município, idade) , mas devido ser um grande conjunto de dados, foi necessário realizar uma seleção de quais dados deveriam ser usados através de uma cidade alvo, a qual pudesse ser representada pelas dez classes criminais com maior ocorrência na região, com Altamira/PA sendo o município escolhido. Nas Figuras 2 e 3 é possível

observar os quantitativos dessas classes criminais em dois cenários diferentes, sendo eles, respectivamente: período de 2019 a 2021, e o período total contido na base de dados, 2019 a 2023.

A divisão supracitada foi realizada devido parte dos dados terem sidos coletados em um ambiente atípico, que foi influenciado por eventos extraordinários e condições sociais em mudança. Logo, tem-se a oportunidade de avaliar os algoritmos em dois cenários distintos. O primeiro deles compreende os dados registrados de 2019 a 2021, marcados pela ocorrência do Coronavírus 2019 (COVID-19), onde em fevereiro de 2020 foi registrado o primeiro caso no Brasil [Farias 2020]. O segundo cenário engloba os dados registrados de 2019 a 2023, resultando na avaliação de todos os relatos criminais registrados dentro das dez classes mais frequentes em Altamira/PA.

É importante destacar o comunicado realizado em 2022 pelo Ministério da Saúde, o qual informava o fim do estado emergencial de saúde causado pela referida pandemia [Brasil 2022]. Com essa declaração pública, podemos considerar o ano de 2022 como o início da pós-pandemia no Brasil. Dessa forma, os algoritmos devem se adaptar a dados que foram coletados em ambientes pandêmicos e pós-pandêmicos.

Figura 2. As dez classes criminais mais frequentes no período de 2019 a 2021 na Cidade de Altamira/PA.

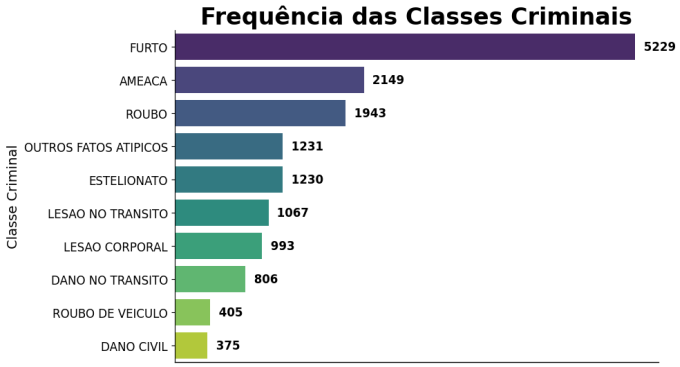
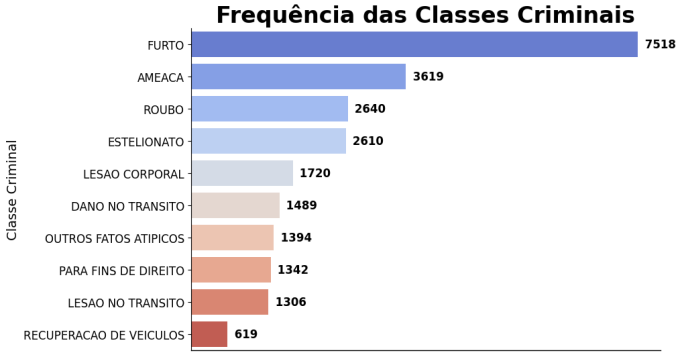


Figura 3. As dez classes criminais mais frequentes no período de 2019 a 2023 na Cidade de Altamira/PA.



3.3. Limpeza e construção

No geral, com a seleção de Altamira/PA como cidade alvo, foram contabilizadas 37.909 amostras disponíveis e 72 atributos. Entretanto, é necessário que os dados a serem utilizadas pelos algoritmos passem por um pré-processamento para que sejam removidos

conteúdos indesejados que possam prejudicar os resultados. Para tanto, dois atributos serão fundamentais para o desenvolvimento do classificador, um atributo textual e um atributo com classe alvo, sendo eles o "relato" (descrição do acontecimento) e o "consolidado" (classe final atribuída para um boletim de ocorrência), respectivamente.

Presentes nesta base de dados, existem valores duplicados ou irrelevantes que acabam inflando o número de amostras que foi apresentado, sendo necessário realizar uma etapa de limpeza nos dados para a remoção desses valores. A seguir serão descritas as etapas que foram realizadas:

- Amostras duplicadas: Crimes com múltiplas vítimas/autores em um mesmo evento, possuem o campo "nro_ bop" como seu identificador único, pelo que podemos preservar apenas a primeira ocorrência e a duplicidade da coluna "relato" pode ser descartada.
- Caracteres especiais e *stopwords*:Para simplificar o conteúdo textual foram removidas as *tags* HTML presentes no texto, as quais são provenientes da página web onde os relatos são realizados. Além disso, foram removidos acentos, pontuações e sinais, e eliminadas *stopwords*, i.e., palavras que não influenciam no contexto, como artigos e preposições.
- Informação sensível: Para evitar qualquer enviesamento na análise de palavras, foram ocultadas algumas informações, como vítimas, dados pessoais, localidades, endereços, etc., as quais foram substituídas por etiquetas (*e.g.*, RG, CPF e LOCAL) para minimizar o efeito da ausência dessas palavras.

Após a remoção dos valores indesejados, todos os caracteres foram convertidos para minúsculos. Ao final desta etapa, 18.372 amostras estavam devidamente preparadas e foram transformadas em um *Hugging Face dataset*. A Figura 4 ilustra o avanço progressivo na preparação dos dados, desde a fase de seleção até a conclusão do pré-processamento, resultando em amostras prontas para análise pelos classificadores.

Figura 4. Histórico de preparação dos dados.

3.4. Modelagem

Para a classificação de boletins de ocorrência, foram utilizados os algoritmos Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al. 2018] e sua variante BERTimbau [Souza et al. 2020]. Ambos os modelos utilizam uma abordagem bidirecional que, ao processar um relato criminal, considera o contexto completo de uma palavra ao levar em conta os termos que a antecedem e a sucedem, o que resulta em uma compreensão mais profunda e precisa de seu significado. Na tarefa de classificação, os modelos utilizam *tokens* especiais, como o [CLS], que representa a sequência inteira de palavras, e o [SEP], que delimita o seu fim, assim facilitando a previsão das classes correspondentes aos dados de entrada. A principal diferença entre o BERT e o BERTimbau reside nos

dados de pré-treinamento: enquanto o BERT foi treinado com uma ampla variedade de textos em inglês, o BERTimbau, por sua vez, foi especificamente pré-treinado em grandes volumes de texto em português extraídos da Wikipédia, o que o torna mais adequado para lidar com essa língua em particular.

Os dados pós-processados foram divididos em três subconjuntos estratificados, sendo 80% alocados para o treinamento, 10% para a validação e 10% para o teste. Todos os modelos foram treinados sob os mesmos hiperparâmetros: "Tamanho de lote para avaliação" fixado em 1, "Taxa de decaimento" ajustada para 0,01 e o "Número de épocas de treinamento" estabelecido em 3, e para finalizar, cada modelo assume médias de vinte execuções para a análise dos resultados.

4. Resultados

As especificações de *hardware* utilizada neste estudo foram: CPU AMD Ryzen 7 3700 X, 8 núcleos de 4,4 GHz; Placa de vídeo Nvidia RTX 3060 de 12 GB; Memória RAM de 64 GB e SSD de 500 GB. Ademais, como os modelos empregados são supervisionados, serão aplicadas as seguintes métricas para avaliação: *Accuracy, Precision, Recall e f1-score.*

Tabela 1: Resultado dos modelos para dados de 2019 a 2021 em Altamira/PA: Subconjunto de teste.

Métrica	BERT	BERTimbau
Ameaça		
Precision	0,87	0,89
Recall	0,89	0,91
f1-score	0,88	0,90
Dano Civil		
Precision	0,00	0,42
Recall	0,00	0,26
f1-score	0,00	0,32
Dano no Trânsito		
Precision	0,93	0,86
Recall	0,96	0,91
f1-score	0,94	0,88
Estelionato		
Precision	0,91	0,94
Recall	0,95	0,91
f1-score	0,93	0,92
Furto		
Precision	0,97	0,97
Recall	0,95	0,96
f1-score	0,96	0,97
Lesão Corporal		
Precision	0,75	0,87
Recall	0,77	0,89
f1-score	0,76	0,88

Lesão no Trânsito		
Precision	0,90	0,89
Recall	1,00	0,98
f1-score	0,95	0,93
Outros Fatos Atípicos		
Precision	0,67	0,73
Recall	0,69	0,70
f1-score	0,68	0,71
Roubo		
Precision	0,88	0,90
Recall	0,91	0,94
f1-score	0,89	0,92
Roubo de Veículo		
Precision	0,85	0,77
Recall	0,76	0,81
f1-score	0,80	0,79
Accuracy	0,89	0,90

A Tabela 1 exibe as métricas iniciais da análise dos classificadores, com várias classes criminais apresentando métricas bastante semelhantes entre os modelos. Isso resultou em uma acurácia quase idêntica para o BERTimbau e o BERT, evidenciando a notável capacidade dos modelos em diferenciar entre as classes. Embora categorias criminais como "Estelionato" e "Lesão no Trânsito" não possuam a maior quantidade de dados, elas foram suficientemente representadas para que os modelos atingissem níveis satisfatórios de *precision* e *recall*, com o BERTimbau superando o BERT por uma margem pequena. Entretanto, a classe "Dano Civil" se destacou como um desafio maior para a classificação, principalmente devido à limitada quantidade de amostras disponíveis. Apesar das métricas insatisfatórias do modelo BERTimbau para essa classe, ele ainda demonstrou uma capacidade de classificação em conjuntos de dados reduzidos.

Tabela 2: Resultado dos modelos para dados de 2019 a 2023 em Altamira/PA: Subconjunto de teste.

Métrica	BERT	BERTimbau
Ameaça		
Precision	0,81	0,89
Recall	0,87	0,91
f1-score	0,84	0,91
Dano no Trânsito		
Precision	0,54	0,93
Recall	0,82	0,95
f1-score	0,65	0,94
Estelionato		
Precision	0,46	0,90

Recall	0,95	0,92
f1-score	0,62	0,91
Furto		
Precision	0,90	0,96
Recall	0,87	0,94
f1-score	0,89	0,95
Lesão Corporal		
Precision	0,00	0,87
Recall	0,00	0,92
f1-score	0,00	0,89
Lesão no Trânsito		
Precision	0,00	0,91
Recall	0,00	0,97
f1-score	0,00	0,94
Outros Fatos Atípicos		
Precision	0,00	0,42
Recall	0,00	0,36
f1-score	0,00	0,39
Para Fins de Direito		
Precision	0,23	0,56
Recall	0,02	0,51
f1-score	0,40	0,53
Recuperação de Veículos		
Precision	0,00	0,79
Recall	0,00	0,86
f1-score	0,00	0,82
Roubo		
Precision	0,62	0,95
Recall	0,90	0,96
f1-score	0,73	0,95
Accuracy	0,68	0,88

Ao realizar uma análise comparativa das métricas apresentadas nas Tabelas 1 e 2, as limitações do modelo BERT na classificação de boletins de ocorrência tornaram-se evidentes. Conforme mostrado na Tabela 2 , o modelo apresentou dificuldades significativas, obtendo resultados nulos em categorias como “Lesão Corporal”,“Lesão no Trânsito”,“Outros Fatos Atípicos” e “Recuperação de Veículos”. Essas dificuldades podem ser atribuídas às diferenças contextuais entre os dados coletados durante e após a pandemia de Covid-19, o que resulta em variações significativas nos relatos e nos padrões específicos de cada ambiente. Essa diversidade exige uma análise mais minuciosa, que se torna desafiadora para um modelo de linguagem natural que não está totalmente adaptado à linguagem específica dos relatos criminais.

Por outro lado, o BERTimbau, com sua especialização na língua portuguesa, su-

perou essas limitações ao captar nuances mais sutis sobre os relatos criminais, alcançando métricas factíveis em grande parte das classes. Entretanto, o modelo encontrou dificuldades em classificar categorias mais abrangentes como "Outros Fatos Atípicos" e "Para Fins de Direito", devido a seus relatos descreverem ocorrências que são mais amplas. Com isso, os resultados demonstraram que um maior domínio da língua contribui para uma melhor diferenciação entre classes e, consequentemente, uma maior eficiência na classificação.

5. Conclusões e Trabalhos Futuros

Este estudo detalhou a implementação de um classificador supervisionado empregado em uma base de dados de registros policiais fornecida pela SIAC, com foco no modelo BERTimbau, que obteve acurácia de 88% a 90% nos testes realizados.

A literatura existente apresenta diversos estudos que exploraram a classificação de boletins de ocorrências criminais, com o objetivo de empregar a tecnologia para automatizar essa tarefa. Isso permite que as forças policiais concentrem seus esforços em atividades de maior relevância para a sociedade. A arquitetura BERT e suas variantes surgem como uma área promissora para esse propósito, oferecendo a possibilidade de desenvolver ferramentas valiosas para a administração pública e outras agências governamentais, devido à sua excepcional capacidade de lidar com tarefas de PLN.

Para trabalhos futuros, é possível modificar hiperparâmetros como o número de épocas e a quantidade de repetições, utilizando as métricas apresentadas neste estudo como referência para comparação. Além disso, é viável explorar diferentes modelos de PLN e variantes do modelo BERT, como o BERT Multilingual e o DistilBERT. Se o objetivo for desenvolver uma ferramenta de valor para o serviço público, o aprimoramento contínuo dessa arquitetura e o investimento dedicado a sua evolução poderão justificar o esforço e os recursos aplicados, beneficiando assim a sociedade.

Referências

Alves, D., Marques, M., Santos, R., and Santos, A. (2024). Classificação de boletins de ocorrências através de modelos de linguagem baseados em bert. In *Anais do XII Workshop de Computação Aplicada em Governo Eletrônico*, pages 169–179, Porto Alegre, RS, Brasil. SBC.

Assad, F. J. P. and Chagas, J. F. C. (2019). Análise preditiva de manchas criminais no estado de são paulo.

Brasil (2019). O sistema nacional de informações de segurança pública, prisionais, de rastreabilidade de armas e munições, de material genético, de digitais e de drogas (sinesp). Disponível em: `https://www.gov.br/mj/pt-br/assuntos/sua-seguranca/seguranca-publica/sinesp-1/`. Acesso em : 27 de julho de 2024.

Brasil (2022). Após dois anos, chega ao fim estado de emergência em saúde pública por conta da covid-19 no brasil. Disponível em: `https://www.gov.br/saude/pt-br/assuntos/noticias/2022/maio/apos-dois-anos-chega-ao-fim-estado-de-emergencia-em-saude-publica-por-conta-da-covid-19-no-brasil`. Acesso em : 27 de julho de 2024.

Brasil (2023). Agu inova no uso de inteligência artificial para aprimorar eficiência e prestação de serviços à sociedade. Disponível em: https://www.gov.br/agu/pt-br/comunicacao/noticias/agu-inova-no-uso-de-inteligencia-artificial-para-aprimorar-eficiencia-e-prestacao-de-servicos-a-sociedade. Acesso em : 18 de agosto de 2024.

Brasil (2024). Anuário brasileiro de segurança pública 2024. Disponível em: https://apidspace.universilab.com.br/server/api/core/bitstreams/80177eeb-4a88-40f6-98f5-c476dea0f3db/content. Acesso em : 27 de julho de 2024.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., et al. (2000). Crisp-dm 1.0: Step-by-step data mining guide. *SPSS inc*, 9(13):1–73.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Farias, H. S. d. (2020). O avanço da covid-19 e o isolamento social como estratégia para redução da vulnerabilidade. *Espaço e Economia. Revista brasileira de geografia econômica*, (17).

Koegl, C. J. and Day, D. M. (2019). The monetary costs of crime for a sample of offenders in ontario. *Canadian Journal of Criminology and Criminal Justice*, 61(3):21–44.

Ratul, M. A. R. (2020). A comparative study on crime in denver city based on machine learning and data mining. *arXiv preprint arXiv:2001.02802*.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Souza, S. L. (2022). Mineração de dados em banco de dados de segurança pública no estado do pará, brasil. Dissertação de Mestrado, Programa de Pós-Graduação em Segurança Pública. Universidade Federal do Pará.

Trindade, E. A. R. A. (2019). Homicídios na região metropolitana de Belém: práticas para contenção e vulnerabilidades. Dissertação de Mestrado, Programa de Pós-Graduação em Segurança Pública. Universidade Federal do Pará.