MODELO DE CLASSIFICAÇÃO DE TTP BASEADO EM TRANSFORMADAS BERT

Paulo M. M. R. Alves, Geraldo P. R. Filho e Vinícius P. Gonçalves
Universidade de Brasília, Brasil

RESUMO

Informações relativas às Táticas Técnicas e Procedimentos (TTP) observados em um ataque são costumeiramente apresentadas na forma de textos não estruturados. A aplicação de técnicas de aprendizagem de máquina junto ao Processamento de Linguagem Natural (NLP) pode auxiliar na identificação dessas TTP. Esse trabalho propõe o enfrentamento desse problema por meio da utilização de modelos BERT (*Bidirectional Encoder Representations from Transformers*), estado da arte no campo de NLP. O *dataset* utilizado inicialmente é a base de sentenças do instituto MITRE, sendo posteriormente realizada validação em conjunto de sentenças manualmente anotadas extraído de relatórios de CTI (*Cyber Threat Intelligence*) públicos. São testados onze modelos e também se conduz uma investigação dos efeitos de alguns parâmetros no ajuste fino do modelo. O objetivo é identificar o modelo e a combinação de parâmetros que melhor se adequam à tarefa de classificação de acordo com as convenções sobre TTP do framework ATT&CK do MITRE. Como resultado, verificou-se que os melhores modelos apresentaram acurácia de 0,8264 e 0,7875 nos dois conjuntos de dados utilizados, demonstrando a viabilidade e a relevância do uso dos modelos BERT nessa complexa tarefa do domínio cibernético.

PALAVRAS-CHAVE

Aprendizagem de Máquina, NLP, BERT, TTP

1. INTRODUÇÃO

Os ataques cibernéticos têm aumentado não apenas em volume, mas também em complexidade. Em 2020, ataques por *malwares* em geral *e ransonwares* tiveram um crescimento de 358% e 435% respectivamente (World Economic Forum, 2022). Os últimos anos tem presenciado alguns dos ataques mais sofisticados e críticos já ocorridos. Defender redes de computadores nesse cenário é uma tarefa desafiadora. Os defensores necessitam de informação acionável, precisa e oportuna. Informações sobre ameaças cibernéticas normalmente podem ser encontradas em relatórios de Inteligência de Ameaças Cibernéticas (CTI, na sigla em inglês). Esses documentos difundem informações sobre o *modus operandi* dos atacantes, particularmente Indicadores de Comprometimento (IOC) e Táticas, Técnicas e Procedimentos (TTP).

Indicadores de Comprometimento apresentam informações sobre dados brutos específicos (IPs, *hashes*, domínios etc). Muitos *feeds* de inteligência de ameaças focam nos IOCs e muitas soluções tradicionais de segurança são baseadas em IOCs. Contudo, esses dados não são suficientes para uma proteção adequada (You, et al., 2022). Carecem de contexto informacional adequado para melhor descrever o padrão de ataque. Além disso, Ameaças Persistentes Avançadas (APT), muitas vezes, possuem a capacidade de modificar seus próprios IOCs (Zhu & Dumitras, 2018) (Strom, et al., 2017).

Dessa forma, informações sobre TTPs são cada vez mais relevantes, pois descrevem o comportamento do atacante, sendo, portanto, menos voláteis. A capacidade de proteger a rede e prevenir contra TTPs aumenta consideravelmente a barreira de custo do ataque, visto que o atacante precisará aprender novos padrões de ataque. O conhecido modelo da Pirâmide da Dor (Bianco, 2013), por um lado considera IOCs dados mais simples e triviais, que proporcionam inteligência de menor valor. Por outro lado, TTPs situam-se no ápice da pirâmide, sendo considerados a informação mais valiosa para os gestores de segurança.

A organização MITRE elaborou o *framework* ATT&CK, uma base de conhecimento de comportamento dos atacantes que descreve as TTPs conhecidas (Strom, et al., 2020). A atual versão do ATT&CK (divulgada em abril de 2022) inclui 14 táticas, 191 técnicas e 386 subtécnicas (The MITRE Corporation, 2022). Essa

matriz visa criar uma padronização para esses dados. Contudo, como as TTPs são normalmente divulgadas em relatórios de CTI na forma de textos não estruturados, classificar esses textos em centenas de técnicas e subtécnicas permanece sendo uma tarefa desafiadora.

Apesar da evolução no NLP com a aplicação de modelos de aprendizagem de máquina, a segurança cibernética ainda não parece ter se beneficiado completamente desses avanços (Ponemon Institute, 2021). Nosso trabalho inova ao modelar o BERT para aplicação no problema de classificação de texto específico para mapeamento de TTPs junto a um framework estruturado (MITRE ATT&CK). As principais contribuições desse trabalho são: a) emprega o estado da arte em NLP (BERT) utilizando 11 modelos diferentes para classificar sentenças em 253 TTPs distintas; b) conduz uma varredura de diferentes combinações de parâmetros selecionados de ajuste fino para otimização de desempenho e avalia a correlação dos parâmetros com a performance; c) identifica a melhor configuração dos parâmetros escolhidos e o melhor modelo BERT para classificação de TTPs presentes em textos.

O restante desse artigo está estruturado como explanado a seguir. A seção 2 (Trabalhos Relacionados) promove uma revisão da literatura relacionada. A seção 3 (Metodologia) apresenta a metodologia empregada, explicitando como foi feita a preparação dos dados e os modelos e configurações utilizados no experimento. A seção 4 (Resultados e Discussão) expõe e discute os resultados alcançados com a metodologia proposta. Ao final, a seção 5 (Conclusões e Trabalhos Futuros) apresenta as conclusões do experimento e discute possibilidade de desenvolvimentos futuros convergentes com a linha de pesquisa desse trabalho.

2. TRABALHOS RELACIONADOS

Uma das maiores dificuldades no uso de NLP no domínio cibernético é a pouca disponibilidade de bases de dados consistentes e anotadas (Tikhomirov, et al., 2020). Conjuntos de dados relacionados a TTPs são ainda mais raros e essa escassez dificulta o avanço de pesquisas em classificação de TTPs (You, et al., 2022) (Riera, et al., 2022). A despeito da importância, ainda há pouca pesquisa voltada ao problema da extração de TTPs de textos não estruturados (Rahman, et al., 2020). Legoy et al (2019) testa múltiplos métodos de representação de texto com diferentes classificadores, encontrando como melhor combinação a representação pelo método TF-IDF acompanhado do classificador Linear Support Vector (LinearSVC).

Husari et al (2017) propõem o TTPDrill, uma abordagem que emprega TF-IDF com uma versão modificada do algoritmo BM25. Os mesmos autores posteriormente propõem o ActionMiner, modelagem que faz uso dos conceitos de entropia e informação mútua da Teoria da Informação. Outros trabalhos similares buscam por ações maliciosas utilizando uma variedade de técnicas, como grafos de proveniência (Satvat, et al., 2021) e aprendizado de máquina supervisionado (Ghazi, et al., 2018).

A pesquisa de Ayoade et al (2018), por sua vez, emprega métodos de correção de viés, de propagação de confiança e de estimativa de importância de pesos para fazer predições de táticas e técnicas presentes em relatórios de CTI. You et al (2022) propõem o framework Threat Intelligence Mining (TIM), desenvolvendo a ferramenta TCENet. A solução faz sua análise agrupando conjuntos de três sentenças de modo a buscar mais contexto. Também busca combinar IOCs mostrados nos relatórios com as técnicas, de modo a enriquecer o conhecimento das TTPs com mais dados contextuais.

Outro trabalho relevante nesse campo é o TRAM (Threat Report ATT&CK Mapper), feito pela instituição MITRE. Essa ferramenta aplica Regressão Logística na predição de técnicas relacionadas a cada sentença, utilizando as sentenças exemplos da base do MITRE para treinamento do algoritmo. Cada classificação proposta deve ser manualmente revisada por um analista humano (Yoder & Lasky, 2019).

A multiplicidade de fontes de CTI produz sobrecarga de informações e torna impraticável ao analista extrair manualmente TTPs dessa massa de relatórios (Ranade, et al., 2021) (Legoy, et al., 2019) (Husari, et al., 2017). Buscando solucionar esse problema, pesquisadores de segurança cibernética tem recorrido cada vez mais a técnicas de Processamento de Linguagem Natural (NLP) e Recuperação de Informação (IR). Percebeu-se que também a automação é essencial e muitos estudos recentes tem combinado métodos de inteligência artificial com NLP (Harel, et al., 2017) (Ghazi, et al., 2018) (Rahman, et al., 2020).

O campo do Processamento de Linguagem Natural obteve grandes avanços com a incorporação de técnicas de aprendizagem de máquina. O trabalho de Conneau et al (2017) sobre representação universal de sentenças mostrou que métodos de transferência de aprendizagem possuem aplicabilidade em tarefas NLP. Vaswani

et al (2017) propuseram a arquitetura de Transformadas, um modelo baseado em mecanismo de auto-atenção para representar entradas e saídas.

Empregando conjuntamente técnicas de transferência de aprendizagem e a arquitetura de Transformadas, Devlin et al (2019) propuseram o BERT (*Bidirectional Encoder Representations from Transformers*), um modelo em duas etapas, pré-treinamento e ajuste fino, cujo desempenho o alçou a condição de estado da arte para diversas tarefas NLP. Prottasha et al (2022) confirmaram essa condição ao testar diferentes modelos de representação (Word2Vec, GloVe, FastText e BERT) e demonstrar que o BERT, com o ajuste fino adequado supera os demais modelos em diversas tarefas de NLP.

3. METODOLOGIA

3.1 Preparação de Dados

Desde o lançamento do framework ATT&CK, em 2015, o MITRE mantém uma base de conhecimento anotada manualmente de informações extraídas de relatórios de CTI (Strom, et al., 2020). Esse repositório, entre outros dados, conta com 10360 sentenças ilustrativas de TTPs. A Tabela **1** abaixo mostra alguns exemplos:

Tabela 1. Exemplo de sentenças exemplo da base do MITRE com as correspondentes técnicas ou subtécnicas

Sentença	*ID da técnica*	*Nome da técnica*
The NETWIRE payload has been injected into benign Microsoft executables via process hollowing.	T1055.012	Process Injection: Process Hollowing
Sykipot contains keylogging functionality to steal passwords.	T1056.001	Input Capture: Keylogging
Mosquito deletes files using DeleteFileW API call.	T1070.004	Indicator Removal on Host: File Deletion
RTM has initiated connections to external domains using HTTPS.	T1071.001	Application Layer Protocol: Web Protocols
Dragonfly has compromised user credentials and used valid accounts for operations.	T1078	Valid Accounts
Patchwork payloads download additional files from the C2 server.	T1105	Ingress Tool Transfer
PlugX has a module to create, delete, or modify Registry keys.	T1112	Modify Registry
Sandworm Team's CredRaptor tool can collect saved passwords from various internet browsers.	T1555.003	Credentials from Password Stores: Credentials from Web Browsers
TA551 has sent spearphishing attachments with password protected ZIP files.	T1566.001	Phishing: Spearphishing Attachment
APT33 has sent spearphishing emails containing links to .hta files.	T1566.002	Phishing: Spearphishing Link

Entre as 576 técnicas e subtécnicas, 466 possuem pelo menos uma sentença ilustrativa. Em razão da necessidade de dados para treinamento dos modelos de aprendizagem de máquina, limitamos o escopo às técnicas e subtécnicas que apresentam pelo menos 5 exemplos. Sob esse critério, aproveitamos 9909 sentenças exemplo e trabalhamos com as 253 técnicas mais comuns. Dadas as especificidades da base de dados do MITRE, modelamos o problema de classificação de TTPs usando a abordagem multiclasse (cada amostra recebe um único rótulo), na qual cada uma dessas 253 técnicas constituirá uma classe.

Uma peculiaridade desse conjunto de dados é o desbalanceamento entre as classes, problema comum em bases de dados textuais (Padurariu & Breaban, 2019). Na base ora utilizada, a maior classe apresenta 371 exemplos enquanto as menores, pelas restrições experimentais impostas, possuem 5 sentenças. Felizmente, contudo, pesquisas mostram que BERT lida bem com bases desbalanceadas e estratégias de extensão de dados (*data augmentation*) não impactam significativamente a performance (Tikhomirov, et al., 2020) (Madabushi, et al., January 2019) (Iikura, et al., 2020) (Oak, et al., 2019).

Para utilizar o modelo BERT, as sentenças precisam ser individualmente "tokenizadas" (separadas em *tokens* que representam palavras ou parte de palavras, além de alguns *tokens* de controle) e codificadas (os tokens devem ter representações numéricas, pois o modelo, na verdade, enxerga conjuntos de matrizes numéricas representando o texto). Após isso, dividimos o conjunto de dados em *datasets* de treinamento, validação e teste na proporção 60:20:20. Adotamos uma estratégia de amostragem estratificada, a qual garante que cada *dataset* possua exemplos de todas as técnicas.

Para fins de validação, após realizar treinar os modelos BERT nas sentenças exemplo do repositório MITRE, utilizamos os modelos para realizar predições em um conjunto de 80 sentenças extraídas e anotadas manualmente de relatórios de CTI públicos (*dataset* de inferência).

3.2 Modelos e Configurações

Inicialmente estabelecemos uma linha de base por meio da utilização de um modelo que combina a representação TF-IDF com classificador por Regressão Linear. Posteriormente, processamos os dados utilizando onze versões de BERT: BERT Base Cased, BERT Base Uncased, BERT Large Cased, BERT Large Uncased, RoBERTa Base, RoBERTa Large, DistilRoBERTa, DistilBERT Uncased, DistilBERT Cased, SecBERT e SecRoBERTa. Os dois últimos modelos são modelos treinados em dados cibernéticos que incluem relatórios de CTI e apresentam vocabulário expandido.

Para estabelecer a parametrização inicial, utilizamos recomendações de pesquisas anteriores (Devlin, et al., 2019) (Sun, et al., 2019) e experimentação. Definimos o tamanho máximo de cada amostra individual enviada ao BERT em 256, a taxa de aprendizagem em 2e-5 e o tamanho de cada lote de amostras em 16. Dado o alto número de classes (253) de nosso experimento, decidimos, inicialmente, treinar por 30 épocas.

Conduzimos também uma análise de parametrização buscando potenciais otimizações e examinando o efeito das alterações de parâmetro no desempenho do modelo. Devido ao alto custo computacional da tarefa, selecionamos os parâmetros de taxa de aprendizado (utilizando as configurações 1e-4, 5e-5, 2e-5 e 1e-5) e tamanho do lote (com as configurações 8, 16, 34 e 32) e aplicamos todas as combinações entre esses parâmetros por 10 épocas para cada par de valores. O modelo escolhido foi BERT Base Uncased, de tamanho mediano entre as diferentes versões. Após esse procedimento, aplicamos o melhor ajuste identificado aos modelos BERT para examinar o efeito no desempenho. A Figura **1** proporciona uma visão geral da nossa abordagem:

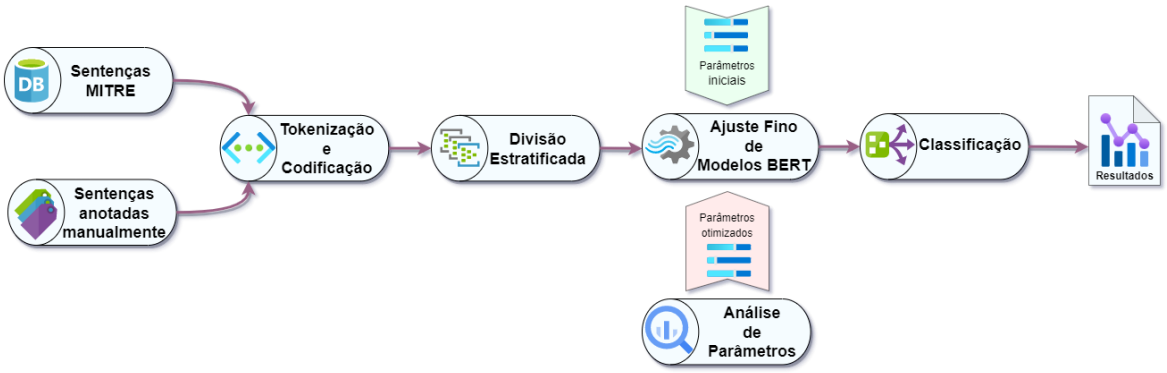

Figura 1. Visão geral da abordagem utilizada para classificação de TTPs

4. RESULTADOS E DISCUSSÃO

Optamos por utilizar a métrica da acurácia para avaliação. Em problemas do tipo multiclasse, as médias micro de precisão, *recall* e *F-measure* igualam-se à acurácia. As médias macro, por sua vez, são bastante afetadas pelo desbalanceamento de classes. Como na base MITRE não há uma dominância de classes (a maior classe representa apenas 3,58% do total) e nosso problema de classificação não apresenta preferência ou precedência entre as classes, a métrica da acurácia proporciona boa compreensão do desempenho global. A acurácia é definida conforme a formula abaixo:

$$acurácia = \frac{VP + VN}{VP + FN + VN + FP}$$

na qual VP significa Verdadeiro Positivo; VN, Verdadeiro Negativo; FN, Falso Negativo; e FP, Falso Positivo.

O modelo TF-IDF/Regressão Linear, utilizado como linha de base, obteve uma acurácia de 0,6051 no *dataset* de teste e 0,4770 no *dataset* de inferência. Aplicamos então os onze modelos BERT com os parâmetros iniciais. A Figura 2 mostra, à guisa de exemplo, as curvas de acurácia e função de perda (*training loss*) para o modelo BERT Base Uncased treinado por 30 épocas.

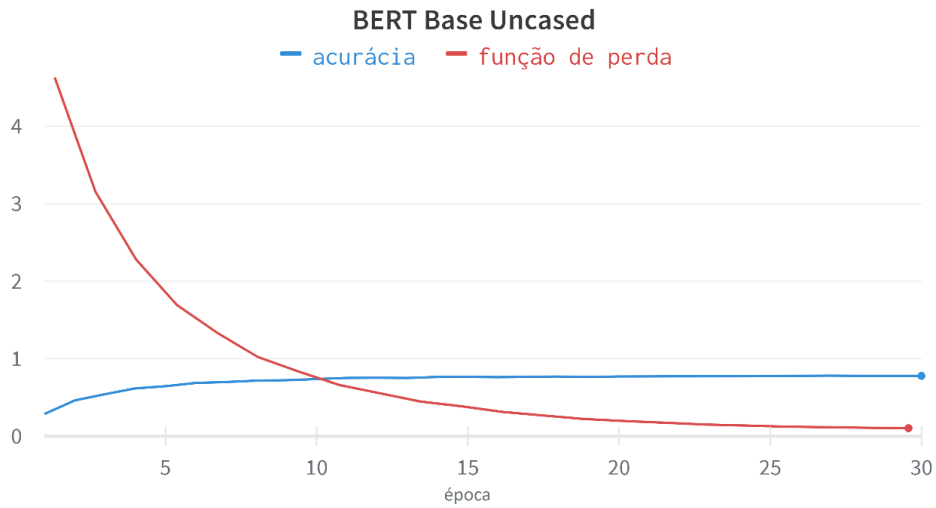

Figura 2. Acurácia e Função de Perda para o modelo BERT Base Uncased

Observe-se que o eixo y da Figura 2 não apresenta rótulo, pois a função de perda não apresenta unidade e a acurácia é medida entre 0 e 1 (ou sua percentagem correspondente). As curvas apresentam o comportamento esperado para o treinamento de aprendizagem de máquina, com a acurácia percorrendo uma curva ascendente e a função de perda decaindo. A Tabela **2** apresenta a acurácia obtida nos *datasets* de teste e de inferência:

Tabela 2. Acurácia dos modelos BERT na classificação de TTPs nos *datasets* de teste e de inferência utilizando os parâmetros iniciais

Modelos	*Dataset* de teste	*Dataset* de inferência
BERT Base Uncased	0,7719	0,6375
BERT Base Cased	0,7906	0,7125
BERT Large Uncased	0,8143	0,7250
BERT Large Cased	0,8032	**0,7875**
RoBERTa Base	0,7951	0,7000
RoBERTa Large	**0,8264**	0,7750
DistilRoBERTa Base	0.7931	0.6500
DistilBERT Base Uncased	0.7840	0.7125
DistilBERT Base Cased	0.7729	0.6750
SecBERT	0,7830	0,7000
SecRoBERTa	0,7633	0,7000

Todos os modelos foram treinados por 30 épocas para ajuste fino. Os modelos que obtiveram o melhor desempenho foram RoBERTa Large e BERT Large Cased, com uma acurácia de 0,8264 e 0,7875 nos *datasets* de teste e inferência, respectivamente. Ambos são modelos grandes, com redes neurais de 24 camadas e pré-treinados com 355 milhões e 340 milhões de parâmetros, respectivamente. Na arquitetura de transformadas do BERT, o tamanho afeta o desempenho, ainda que não de forma drástica (Devlin, et al., 2019).

Percebe-se, pela Tabela 2 que, assim como nosso modelo de linha de base (TF-IDF/Regressão Linear), os resultados das predições nos dados de inferência são piores que os obtidos para o *dataset* de teste. Avaliamos que esse resultado se deve ao fato de que as sentenças retiradas de relatórios de CTI (inferência) são mais longas e mais complexas que os exemplos da base do MITRE (teste). Além disso, diferentes organizações e analistas apresentam padrões, convenções e estilos de escrita distintos, tornando os dados mais heterogêneos.

Constata-se ainda que os modelos treinados em textos do domínio cibernético não apresentaram resultados superiores, contrariando as expectativas. Observa-se também que os modelos BERT apresentam menor diferença entre o desempenho no teste e na inferência. A linha de base apresentou uma diferença de 12,8 pontos percentuais no desempenho, enquanto os modelos BERT com as configurações iniciais tiveram uma diferença média de 8,4 pontos percentuais. BERT aprimora a compreensão de sentenças mais longas e complexas do domínio cibernético.

BERT alcança boa performance no problema de classificação de TTPs. Comparando o melhor modelo (RoBERTa Large) ao nosso modelo de linha de base, percebemos um incremento de 22,1 pontos percentuais na acurácia. A comparação entre os nossos resultados e outros trabalhos precedentes é dificultada pelo fato de que, a despeito do objetivo ser semelhante, a similaridade entre os trabalhos é limitada por diferentes premissas iniciais.

O TCENet, mostrou classificação de TTPs com uma acurácia de 94,1%. No entanto, os testes do TCENet envolveram apenas as cinco técnicas (ou subtécnicas) mais populares, uma tática e as respectivas classes negativas (You, et al., 2022). Nossa pesquisa aplicou modelos BERT para classificação de 253 técnicas e subtécnicas. Husari et al [13] alega ter obtido precisão de 84% e *recall* de 82%, no entanto o experimento foi realizado sob pressupostos significativamente diferentes. A abordagem daquele trabalho não empregava aprendizagem de máquina e baseava-se em uma ontologia previamente construída que precisaria ser refeita manualmente a cada atualização do framework ATT&CK.

Buscando otimizar ainda mais o ajuste fino, conduzimos uma análise de parametrização na qual testamos 16 possíveis combinações de pares taxa de aprendizado/tamanho do lote. A Figura 3 abaixo apresenta os resultados:

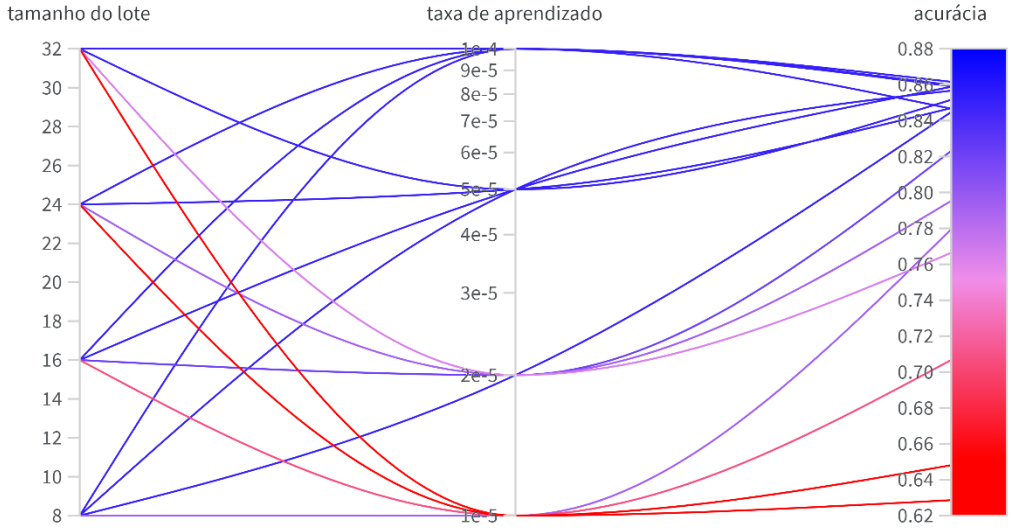

Figura 3. Análise de parametrização para taxa de aprendizado e tamanho do lote.

Essa análise permitiu observar que a taxa de aprendizado apresenta correlação positiva de 0,670 com a acurácia. Esse dado informa que, nas condições do nosso experimento, é provável que maiores taxas de aprendizado impliquem acurácia mais elevada. O tamanho do lote, contudo, não apresentou correlação significativa (-0,283) com a métrica escolhida. Esses resultados vão ao encontro do conceito de Goodfellow et al (2016) de que a taxa de aprendizado é o mais importante parâmetro a ser ajustado em modelos de aprendizagem de máquina.

A melhor combinação observada dados os parâmetros testados foi: taxa de aprendizado de 1e-4 e tamanho de lote 24. No entanto, ao aplicar essa taxa, todos os modelos do tipo “Large” incorreram no chamado esquecimento catastrófico. Essa situação consiste na incapacidade da rede neural de reter informações antigas quando apresentada a informações novas. Constitui problema comum na aprendizagem de máquina no campo de NLP, particularmente quando se utilizam taxas de aprendizado mais altas (Kaushik, et al., 2021) (Sun, et al., 2019).

Aplicando essa configuração aos 11 modelos BERTs estudados, verificamos os resultados explicitados na Tabela **3**:

Tabela 3. Acurácia dos modelos BERT na classificação de TTPs nos datasets de teste e de inferência utilizando os parâmetros otimizados. EC corresponde a situações de esquecimento catastrófico

Modelos	*Dataset* de teste	*Dataset* de inferência
BERT Base Uncased	0,7996	0,7000
BERT Base Cased	0,7840	0,7250
BERT Large Uncased	EC	EC
BERT Large Cased	EC	EC
RoBERTa Base	0,8007	0,6875
RoBERTa Large	EC	EC
DistilRoBERTa Base	**0,8012**	0,7538
DistilBERT Base Uncased	0,7825	**0,7625**
DistilBERT Base Cased	0,7936	0,7125
SecBERT	0,7926	0,6750
SecRoBERTa	0,7845	0,7000

Percebe-se uma pequena tendência de melhoria de desempenho. Os modelos "destilados" (DistilBERT Cased e Uncased e DistiRoBERTa, modelos mais enxutos, com redes neurais de 6 camadas e treinados com 65, 66 e 85 milhões de parâmetros respectivamente) apresentaram os maiores avanços. Contudo, nenhum dos modelos menores atingiu as marcas de desempenho dos modelos Large alcançados com os parâmetros iniciais.

5. CONCLUSÕES E TRABALHOS FUTUROS

Analistas de segurança cibernética precisam de recursos que facilitem a aquisição de informações essenciais ao seu trabalho. Nossa pesquisa contribui para isso ao aplicar modelos BERT, estado da arte em NLP, ao problema de classificar TTPs. Utilizamos a base de sentenças rotuladas do MITRE com uma estratégia de amostragem estratificada e convertemos o dataset em um formato adequado ao BERT.

Empregamos 11 diferentes modelos BERT, obtendo as melhores acurácias com o RoBERTa Large (*dataset* de teste) e BERT Large Cased (*dataset* de inferência). Conduzimos uma "varredura" de diversas combinações de parâmetros. Também investigamos o efeito da taxa de aprendizagem e tamanho do lote na acurácia, buscando optimização. Constatamos que o tamanho do lote não apresentou correlação com a acurácia em nosso experimento. Verificamos que a taxa de aprendizado pode produzir pequenas melhorias na acurácia, mas sob o risco do fenômeno do esquecimento catastrófico, observado nos modelos maiores (Large) para a taxa mais alta.

O presente trabalho demonstrou que a arquitetura BERT de transformadas constitui uma ferramenta útil e relevante para equacionar o problema de classificação de TTPs retirados de bases textuais. No futuro, é possível entender esse trabalho para uma modelagem multilabel, mitigando o problema de sentenças longas com múltiplos TTPs descritos. Além disso, os efeitos de outros parâmetros (como decaimento de pesos ou taxa de dropout) podem ser investigados.

REFERÊNCIAS

Ayoade, G. et al., 2018. *Automated Threat Report Classification Over Multi-Source Data.* Philadelphia, PA, USA, IEEE, p. 11.

Bianco, D. J., 2013. *The Pyramid of Pain.* [Online] Available at: https://detect-respond.blogspot.com/2013/03/the-pyramid-of-pain.html [Acesso em 20 junho 2022].

Conneau, A. et al., 2017. *Supervised Learning of Universal Sentence Representations from Natural Language Inference Data.* Copenhagen, Denmark, Association for Computational Linguistics, pp. 670-680.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K., 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Minneapolis, Minnesota, USA, Association for Computational Linguistics, pp. 4171-4186.

Ghazi, Y. et al., 2018. *A Supervised Machine Learning Based Approach for Automatically Extracting High-Level Threat Intelligence from Unstructured Sources.* Islamabad, Pakistan, IEEE, pp. 129-134.

Goodfellow, I., Bengio, Y. & Courville, A., 2016. *Deep Learning (Adaptive Computation and Machine Learning series) Illustrated Edition.* Illustrated ed. s.l.:The MIT Press.

Harel, Y., Gal, I. B. & Elovici, Y., 2017. Cyber security and the role of intelligent systems in addressing its challenges. *ACM Transactions on Intelligent Systems and Technology (TIST) - Special Issue: Cyber Security and Regular Papers*, May, 8(4), p. 12.

Husari, G. et al., 2017. *TTPDrill: Automatic and Accurate Extraction of Threat Actions from Unstructured Text of CTI Sources.* Orlando, Estados Unidos, s.n., pp. 103-115.

Iikura, R., Okada, M. & Mori, N., 2020. *Improving BERT with Focal Loss for Paragrah Segmentation of Novels.* L´Aquila, Italy, Springer, Cham, pp. 21-30.

Japkowicz, N. & Shah, M., 2011. *Evaluating learning algorithms: a classification perspective.* s.l.:Cambridge University Press.

Kaushik, P., Gain, A., Kortylewski, A. & Yuille, A., 2021. *Understanding Catastrophic Forgetting and Remembering in Continual Learning with Optimal Relevance Mapping.* Virtual, s.n.

Legoy, V., Caselli, M., Seifert, C. & Peter, A., 2019. *Automated Retrieval of ATT&CK Tactics and Techniques for Cyber Threat Reports,* Enschede: University of Twente.

Madabushi, H. T., Kochkina, E. & Castelle, M., January 2019. *Cost-Sensitive BERT for Generalisable Sentence Classification with Imbalanced Data.* Hong Kong, China, Association for Computational Linguistics, pp. 125-134.

Mithun, M. P., Suntwal, S. & Surdeanu, M., 7-11 November 2021. *Students Who Study Together Learn Better: On the Importance of Collective Knowledge for Domain Transfer In Fact Verification.* s.l., Association for Computational Liguistics, pp. 6968-6973.

Oak, R. et al., 2019. *Malware Detection on Highly Imbalanced Data through Sequence Modeling.* London, United Kingdom, Association for Computing Machinery, pp. 37-48.

Padurariu, C. & Breaban, M. E., 2019. Dealing with Data Imbalance in Text Classification. *Procedia Computer Science. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES2019.*, Volume 159, pp. 736-745.

Ponemon Institute, 2021. *6th Cyber Resilient Organization Study* , s.l.: IBM Security.

Rahman, R., Mahdavi-Hezaveh, R. & Williams, L., 2020. *A Literature Review on Mining Cyberthreat Intelligence from Unstructured Texts.* Sorrento, Italy, s.n.

Ranade, P., Piplai, A., Joshi, A. & Finin, T., 2021. *CyBERT: Contextualized Embeddings for the Cybersecurity Domain.* s.l., IEEE, pp. 334-3342.

Riera, T. S. et al., 2022. A new multi-label dataset for Web attacks CAPEC classification using machine learning techniques. *Computers & Security*, 05 06, Volume 120, p. 102788.

Satvat, K., Gjomemo, R. & Venkatakrishnan, V. N., 2021. *Extractor: Extracting Attack Behavior from Threat Reports.* Vienna, Austria, IEEE, pp. 598-615.

Strom, B. E. et al., 2020. *MITRE ATT&CK: Design and Philosophy,* McLean, VA: The MITRE Corporation.

Strom, B. E. et al., 2017. *Finding Cyber Threats with ATT and CK(registered trademark)-Based Analytics,* Annapolis Junction, MD: The MITRE Corporation..

Sun, C., Qiu, X., Xu, Y. & Huang, X., 2019. *How to Fine-Tune BERT for Text Classification?.* Kunming, China, Springer International Publishing, pp. 194-2016.

The MITRE Corporation, 2022. *Updates - April 2022.* [Online] Available at: https://attack.mitre.org/resources/updates/ [Acesso em 28 May 2022].

Tikhomirov, M., Loukachevitch, N., Sirotina, A. & Dobrov, B., 2020. *Using BERT and Augmentation in Named Entity Recognition for Cybersecurity Domain.* Saarbrücken, Germany, Springer, Cham, pp. 16-24.

World Economic Forum, 2022. *The Global Risks Report 2022: 17th Edition,* s.l.: s.n.

Yoder, S. & Lasky, J., 2019. *Automating Mapping to ATT&CK: The Threat Report ATT&CK Mapper (TRAM) Tool.* [Online] Available at: https://medium.com/mitre-attack/automating-mapping-to-attack-tram-1bb1b44bda76 [Acesso em 12 fevereiro 2022].

You, Y. et al., 2022. TIM: threat context-enhanced TTP intelligence mining on unstructured threat data. *Cybersecurity,* 5(3), p. 17.

Zhu, Z. & Dumitras, T., 2018. *ChainSmith: Automatically Learning the Semantics of Malicious Campaigns by Mining Threat Intelligence Reports.* London, UK, IEEE, pp. 458-472.