CENTRO UNIVERSITÁRIO FEI

HUGO LINHARES OLIVEIRA
JOÃO PEDRO ROSA CEZARINO
THALES DE OLIVEIRA LACERDA
VITOR MARTINS OLIVEIRA

DO TWEET À AMEAÇA: UM ESTUDO SOBRE PADRÕES DE DETECÇÃO DE AMEAÇAS CIBERNÉTICAS ATRAVÉS DE PROCESSAMENTO DE LINGUAGEM NATURAL

São Bernardo do Campo
2023

HUGO LINHARES OLIVEIRA
JOÃO PEDRO ROSA CEZARINO
THALES DE OLIVEIRA LACERDA
VITOR MARTINS OLIVEIRA

DO TWEET À AMEAÇA: UM ESTUDO SOBRE PADRÕES DE DETECÇÃO DE AMEAÇAS CIBERNÉTICAS ATRAVÉS DE PROCESSAMENTO DE LINGUAGEM NATURAL

Trabalho de Conclusão de Curso, apresentado ao Centro Universitário FEI, como parte dos requisitos necessários para obtenção do título de Bacharel em Ciência da Computação. Orientado pelo Profº. Dr. Charles Henrique Porto Ferreira

São Bernardo do Campo
2023

DO TWEET À AMEAÇA: UM ESTUDO SOBRE PADRÕES DE DETECÇÃO DE AMEAÇAS CIBERNÉTICAS ATRAVÉS DE PROCESSAMENTO DE LINGUAGEM NATURAL / Hugo Linhares Oliveira...[et al.]. São Bernardo do Campo, 2023.
50 p. : il.

Trabalho de Conclusão de Curso - Centro Universitário FEI.
Orientador: Prof. Dr. Charles Henrique Porto Ferreira.

1. Segurança cibernética. 2. Redes sociais. 3. Machine Learning. 4. Processamento de texto. 5. Análise de sentimentos. I. Oliveira, Hugo Linhares. II. Cezarino, João Pedro Rosa. III. Lacerda, Thales de Oliveira. IV. Oliveira, Vitor Martins. V. Ferreira, Charles Henrique Porto, orient. VI. Título.

Elaborada pelo sistema de geração automática de ficha catalográfica da FEI com os dados fornecidos pelo(a) autor(a).

Hugo Linhares Oliveira
João Pedro Rosa Cezarino
Thales de Oliveira Lacerda
Vitor Martins Oliveira

DO TWEET À AMEAÇA: UM ESTUDO SOBRE PADRÕES DE DETECÇÃO DE AMEAÇAS CIBERNÉTICAS ATRAVÉS DE PROCESSAMENTO DE LINGUAGEM NATURAL

Trabalho de Conclusão de Curso, apresentado ao Centro Universitário FEI, como parte dos requisitos necessários para obtenção do título de Bacharel em Ciência da Computação.

Comissão julgadora

_____

Profº. Dr. Charles Henrique Porto Ferreira

_____

Profº. Dr. Paulo Sergio Silva Rodrigues

_____

Profº. Dr. Luciano Rossi

São Bernardo do Campo
2023

Dedicamos este trabalho às famílias de cada um dos integrantes, pilares de nosso crescimento, e ao Profº. Dr. Charles Henrique Porto Ferreira, bússola de nossa jornada neste projeto, com profunda gratidão e respeito.

AGRADECIMENTOS

Neste momento especial de conclusão da jornada acadêmica, gostaríamos de expressar nossa mais profunda gratidão àqueles que tornaram esta caminhada possível e enriquecedora. Em primeiro lugar, estendemos nossos sinceros agradecimentos às nossas famílias, cujo apoio incondicional e compreensão foram fundamentais para superarmos os desafios e alcançarmos nossos objetivos. Um agradecimento especial é reservado ao Profº. Dr. Charles Henrique Porto Ferreira, nosso orientador, cuja sabedoria, paciência e orientação criteriosa nos guiaram por este projeto. Sua capacidade de nos desafiar e, ao mesmo tempo, oferecer suporte, foi crucial para o desenvolvimento do nosso trabalho. Sua influência estende-se além das páginas deste trabalho, marcando nossa formação acadêmica e profissional de maneira definitiva.

“Mesmo a defesa cibernética mais corajosa experimentará derrotas quando as fraquezas são negligenciadas.”

Stéphane Nappo

RESUMO

Com o aumento dos ataques cibernéticos, a segurança digital torna-se crucial. Redes sociais, especialmente o Twitter, são plataformas onde hackers expressam intenções. Este trabalho propõe um método de extração e análise de dados dessas redes usando técnicas de *Machine Learning* e pré-processamento de textos para identificar padrões indicativos de ameaças cibernéticas. Diante desse desafio, são apresentadas duas abordagens: na primeira, há a combinação de sentimentos, entidades e similaridade com palavras-chave de segurança da informação em uma única representação vetorial, juntamente com os resultados de um algoritmo de classificação. Já na segunda abordagem, é utilizado um score ponderado para cada atributo do mecanismo de análise, visando uma abordagem mais refinada na detecção de possíveis ameaças. Os resultados destacam a importância da análise de entidades na melhoria da precisão do modelo, onde identificou-se que datas e números são mais prevalentes em comunicações que contêm ameaças. Além disso, os resultados obtidos questionam a eficácia da análise de sentimentos como indicador confiável, desafiando a premissa de que a polaridade do sentimento é um sinal seguro de conteúdo mal-intencionado na identificação de tweets potencialmente perigosos. Neste cenário, o algoritmo Random Forest se destacou, alcançando uma acurácia de até 79,59% na classificação de tweets como ameaças, contra 79,25% de baseline.

Palavras-chave: Segurança cibernética, Ataques cibernéticos, Redes sociais, Machine Learning, Processamento de texto, Rastreamento de dados, Análise de sentimentos

ABSTRACT

"With the increase in cyber attacks, digital security becomes crucial. Social networks, especially Twitter, are platforms where hackers express their intentions. This work proposes a method for extracting and analyzing data from these networks using Machine Learning techniques and text preprocessing to identify patterns indicative of cyber threats. In the face of this challenge, two approaches are presented: the first combines sentiment analysis, entity recognition, and similarity with information security keywords into a single vector representation, along with the results of a classification algorithm. The second approach uses a weighted score for each attribute of the analysis mechanism, aiming for a more refined approach in detecting potential threats. The results highlight the importance of entity analysis in improving the model's accuracy, where it was found that dates and numbers are more prevalent in communications containing threats. Moreover, the obtained results challenge the effectiveness of sentiment analysis as a reliable indicator, defying the premise that sentiment polarity is a sure sign of malicious content in the identification of potentially dangerous tweets. In this scenario, the Random Forest algorithm stood out, achieving an accuracy of up to 79.59% in classifying tweets as threats, compared to a 79.25% baseline."

Keywords: Cybersecurity, Cyber attacks, Social networks, Machine Learning, Text processing, Data tracking, Sentiment analysis

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A segurança da informação é a prática que visa proteger computadores, servidores, dispositivos móveis, sistemas eletrônicos, softwares, redes e dados contra-ataques maliciosos e acessos indesejados (GUTTMAN; ROBACK, 1995). Tais ataques maliciosos e acessos indesejados podem assumir diversas formas e impactar diferentes componentes da infraestrutura tecnológica de uma organização. Por exemplo, o ataque de negação de serviço em massa (DDoS/DOS) visa sobrecarregar os servidores e dispositivos de rede com um grande volume de tráfego, tornando-os inacessíveis para os usuários. Já a encriptação de dados (Ransomware) é um tipo de malware que criptografa os dados do usuário e exige o pagamento de um resgate para desbloqueá-los. Todos esses ataques têm em comum o fato de representarem uma ameaça à segurança da informação e, portanto, devem ser prevenidos ou mitigados por meio de medidas de segurança apropriadas. Além disso, a segurança da informação não se limita apenas a ataques externos, mas também envolve a proteção contra erros humanos e falhas técnicas.

Os três pilares da segurança da informação são a Confidencialidade, a Integridade e a Disponibilidade (Tríade "CIA", do termo em inglês *confidentiality*, *integrity* and *availability*) (STALLINGS, 2017). Cada um desses pilares tem a importante função de assegurar a confiabilidade das informações que serão disponibilizadas. A confidencialidade garante a restrição do acesso à informação por usuários não previamente autorizados. A integridade é responsável por garantir que a informação esteja completa e exata, sem modificações inesperadas. Por fim, a disponibilidade fica responsável por garantir que a informação esteja acessível e utilizável sempre que necessário.

O constante crescimento e sofisticação da internet e dos sistemas distribuídos trazem uma grande melhora na eficiência, automatização de tarefas e distribuição de informação. No entanto, a crescente interdependência entre estes serviços tem aumentado exponencialmente os riscos aos quais tais sistemas estão sujeitos. Nesse contexto, ficou evidente nos últimos anos o ganho de relevância e atenção que grupos hackers como "Anonymous", "Conti", "BlackCat" e "Lockbit" têm recebido por comprometer a segurança de grandes empresas (INDUSTRIALCYBER, 2020). Entre esses grupos, sempre existe um pilar motivacional que se destaca durante o ataque: o **ganho de atenção**. A reivindicação da autoria de grandes ataques através das redes sociais e grandes meios de comunicação reitera a presença e a imponência de um grupo acima dos outros.

As redes sociais desempenham um papel muito importante na comunidade de Segurança da Informação, uma vez que representam um ambiente para a busca de informações abertas, seja de forma automatizada ou manual. Diversos membros de grupos hackers, entusiastas, especialistas em segurança e pesquisadores da área se beneficiam das redes sociais para compartilhar conhecimento e se atualizarem sobre novas ameaças, vulnerabilidades e possíveis ataques cibernéticos. Plataformas como o Twitter são extremamente populares, sendo atualmente utilizadas para dar voz às mais diversas opiniões e sentimentos, sem pré-avaliação ou julgamento dos dados informados.

A análise de sentimentos a partir das percepções inseridas pelos usuários nestas redes sociais tem sido frequentemente utilizada para estudos com foco em Mineração de Opiniões e predição de possíveis eventos. A análise de sentimentos é uma técnica que utiliza recursos computacionais para reconhecer, extrair, ponderar e identificar estados emocionais implícitos em "posts" presentes nestas plataformas (ALSAEEDI; KHAN, 2019). Através da filtragem de termos específicos, é possível compreender o humor e a intenção de um usuário e utilizá-lo como métrica para ponderar novas pesquisas, estudar fenômenos sociais e até mesmo estudos psicológicos.

Das mais variadas plataformas de compartilhamento de opinião, o Twitter tem se destacado como plataforma preferida de organizações de segurança, pesquisadores da área, entusiastas, grupos hackers e outros perfis de usuários relacionados ao tema (SMITH, 2022). A natureza de tempo real do Twitter permite que pesquisadores extraiam conclusões relevantes durante eventos de grande impacto e ajam em tempo hábil. Essa rede social é constantemente utilizada para prever eventos como terremotos e desastres naturais (SAKAKI; OKAZAKI; MATSUO, 2010). Sendo assim, no contexto de segurança da informação, o Twitter é utilizado para comunicar novas vulnerabilidades, reclamar a autoria de ataques já realizados e coordenar novos ataques cibernéticos para disseminar o pânico no meio digital. No trabalho de (TRIPWIRE, 2018) há um exemplo de como estes *tweets* são extremamente valiosos para uma análise mais aprofundada, em que um usuário revelou uma vulnerabilidade nunca encontrada e mostrou, em seu post, o código capaz de explorar tal vulnerabilidade. Consequentemente, em um curto período, um grupo mal-intencionado conhecido como "PowerPool Gang" começou a utilizar o código maliciosamente, explorando a vulnerabilidade encontrada pelo pesquisador em campanhas maliciosas. Através da análise de sentimentos pode-se recolher grandes volumes de dados eficientemente e aliado a técnicas de processamento de linguagem natural, pode-se identificar padrões relacionados a campanhas como a mencionada, de modo que, futuramente, grandes prejuízos sejam evitados.

Segundo (HERNANDEZ et al., 2016), a análise e processamento de linguagem natural indica que características textuais com carga negativa, associadas a termos lexicais de segurança da informação, frequentemente utilizados por grupos hacktivistas, podem servir como alerta para identificar ciberataques. Consequentemente, dados extraídos do Twitter, relacionados a eventos políticos, religiosos e culturais de grande importância, podem ser utilizados para antecipar ataques desse tipo (ZHANG et al., 2017).

1.1 OBJETIVO

Dada a relevância dos *tweets* na disseminação de informações e seu potencial impacto no mundo da Segurança da Informação, este trabalho foca na identificação de padrões de ameaças cibernéticas por meio de técnicas de processamento de linguagem natural (NLP) e aprendizado de máquina (Machine Learning — ML). O foco central é a construção de um modelo capaz de extrair elementos-chave das conversas, tais como: sentimentos, entidades específicas e avaliação da proximidade dos *tweets* com um conjunto de termos e contextos relacionados à segurança cibernética para avaliar se um determinado *tweet* pode representar uma ameaça cibernética. Além disso, busca-se compreender os fatores que influenciam as predições dos algoritmos de classificação, bem como as palavras e relações gramaticais que contribuem para classificar um *tweet* como potencialmente suspeito.

1.2 ESTRUTURA DO TRABALHO

Para alcançar tal objetivo, o projeto se divide em duas abordagens complementares. A primeira abordagem visa integrar diferentes variáveis para formar um mecanismo mais eficaz na detecção de ameaças virtuais. Esta abordagem contempla elementos como os padrões de conversação, a tonalidade emocional dos textos (categorizados como positivos ou negativos por um algoritmo BiLSTM), a identificação de entidades pertinentes utilizando técnicas de Processamento de Linguagem Natural (NLP) e, por fim, a análise de similaridade contextual para validar o conteúdo dos textos para com um glossário específico de termos de segurança cibernética. Todos esses atributos são posteriormente agregados em um único vetor, combinado com a representação vetorial do *tweet*. Esse vetor consolidado alimenta um algoritmo de classificação padrão, como KNN (K-Nearest Neighbors), Random Forest ou SVM (Support Vector Machine). Essa combinação aprimora significativamente a capacidade do sistema de reconhe-

cer mensagens potencialmente mal-intencionadas, resultando em um aumento da precisão na identificação de conteúdo potencialmente malicioso.

Já a segunda estratégia do projeto foca na elaboração de um sistema de avaliação que utiliza um score de pontuação ponderada. Este score é projetado para analisar o conteúdo textual, atribuindo pesos diferenciados a aspectos cruciais como o sentimento detectado, as entidades reconhecidas e a proximidade semântica com palavras-chave comumente associadas com ameaças cibernéticas. Além disso, incorpora uma camada de avaliação proveniente de um modelo classificador tradicional, que contribui com um valor adicional na análise. Os respectivos pesos atribuídos à cada característica foram decorrentes de experimentos realizados ao longo do projeto. Estes fatores são agregados em uma média ponderada, culminando na geração de uma pontuação que estima a probabilidade de uma determinada frase ser um indicativo de risco no ambiente digital.

2 TRABALHOS RELACIONADOS

Este capítulo apresenta um levantamento dos trabalhos relacionados ao tema desta pesquisa. O principal objetivo é contextualizar o tema e identificar as principais contribuições de estudos anteriores para o desenvolvimento deste trabalho. Para isso, foram selecionados artigos, livros e dissertações que abordam o tema discutido. A seguir, será apresentada a análise dos trabalhos, bem como as principais conclusões alcançadas pelos autores destes.

Deb, Lerman e Ferrara (2018) discutem uma abordagem proativa para a identificação de ameaças cibernéticas. O projeto visa prever futuros eventos cibernéticos, como malwares, endereços maliciosos e e-mails maliciosos, em janelas de tempo específicas. A metodologia envolve o processamento de dados de fóruns na web superficial e na dark web, que discutem tópicos de segurança da informação. Os dados, que abrangem o período de um ano, são utilizados para gerar sinais e alertas sobre possíveis ameaças cibernéticas. Métodos de análise de sentimento são usados para analisar o texto das postagens e o objetivo dos autores é maximizar a soma das semelhanças entre os alertas previstos e os eventos reais. A abordagem do projeto fornece uma maneira mais proativa e eficiente de identificar possíveis ameaças cibernéticas, melhorando assim as medidas de segurança cibernética existentes.

Semelhantemente ao projeto anterior, (NUNES et al., 2016) se concentra na exploração e análise de plataformas sociais da darknet e deepnet, particularmente fóruns e mercados relacionados à atividade de hacking malicioso. O projeto implementa um sistema de coleta de informações, usando técnicas de mineração de dados e aprendizado de máquina para desenvolver um rastreador. Este rastreador é projetado para identificar e separar informações relevantes de possíveis ruídos, lidando com desafios como limpeza de texto, erros ortográficos, variações de palavras e grande espaço de recursos. Os dados coletados são então disponibilizados para profissionais de segurança para identificar ameaças e tendências cibernéticas emergentes.

Hernandez et al. (2016) discute um modelo de previsão de ataques de segurança com base na análise de sentimentos de usuários extraídos de posts do Twitter. O modelo usa técnicas de aprendizado de máquina para analisar as emoções e opiniões expressas pelos usuários no Twitter para identificar padrões que possam indicar uma possível ameaça à segurança. Os resultados mostram que o modelo é eficaz na previsão de ataques de segurança e poderia ser usado como uma ferramenta complementar às medidas de segurança tradicionais.

O estudo Hernandez-Suarez et al. (2018), apresenta um método de previsão de ciberataques usando dados de sentimentos extraídos do Twitter. Os autores apresentam um modelo

baseado em aprendizado de máquina que utiliza uma técnica de regularização ℓ_1 para selecionar os atributos mais relevantes para a predição. O modelo é treinado com uma base de dados de *tweets* relacionados à cibersegurança e, em seguida, colocado à prova em um conjunto de *tweets* relevantes ao contexto de Segurança da Informação. O estudo conclui que o modelo proposto supera outros métodos de predição em termos de precisão e eficiência.

O projeto de Dionísio et al. (2019) propõe arquiteturas de redes neurais profundas para implementar as tarefas centrais de um pipeline de processamento para obter informações de segurança relevantes, oportunas e direcionadas do Twitter. O sistema proposto consegue coletar *tweets* de um conjunto de contas, filtrá-los com base em um conjunto de palavras-chave que definem uma infraestrutura a ser monitorada, selecionar os *tweets* que contêm informações relevantes e identificar informações úteis nesses *tweets*. Para isso, foram implementadas redes neurais convolucionais e redes de memória de longo prazo bidirecionais (BiLSTM). A abordagem proposta supera metodologias bem estabelecidas em termos de desempenho.

Em (BEHZADAN et al., 2018) os autores visam abordar o desafio de monitorar e analisar eficientemente grandes volumes de dados gerados em plataformas de mídia social para identificar possíveis ameaças cibernéticas em tempo real. Eles desenvolvem uma coleção de textos selecionados e organizados que servem como uma base de dados para treinamento e teste de algoritmos de aprendizado de máquina. A utilidade de tal corpus é crucial por conter exemplos anotados de indicadores de ameaças, que o classificador de aprendizado profundo pode aprender a reconhecer. O objetivo é fornecer uma ferramenta automatizada que possa ajudar analistas de segurança e pesquisadores a identificar rapidamente potenciais ameaças cibernéticas, permitindo uma resposta mais rápida a incidentes de segurança. O estudo é significativo no contexto de segurança cibernética, pois redes sociais como o Twitter se tornaram fontes importantes de informações em tempo real, que podem incluir anúncios de novas vulnerabilidades, ataques em andamento ou discussões sobre falhas de segurança. Ao utilizar técnicas de aprendizado profundo, o projeto visa melhorar as capacidades de detecção e análise de ameaças em meio a um grande volume de dados, representando um avanço importante na caça a ameaças cibernéticas por meio de fontes abertas de inteligência.

3 CONCEITOS FUNDAMENTAIS

Nesta seção, serão explorados conceitos fundamentais para o projeto, fornecendo uma compreensão clara da metodologia empregada e de como ela é aplicada ao processamento e análise de *tweets*. Esses conceitos são importantes não apenas para assimilar o método utilizado, mas também para interpretar corretamente os resultados alcançados. Assim, será aprofundado o entendimento específico da análise de sentimentos, identificação de contexto e reconhecimento de entidades nomeadas em *tweets*, o que possibilitará a extração de percepções significativas.

3.1 ATAQUES CIBERNÉTICOS

Um ataque cibernético, no contexto do projeto, pode ser definido como uma atividade maliciosa, a qual é indicada ou discutida em *tweets* que visa comprometer a integridade, confidencialidade ou disponibilidade de recursos de informação digitais. Estes ataques podem variar desde a exploração de vulnerabilidades em software ou hardware, phishing, ataques de negação de serviço (DoS), até campanhas de desinformação e espionagem cibernética.

3.2 APRENDIZADO DE MÁQUINA

O Aprendizado de Máquina, ou *Machine Learning*, é um conjunto de técnicas que utilizam modelos estatísticos para aprenderem com dados fornecidos. Trata-se de uma abordagem computacional que capacita um algoritmo a aprender com exemplos e tomar decisões ou realizar tarefas sem ser explicitamente programado para cada situação. Neste contexto, algoritmos são treinados com exemplos rotulados de *tweets* para aprender padrões e realizar a classificação automatizada do sentimento expresso. Neste âmbito, destacam-se dois aspectos importantes: o aprendizado supervisionado e a análise de sentimentos.

3.2.1 Aprendizado Supervisionado

O Aprendizado Supervisionado, uma abordagem de aprendizado de máquina, envolve o treinamento de um modelo com dados previamente rotulados, ou seja, nos quais as respostas corretas já são conhecidas. O modelo é treinado com *tweets* já rotulados com sentimentos

positivos, negativos ou neutros, aprendendo a classificar o sentimento em novos *tweets*. Esta abordagem se relaciona diretamente com outro conceito fundamental: a análise de sentimentos.

3.2.2 Análise de Sentimentos

A Análise de Sentimentos é uma técnica cujo objetivo é a identificação, extração e quantificação da opinião, emoção ou sentimento expresso em um texto. No caso dos *tweets*, a análise visa entender a polaridade do sentimento (positivo, negativo ou neutro) dos usuários sobre um determinado assunto, por meio de técnicas de processamento de linguagem natural e aprendizado de máquina.

3.2.3 Algoritmos de Classificação

Os Algoritmos de Classificação desempenham um papel crucial no processo de categorização de conjuntos de dados, utilizando conjuntos de regras e procedimentos definidos para rotular dados com base em suas características intrínsecas. Existem diferentes algoritmos que podem ser usados, como Naive Bayes (WANG; MANNING, 2012), SVM (Support Vector Machines) (NAYAK; NAIK; BEHERA, 2015), Random Forest (BIAU; SCORNET, 2016), entre outros. Cada um destes algoritmos aplica diferentes abordagens e técnicas para realizar a classificação dos *tweets*.

Finalmente, um Modelo Classificador é gerado a partir do treinamento de um algoritmo de classificação com dados rotulados. Este modelo, com base nos padrões aprendidos durante o treinamento, pode classificar novos exemplos, determinando a classe do *tweet*. Uma vez que os modelos estejam treinados e prontos para análise, a aquisição de dados em tempo real torna-se vital para uma predição de possíveis acontecimentos.

3.2.4 BiLSTM — Bidirectional Long Short-Term Memory

BiLSTM, ou Bidirectional Long Short-Term Memory, é uma variação avançada dos tradicionais modelos de Redes Neurais Recorrentes (RNNs), projetada para capturar contextos de longo alcance em sequências de dados. Diferentemente das RNNs padrão, que processam a informação sequencialmente em uma única direção, os modelos BiLSTM consideram as informações em ambas as direções, tanto passadas quanto futuras, em cada ponto da sequência (GRAVES; SCHMIDHUBER, 2005).

Esta característica bidirecional é particularmente útil na análise de textos, como os *tweets*, nos quais o contexto posterior pode fornecer informações essenciais para a compreensão do conteúdo. No âmbito da classificação, os BiLSTMs têm mostrado resultados promissores, especialmente em tarefas que envolvem a compreensão de linguagem natural, podendo detectar nuances e dependências de longa distância nas sequências de texto (XU et al., 2019).

A arquitetura de uma rede BiLSTM consiste em duas camadas LSTM que são dispostas paralelamente. Uma processa a sequência do início ao fim (forward LSTM), enquanto a outra processa a sequência em reverso, do fim para o início (Backward LSTM). As saídas de ambas as camadas são combinadas em cada passo de tempo, permitindo à rede, ter visão completa do contexto durante o processo de classificação. Isso significa que ele não apenas considera a sequência padrão de palavras, mas também sua sequência inversa, possibilitando a captura de contextos tanto anteriores quanto posteriores à palavra em questão.

3.2.5 Word2Vec

O modelo de Word2Vec considera o contexto em que cada palavra aparece para gerar representações vetoriais, ou *word embeddings*. Estes vetores capturam informações semânticas e relacionamentos sintáticos entre as palavras, fornecendo uma visão mais rica e detalhada do texto. No Word2Vec, palavras com significados semelhantes tendem a ter representações vetoriais próximas umas das outras no espaço vetorial, o que permite capturar nuances e similaridades semânticas. Esta técnica é particularmente útil para aplicações que requerem um entendimento mais profundo da linguagem, como a tradução automática ou o reconhecimento de voz. Além disso, os vetores gerados pelo Word2Vec podem ser utilizados como entradas para algoritmos de aprendizado de máquina, facilitando a realização de tarefas complexas de processamento de linguagem natural.

3.3 PRÉ-PROCESSAMENTO DE TEXTO

O Pré-Processamento de Textos é uma área fundamental no desenvolvimento de técnicas e algoritmos para lidar com textos escritos em linguagem natural. Com o crescente volume de dados textuais disponíveis, esta área desempenha um papel importante na extração de informações, compreensão de textos em larga escala e tomada de decisões. Ao automatizar tarefas e identificar padrões, o processamento de textos impulsiona avanços em diversos campos, permitindo o desenvolvimento de soluções inteligentes baseadas em texto.

Uma etapa inicial nesse processo é a Tokenização, que consiste em dividir um texto em unidades significativas, chamadas de tokens. No contexto da análise de sentimentos de *tweets*, a tokenização divide o texto do *tweet* em palavras individuais, emoticons, hashtags ou símbolos de pontuação. Estas unidades semânticas são fundamentais para determinar a polaridade do sentimento expresso no *tweet* e são a base para análises posteriores.

Em seguida, há a Limpeza dos Dados, a qual é uma técnica voltada para a eliminação de informações irrelevantes ou redundantes presentes nos *tweets*, como URLs, menções a outros usuários do Twitter, caracteres especiais e hashtags irrelevantes. O principal objetivo é reduzir o ruído nos dados, descartando elementos que podem impactar negativamente a precisão do modelo ao não contribuir para a análise de sentimentos.

Adicionalmente, a estratégia de remoção de "stop-words" é incorporada ao processo. "Stop-words" são termos frequentemente encontrados no idioma, como artigos, preposições e pronomes, que geralmente não carregam significados distintos em si, mas podem ofuscar a clareza em tarefas como a análise de sentimentos. Vale ressaltar que a remoção destas palavras em um processo de análise de sentimentos pode afetar a acurácia dos resultados devido à perda de contexto, a importância de palavras-chave relevantes, a redução da variedade de vocabulário e a dependência de contexto. Embora a remoção de "stop-words" seja útil em muitos casos, é importante considerar o impacto dessas palavras na interpretação correta do sentimento expresso nos *tweets*. A decisão de remover ou manter estas palavras deve ser baseada nas características dos dados e nos objetivos da análise, visando alcançar resultados mais precisos.

3.4 PROCESSAMENTO DE LINGUAGEM NATURAL — NLP

3.4.1 Word Embeddings

"Word embeddings" são representações vetoriais de palavras que capturam o significado semântico de uma palavra com base em seu contexto de uso em grandes volumes de texto. Estas representações são geradas por modelos como "Word2Vec", "GloVe", "FastText", entre outros. Ao treinar com grandes conjuntos de dados textuais, estes modelos conseguem capturar relações semânticas entre palavras, de modo que palavras com significados semelhantes ou relacionados tendem a ter vetores próximos no espaço vetorial.

Os embeddings de palavras revolucionaram diversas tarefas de processamento de linguagem natural, uma vez que fornecem uma forma mais rica de representar palavras em comparação com abordagens mais antigas. Ao invés de representar cada palavra como um vetor

esparso e único, os embeddings representam palavras em um espaço vetorial contínuo, no qual a posição de cada palavra é determinada pelo seu significado e contexto. Isso permite que algoritmos de aprendizado de máquina operem em dados textuais de uma maneira mais eficiente e intuitiva.

3.4.2 Named Entity Recognition

O Reconhecimento de Entidades Nomeadas, frequentemente abreviado como NER (Named Entity Recognition), é uma tarefa fundamental no processamento de linguagem natural. A técnica se concentra na identificação e classificação de entidades nomeadas em um texto, como nomes de pessoas, organizações, datas, locais, entre outros. Esta técnica pode ser utilizada em várias aplicações, como extração de informações de documentos, indexação de conteúdo, resumo automático de texto, entre outras aplicações. O *NER* utiliza informações contextuais e sintáticas para determinar com precisão quais partes do texto correspondem a entidades nomeadas, contribuindo para uma compreensão mais profunda do conteúdo textual.

3.4.3 Similaridade de Contexto

A Similaridade de Contexto é uma técnica que visa medir o grau de semelhança entre duas palavras, frases ou documentos com base em seu contexto de uso. Em um contexto de processamento de linguagem natural, a similaridade de contexto é frequentemente calculada com base nas representações vetoriais de palavras, como as geradas por modelos de word embeddings, como o Word2Vec mencionado anteriormente. Esta abordagem permite avaliar o quão semanticamente próximas duas palavras ou frases estão, considerando o contexto em que aparecem nos dados de treinamento. A similaridade de contexto é amplamente utilizada em tarefas como recuperação de informações, agrupamento de documentos e até mesmo na detecção de plágio.

3.5 MÉTRICAS DE AVALIAÇÃO DE DESEMPENHO

As métricas de avaliação de desempenho são utilizadas para mensurar a eficácia com que um sistema ou modelo estatístico desempenha tarefas específicas. Em um campo onde a complexidade e a variedade dos dados são vastas, tais métricas são indispensáveis para um diagnóstico preciso e para guiar o aprimoramento contínuo dos modelos. Além da acurácia, é

fundamental considerar métricas complementares — como precisão, recall, F1-score e a matriz de confusão — para obter um panorama mais detalhado e diferenciado do desempenho. A combinação destas métricas provê visões diferentes sobre os resultados, permitindo um melhor entendimento das técnicas estudadas.

3.5.1 Acurácia da Classificação

No espectro de métricas de desempenho, a acurácia da classificação destaca-se por ser uma das mais intuitivas e comumente empregadas. Esta métrica calcula a relação de predições corretas em relação ao total de predições, oferecendo uma visão geral sobre a eficiência de um dado modelo. Porém, sua aplicabilidade deve ser considerada com cautela. Situações com distribuição desigual de classes, em que categorias minoritárias são sub-representadas, podem distorcer a percepção de desempenho que a acurácia proporciona. Por isso, é recomendável que ela seja analisada em conjunto com outras métricas que possam revelar nuances importantes, como o equilíbrio entre a sensibilidade e a especificidade do modelo em identificar cada classe.

3.6 BASES DE DADOS

3.6.1 Sentiment140

A base de dados utilizada para fazer o treinamento do modelo de análise de sentimentos foi o “Sentiment140”. Essa base de dados foi desenvolvida por meio de pesquisas realizadas pela Universidade de Stanford (GO; BHAYANI; HUANG, 2009). Ela possui 1,6 milhão de *tweets* rotulados, os quais são categorizados em sentimentos negativos (rotulados como 0) e positivos (rotulados como 4), sem contemplar sentimentos neutros. Cada entrada inclui informações como ID do *tweet*, data de postagem, consulta e o texto do *tweet*. Este conjunto foi desenvolvido para tarefas de classificação binária de sentimentos e seu objetivo principal é discernir se um *tweet* exibe uma inclinação positiva ou negativa.

3.6.2 Cybertweets

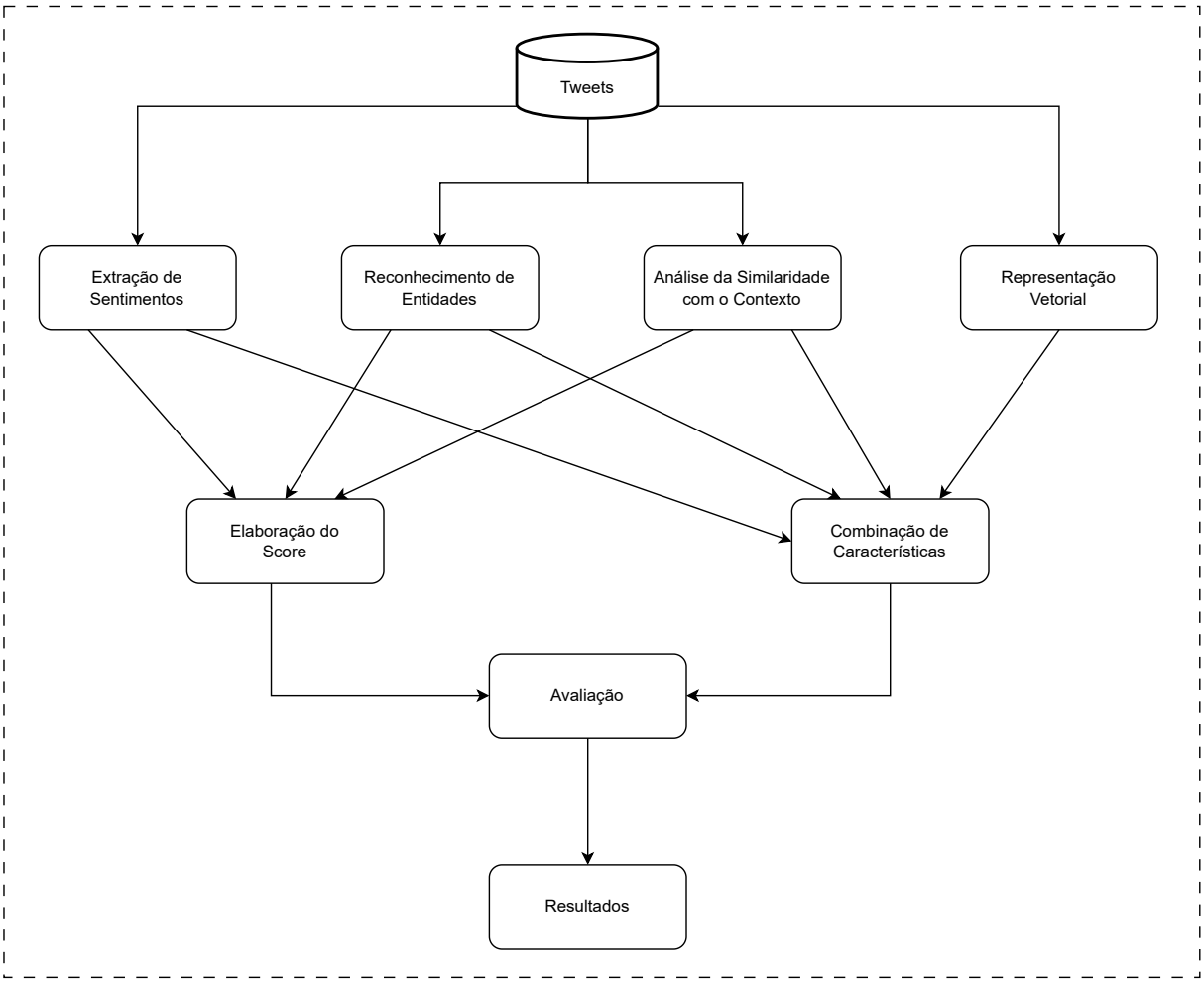
A base de dados “Cybertweets” foi utilizada para validação das propostas apresentadas. Esta base de dados é parte de um projeto maior (BEHZADAN et al., 2018) cujo foco é identificar e descobrir ameaças cibernéticas em tempo real em sistemas computacionais usando

abordagens de aprendizado de máquina aplicadas em conjuntos de dados coletados de diferentes redes online. Nesta base, os *tweets* são rotulados com as seguintes categorias: “Threat”, “Business”, “Irrelevant” e “Don‘t know”. Para efeito do projeto, os textos com o rótulo “Threat” foram categorizados como Ameaça e os demais como Não ameaça.

4 METODOLOGIA

A proposta deste projeto visa utilizar técnicas de processamento de linguagem natural e aprendizado de máquina para auxiliar na identificação de conversas que podem indicar ataques cibernéticos. Neste contexto, duas abordagens diferentes foram propostas: uma delas visa combinar informações como padrões de conversa, sentimentos dos textos e similaridade com palavras recorrentes com a área de segurança da informação para melhorar a acurácia da classificação de textos contendo ameaças cibernéticas. A outra abordagem visa elaborar um score para categorizar um texto como uma possível ameaça mediante um peso aplicado ao tipo de sentimento, entidades envolvidas e similaridade com palavras-chave relacionadas a ameaças cibernéticas. Na Figura 1 é mostrado uma visão geral das propostas enunciadas.

Figura 1 – Pipeline Completo



Fonte: Autores, 2023.

De modo geral, as duas abordagens seguem as seguintes etapas: para fazer a extração de sentimento inicialmente, realiza-se o treinamento de um modelo BILSTM utilizando uma base de dados de *tweets* rotulada. Durante esta fase, os sentimentos dos *tweets* são identificados como positivos ou negativos. Uma vez treinado, o modelo pode ser utilizado para classificar o sentimento de novos *tweets*. Na etapa de "Reconhecimento de Entidades", são identificadas as entidades de cada *tweet* e agrega-se a frequência das mesmas. Durante a "Identificação de Contexto", é mensurado a similaridade dos termos de cada *tweet* com um glossário de termos de Cibersegurança utilizando a técnica de similaridade de cossenos. Nesta fase, tanto os termos do *tweet* quanto do glossário são vetorizados utilizando a técnica de Word2Vec.

Por fim, estes três componentes (Sentimento, Entidade e Contexto) são utilizados nas duas abordagens propostas. Na primeira abordagem eles são incorporados em uma única representação vetorial, a qual é encaminhada para um classificador para fazer a classificação, resultando na predição final. Na segunda abordagem, esses componentes são utilizados em uma métrica para determinar a classe do *tweet* mediante um peso aplicado a cada componente. Espera-se que o modelo consiga aprender os padrões encontrados nos atributos selecionados e associá-los com um determinado rótulo (ameaça ou não ameaça) e, consequentemente, possa classificar novos *tweets* com base nestas características.

4.1 PIPELINE PROPOSTO

Nas próximas subseções serão detalhados os passos de implementação de cada proposta, assim como, dos componentes necessários para a correta aplicação delas.

4.1.1 Proposta 1: Combinação de Características

A primeira abordagem proposta tem por objetivo fazer a combinação de diversas características dos *tweets*, visando criar uma representação que condense as diversas características dos textos para auxiliar na classificação de ameaças virtuais. Primeiramente, são analisados os padrões de conversação presentes nos textos, identificando nuances e características que podem indicar possíveis ameaças. Além disso, é avaliado o sentimento dos textos, com a categorização dos mesmos como positivos ou negativos, por um algoritmo BiLSTM, descrito na seção 4.1.3.

Outro componente fundamental é a identificação de entidades pertinentes, utilizando a técnica de "Named Entity Recognition", descrita na seção 3.4.2. A qual permite identificar palavras, frases ou termos específicos relacionados à segurança cibernética, sendo frequente-

mente associados a ameaças virtuais. Por fim, realiza-se uma análise de similaridade contextual para validar o conteúdo dos textos em relação a um glossário específico de termos de segurança cibernética. Isso garante que as mensagens sejam avaliadas quanto à sua relevância para o contexto da cibersegurança.

Todos esses atributos extraídos dos textos são posteriormente agregados em uma única representação vetorial, combinada com a representação vetorial de cada *tweet*. Esse vetor consolidado é então alimentado em um algoritmo de classificação, como K-NN (K-Nearest Neighbors), Random Forest ou SVM (Support Vector Machine). A combinação das características extraídas dos *tweets* auxilia os algoritmos de classificação a categorizar *tweets* potencialmente mal-intencionados. Como resultado, há um aumento notável na precisão da identificação de conteúdo potencialmente malicioso, fortalecendo a segurança cibernética e a proteção contra ameaças virtuais.

4.1.2 Proposta 2: Elaboração do Score

A segunda estratégia do projeto propõe uma métrica de pontuação ponderada, visando categorizar um *tweet* como ameaça ou não ameaça. Primeiramente, avalia-se o sentimento presente no texto, identificando-os como positivos ou negativos. A estratégia utilizada para realizar a análise de sentimentos está descrita na seção 4.1.3.

Feito isso, buscam-se entidades relevantes, identificando nomes de empresas, organizações, pessoas ou lugares mencionados no texto, uma vez que essas informações muitas vezes desempenham um papel significativo na detecção de ameaças cibernéticas. Os passos desta etapa estão na seção 4.1.4. A proximidade semântica com palavras-chave relacionadas a ameaças cibernéticas também é considerada através da Similaridade de Cossenos descrita na seção 4.1.6 uma vez que, fornece características importantes sobre a natureza do conteúdo observado nos *tweets*.

Por fim, uma camada adicional de característica dos textos é incorporada ao sistema, que se baseia no resultado da predição de um classificador o qual é adicionado e ponderado no cálculo proposto. Esta camada traz uma perspectiva complementar à análise, contribuindo com informações adicionais para a geração da pontuação final. Estes fatores são então agregados em uma média ponderada, resultando na criação de um score que representa a probabilidade de um *tweet* específico indicar uma potencial ameaça no ambiente digital. O score proposto para o cálculo está descrito nas Equações 1 à 4:

$$\text{score} = W_s \times se + W_e \times en + W_c \times co + W_a \times ac \tag{1}$$

$$\text{max\_score} = W_s + W_e + W_c + W_a \tag{2}$$

$$\text{normalized\_score} = \frac{\text{score}}{\text{max\_score}} \tag{3}$$

$$\text{classificação} = \begin{cases} 1 & \text{se normalized\_score} \geq t \\ 0 & \text{caso contrário} \end{cases} \tag{4}$$

Em que:

se : sentimento — W_s : peso sentimento

en : entidades — W_e : peso das entidades

co : contexto com a área de cibersegurança — W_c : peso do contexto

ac : acurácia da classificação — W_a : peso da acurácia

t : threshold

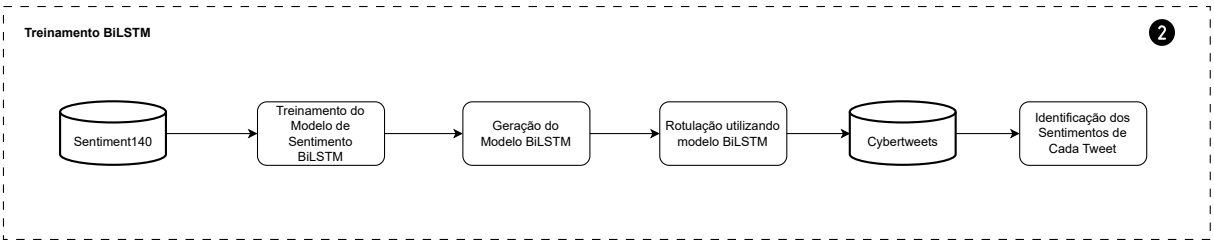
4.1.3 Extração de Sentimentos

Nesta primeira etapa, O BiLSTM, ou Bidirectional Long Short-Term Memory é utilizado para realização da análise de sentimentos nos *tweets*. A capacidade desta arquitetura de considerar a sequência completa — passado, presente e futuro — de um dado é crucial para garantir uma detecção precisa e contextualizada do sentimento. Inicialmente, é utilizado para treinamento a base de textos "Sentiment140", composta por *tweets* previamente rotulados, como sentimentos positivos ou negativos.

A próxima etapa consiste na remoção de ruídos (JURAFSKY; MARTIN, 2014) visando remover informações irrelevantes ou redundantes dos *tweets* que podem afetar negativamente a precisão da análise de sentimentos. Isso pode incluir a remoção de URLs, menções a outros usuários do Twitter, caracteres especiais e *hashtags* irrelevantes. A remoção desses elementos pode ajudar a reduzir o ruído e tornar a análise de sentimento mais precisa, uma vez que as informações irrelevantes não contribuirão para o cálculo do sentimento geral do *tweet*.

Com os dados devidamente preparados, segue-se para o treinamento do modelo BiLSTM. Esse treinamento resulta na geração de um modelo capaz de rotular sentimentos de novos conjuntos de *tweets*. Cada *tweet* é então analisado e rotulado como positivo ou negativo, dependendo do conteúdo identificado pelo modelo. A Figura 2 resume os passos desta etapa.

Figura 2 – Análise de Sentimentos



Fonte: Autores, 2023.

Esta etapa é essencial ao projeto e fornece uma abordagem sistemática para a análise de sentimentos de *tweets*, aproveitando a eficácia da rede BiLSTM e a robustez da base de dados "Sentiment140". Juntos, estes componentes garantem uma rotulação precisa e eficiente dos sentimentos expressos nos *tweets*.

4.1.4 Reconhecimento de Entidades

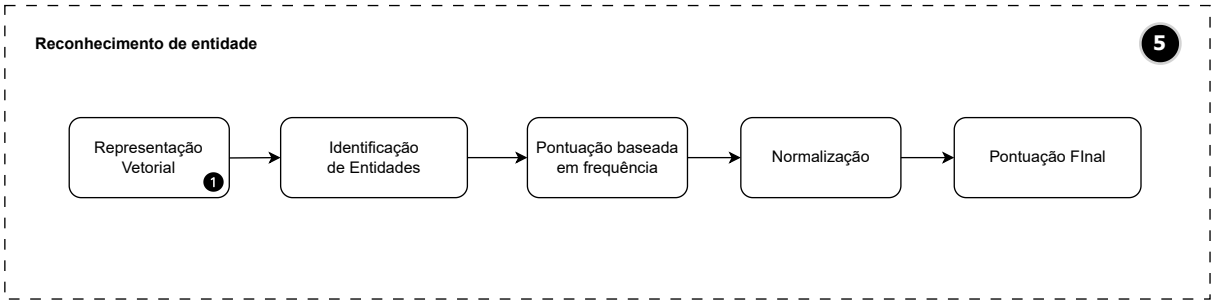
O "Reconhecimento de Entidades" dedica-se a identificar e categorizar entidades nomeadas presentes nos *tweets*, como datas, organizações, pessoas, localizações geográficas, entre outras. Através da técnica de Named Entity Recognition realiza-se o reconhecimento das entidades em cada *tweet*.

Cada entidade reconhecida é posteriormente pontuada com base em critérios de relevância e frequência. A relevância pode ser determinada por fatores como o contexto em que a entidade aparece e sua centralidade na comunicação da mensagem principal do *tweet*. A frequência, por outro lado, pode ser calculada simplesmente pelo número de vezes que a entidade aparece no conjunto de dados analisado. Este processo pode ser desafiador, especialmente ao tratar de linguagem informal e abreviada como a encontrada nos *tweets*, em que gírias, erros ortográficos e uso criativo da linguagem são comuns. O passo a passo desta etapa pode ser visualizado na Figura 3.

Por fim, um processo de normalização é aplicado às pontuações das entidades para garantir que os valores estejam em um intervalo padronizado de 0 até 1. Isso é especialmente útil quando diferentes entidades com diferentes escalas de relevância e frequência precisam

ser comparadas. O resultado é uma pontuação final que sinaliza a importância e relevância da entidade no contexto do *tweet*.

Figura 3 – Reconhecimento de Entidades

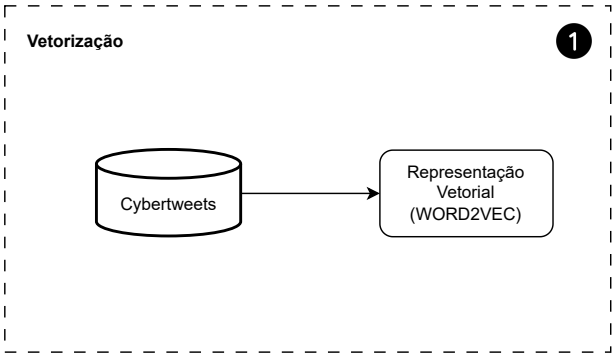


Fonte: Autores, 2023.

4.1.5 Representação Vetorial

A etapa de "Vetorização" desempenha um importante papel para o fluxo da análise. Trata-se de um processo crítico para a transformação de dados textuais em representações numéricas compreensíveis para algoritmos de aprendizado de máquina. Através da técnica "Word2Vec", cada palavra ou frase é mapeada em um vetor numérico. Este vetor é uma representação densa de baixa dimensionalidade que capta a semântica e nuances contextuais das palavras. Este método é fundamental para garantir que as subsequentes operações de análise e aprendizado de máquina possam interpretar e trabalhar eficientemente com o conteúdo dos *tweets*. A Figura 4 mostra o processo de vetorização.

Figura 4 – Vetorização



Fonte: Autores, 2023.

A abordagem fornecida pela técnica "Word2Vec", diferencia-se de métodos tradicionais, como o TF-IDF (Term Frequency-Inverse Document Frequency), por sua capacidade de

capturar contextos e relações semânticas entre palavras de um determinado texto (CHURCH, 2017). Enquanto o TF-IDF se concentra na frequência de uma palavra em um documento em comparação com sua presença em outros documentos, o Word2Vec identifica a similaridade semântica baseando-se na vizinhança contextual da palavra. Isso permite que o Word2Vec capture nuances, analogias e até relações gramaticais entre as palavras. Esta habilidade de reconhecer semelhanças contextuais e semânticas torna o Word2Vec superior ao TF-IDF em muitas tarefas de Processamento de Linguagem Natural, especialmente quando a compreensão do contexto é essencial ao projeto.

4.1.6 Análise da similaridade com o contexto

Posteriormente, após a etapa de vetorização, aplica-se a "Identificação de Contexto". Este processo se destina a identificar a similaridade de um *tweet* com a área de segurança da informação. Para isso, cada *tweet* será transformado em uma representação vetorial, assim como o glossário de palavras-chave e, por fim, comparados através da técnica de similaridade de cossenos.

A técnica de similaridade de cossenos é um método comumente utilizado em análise de textos para medir a similaridade entre dois vetores. Esta técnica baseia-se no cálculo do cosseno do ângulo entre esses vetores. O valor resultante varia de -1 a 1, sendo que um valor de 1 indica que os vetores são idênticos, 0 indica serem ortogonais ou não relacionados e -1 indica serem diametralmente opostos. Matematicamente, a similaridade de cossenos entre dois vetores *A* e *B* pode ser calculada utilizando a Equação 5.

$$\text{similaridade de cossenos}(A, B) = \frac{A \cdot B}{\|A\|\|B\|} \tag{5}$$

Em que:

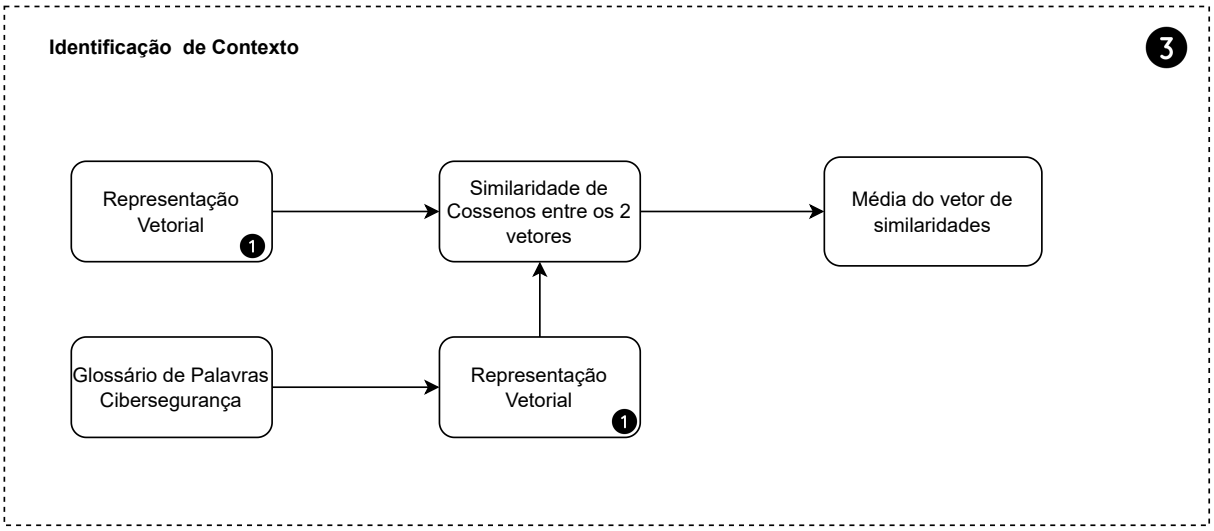
$A \cdot B$ é o produto escalar dos vetores.

$\|A\|$ e $\|B\|$ são as magnitudes (ou normas) dos vetores A e B, respectivamente.

A similaridade de cossenos é especialmente útil em contextos onde a direção dos vetores é mais importante do que sua magnitude. No campo da análise de texto, essa técnica é frequentemente aplicada para comparar documentos, que no cenário deste projeto, compara *tweets* a um glossário de palavras-chave, permitindo identificar o grau de similaridade contextual entre

eles (GUNAWAN; SEMBIRING; BUDIMAN, 2018). O passo a passo desta etapa encontra-se descrito na Figura 5.

Figura 5 – Identificação de Contexto



Fonte: Autores, 2023.

4.1.7 Avaliação Final

A última etapa consiste na classificação do *tweet* em uma ameça ou não ameaça, representando a síntese das características adquiridas ao longo do pipeline. Para a primeira proposta utiliza-se a representação vetorial que combina as diversas informações extraídas dos textos para alimentar um classificador e fazer a predição. Para a segunda proposta, utiliza-se o score para determinar se um *tweet* representa uma ameaça cibernética. A eficácia de ambas as propostas é avaliada em termos de sua precisão, servindo como métrica para a eficácia geral do sistema.

5 RESULTADOS E DISCUSSÕES

Este capítulo é dedicado à exposição dos resultados obtidos através da metodologia proposta, conforme descrito no Capítulo 4. Serão, serão apresentados os dados coletados, as análises realizadas e as evidências encontradas que contribuem para o entendimento e a resposta da questão de pesquisa proposta.

A análise dos resultados é fundamentada em uma abordagem qualitativa, refletindo sobre a aplicabilidade e o impacto dos achados no contexto de identificação de padrões relacionados a ameaças cibernéticas em *tweets*. Será dada especial atenção às relações identificadas entre as variáveis estudadas, bem como à interpretação dos padrões emergentes dos dados. A discussão sobre a relevância dos resultados obtidos é feita em consonância com a literatura consultada, fornecendo uma base sólida para a validação das hipóteses levantadas.

5.1 CONFIGURAÇÕES DOS EXPERIMENTOS

Esta seção descreve as configurações e parâmetros utilizados nos experimentos deste projeto, detalhando características das parametrização dos algoritmos de classificação e os critérios de pontuação empregados.

5.1.1 Algoritmos de Classificação

Para o treinamento de validação dos *tweets* foi feita uma divisão dos dados em conjuntos de treino e teste na proporção de 80/20. Para a classificação foram utilizados os seguintes algoritmos e suas respectivas parametrizações:

a) **SVM (Support Vector Machines):** O modelo foi treinado utilizando as configurações padrão encontrado na biblioteca do sklearn para o kernel e os hiperparâmetros.

b) **Random Forest:** O modelo Random Forest foi configurado com 100 árvores de decisão.

c) **K-NN:** O parâmetro k foi configurado com o valor 5.

d) **Naive Bayes:** Mantido com as configurações padrão da biblioteca.

5.1.2 Baseline de comparação

Para comparar o desempenho das propostas apresentadas nesse projeto foi realizada a comparação dos resultados obtidos com as mesmas em relação ao resultado da classificação direta da base de dados "Cybertweets". Em outras palavras, está sendo avaliado o quanto a combinação de características como entidades, contexto e sentimentos podem ajudar a categorizar um *tweet* em detrimento de apenas treinar um algoritmo de classificação sobre os dados.

5.1.3 Pontuação das Entidades Nomeadas

O sistema de pontuação para o reconhecimento de entidades nomeadas foi definido após uma análise preliminar dos dados. Cada categoria de entidade recebeu um peso específico com base em sua relevância percebida para o contexto de segurança da informação. Estes pesos foram balanceados de modo que a soma totalizasse 1, assegurando uma distribuição proporcional. A Tabela 1 mostra os pesos adotados para cada tipo de entidade.

Tabela 1 – Pesos Atribuídos às Entidades Nomeadas

Entidade	**Peso (%)**
CARDINAL	0,3208
DATE	0,3807
EVENT	0,0002
FAC	0,0003
GPE	0,0392
LANGUAGE	0,0002
LAW	0,0004
LOC	0,0030
MONEY	0,0014
NORP	0,0254
ORDINAL	0,0097
ORG	0,1288
PERCENT	0,0003
PERSON	0,0630
PRODUCT	0,0052
QUANTITY	0,0133
TIME	0,0079
WORK_OF_ART	0,0003

Fonte: Autores, 2023.

Essa estratégia de pontuação foi adotada para refletir a importância relativa de cada tipo de entidade no contexto dos *tweets* analisados, contribuindo para uma análise mais refinada e detalhada.

5.1.4 Word2Vec

Foi utilizado um modelo pré-treinado de Word Embedding que possui vetores de 300 dimensões. O modelo específico foi carregado a partir do conjunto de dados "GoogleNews-vectors-negative300". Este modelo foi treinado no corpus de notícias do Google e são disponibilizados em formatos binários.

5.1.5 Aspectos Técnicos Adicionais

Outras configurações técnicas incluíram:

a) A normalização dos scores de entidades para garantir que os valores finais estivessem no intervalo de 0 a 1.

b) O glossário de termos relacionados à área de Segurança da Informação foi retirado do glossário de palavras de Cibersegurança da empresa Fortinet (CYBER. . . , 2023).

Cada um desses elementos contribuiu para a robustez e a precisão dos experimentos, permitindo uma análise detalhada do conteúdo dos *tweets*.

5.2 ANÁLISE COMPARATIVA DOS RESULTADOS DOS MODELOS DE CLASSIFICAÇÃO

A análise dos resultados obtidos com diferentes algoritmos de aprendizado de máquina fornece percepções significativas sobre a eficácia de várias técnicas na identificação de ameaças cibernéticas em *tweets*. Utilizando a base de dados "Cybertweets", cada algoritmo apresentou uma taxa de acurácia que reflete a sua habilidade em classificar corretamente os *tweets* como ameaças ou não ameaças.

Dentre os modelos de classificação analisados, a abordagem de "Combinação dos atributos utilizando somente o parâmetro de sentimento" permitiu isolar o impacto desse parâmetro na classificação de ameaças cibernéticas em tweets, excluindo a influência de outros atributos. Focando exclusivamente nesse aspecto, a última linha das tabelas de resultados revela a rele-

vância singular do sentimento na eficácia do modelo em comparação com outras estratégias, proporcionando uma compreensão mais profunda da contribuição individual do parâmetro de sentimento na identificação de ameaças cibernéticas.

O algoritmo Support Vector Machine (SVM) obteve um desempenho de base (baseline) de 76,56%, que foi ligeiramente superado ao combinar atributos de sentimento, contexto e entidade, alcançando uma acurácia de 77,89%. Isso sugere que a inclusão de múltiplos atributos pode aumentar a precisão do modelo, embora o aumento seja modesto. Por outro lado, o score ponderado proposto que combinava todos os parâmetros produziu uma tênue melhora, sugerindo que a efetividade de cada atributo pode variar dependendo de como for utilizada.

Tabela 2 – Resultados obtidos utilizando o classificador SVM

Método	**Acurácia**
Baseline utilizando base de dados Cybertweets	76,56%
Combinação dos atributos sentimento, contexto e entidade	**77,89**%
Score ponderando baseline, sentimento, contexto e entidade	76,63%
Combinação dos atributos utilizando somente o parâmetro de sentimento	77,00%

Fonte: Autores, 2023.

Para o algoritmo K-Nearest Neighbors (KNN), a acurácia foi ligeiramente mais alta do que o baseline com 78,08%, e a melhor taxa de acurácia foi obtida com a combinação de atributos, atingindo 78,44%. Este resultado implica que o KNN pode ser mais sensível às nuances dos dados do que o SVM para este conjunto de dados específico.

Tabela 3 – Resultados obtidos utilizando o classificador K-NN

Método	**Acurácia**
Baseline utilizando base de dados Cybertweets	78,08%
Combinação dos atributos sentimento, contexto e entidade	**78,44**%
Score ponderando baseline, sentimento, contexto e entidade	76,63%
Combinação dos atributos utilizando somente o parâmetro de sentimento	78,24%

Fonte: Autores, 2023.

O algoritmo Naive Bayes teve um desempenho inferior ao SVM e KNN com uma acurácia base de 74,52%. Desta vez, a combinação de atributos diminuiu a acurácia para 68,25%, por outro lado, o score ponderado resultou no melhor desempenho para este algoritmo, com 76,63% de acurácia. Isso pode indicar que o Naive Bayes, conhecido por sua simplicidade, pode ser mais eficaz ao utilizar uma abordagem mais direcionada na ponderação de atributos.

Tabela 4 – Resultados obtidos utilizando o classificador Naive Bayes

Método	**Acurácia**
Baseline utilizando base de dados Cybertweets	74,52%
Combinação dos atributos sentimento, contexto e entidade	68,25%
Score ponderando baseline, sentimento, contexto e entidade	**76,63**%
Combinação dos atributos utilizando somente o parâmetro de sentimento	67,86%

Fonte: Autores, 2023.

Finalmente, o algoritmo Random Forest apresentou a acurácia base mais alta entre os baselines avaliados, com 79,25%. A combinação dos atributos de sentimento, contexto e entidade teve um leve aumento para 79,58%, e surpreendentemente, o uso exclusivo do parâmetro de sentimento resultou na acurácia mais alta de 79,59%. Isso destaca o Random Forest como uma escolha robusta para o processamento de dados complexos e sugere que o sentimento pode ser um indicador mais forte do que se presumia anteriormente para a detecção de ameaças em *tweets*.

Tabela 5 – Resultados obtidos utilizando o classificador Random Forest

Método	**Acurácia**
Baseline utilizando base de dados Cybertweets	79,25%
Combinação dos atributos sentimento, contexto e entidade	79,58%
Score ponderando baseline, sentimento, contexto e entidade	76,63%
Combinação dos atributos utilizando somente o parâmetro de sentimento	**79,59**%

Fonte: Autores, 2023.

Esses resultados apontam para a importância de experimentar diferentes combinações de atributos e explorar o desempenho de diferentes algoritmos de classificação para identificar a configuração mais eficaz para a classificação de segurança cibernética, ressaltando a natureza complexa da análise de dados de mídia social para ameaças cibernéticas.

5.3 ANÁLISE EXPLORATÓRIA DAS CARACTERÍSTICAS DA BASE DE DADOS

A análise exploratória da base de dados é um passo crucial para compreender a natureza e a distribuição dos dados coletados, especialmente no que diz respeito à cibersegurança. Esta seção planeja examinar detalhadamente a composição da base de dados, investigando a frequência e a natureza das entidades nomeadas, a semelhança com o glossário de cibersegurança e a

expressão de sentimentos nos *tweets*. Esta análise é fundamental para identificar padrões, anomalias e tendências que podem influenciar a eficácia dos modelos aplicados.

Ao realizar uma inspeção detalhada, procurou-se entender como os diferentes componentes dos dados contribuem para a identificação de ameaças potenciais. A análise das entidades nomeadas, por exemplo, fornece percepções significativas sobre os aspectos mais mencionados nos *tweets*, enquanto a comparação com o glossário de cibersegurança ajuda a contextualizar nossos achados. Ademais, a investigação dos sentimentos expressos nos *tweets* pode revelar padrões emocionais associados a comunicações de ameaças. As subseções a seguir detalham essas análises.

5.3.1 Exploração de Entidades Nomeadas

Os experimentos realizados para investigar as características das ameaças cibernéticas em *tweets* revelaram alguns padrões sobre a natureza dos dados de segurança cibernética. Ao examinar as entidades nomeadas presentes nos textos, identificaram-se traços característicos diferenciadores entre *tweets* que apresentam ameaças em detrimento daqueles que não as têm.

Em relação às entidades, há uma predominância de datas (DATE) em *tweets* com ameaças, representando 38,07%, em comparação com 24,35% em *tweets* sem ameaças, vide Tabela 6. Isso pode indicar que *tweets* mal-intencionados tendem a citar eventos específicos no tempo com mais frequência, talvez para se referir a campanhas de ataques ou vulnerabilidades recentemente descobertas. Outro aspecto notável é a menor presença de entidades geopolíticas (GPE) em *tweets* ameaçadores (3,92%) em comparação com os não ameaçadores (10,11%), o que pode sugerir que as ameaças cibernéticas são menos propensas a mencionar locais específicos ou são mais globais em sua natureza.

Entretanto, em ambos os tipos de *tweets*, organizações (ORG) e pessoas (PERSON) são frequentemente mencionadas, com 12,88% e 6,30% respectivamente em *tweets* com ameaças, contra 14,62% e 10,86% em *tweets* sem ameaças. Isso reflete a tendência de discussões sobre segurança cibernética girarem em torno de entidades corporativas e individuais, possivelmente destacando alvos de ataques ou envolvimento de indivíduos em atividades de segurança.

Curiosamente, há uma maior porcentagem de entidades NORP (grupos nacionais, religiosos ou políticos) em *tweets* não ameaçadores (9,76%) em comparação com aqueles com ameaças (2,54%), o que pode indicar que conversas benignas tendem a envolver mais discussões sobre identidades coletivas. A presença de entidades como CARDINAL, as quais identificam números, é significativamente mais alta em *tweets* ameaçadores, indicando possivelmente

a inclusão de informações técnicas ou quantitativas específicas relacionadas a ameaças cibernéticas.

Tabela 6 – Distribuição percentual de Entidades

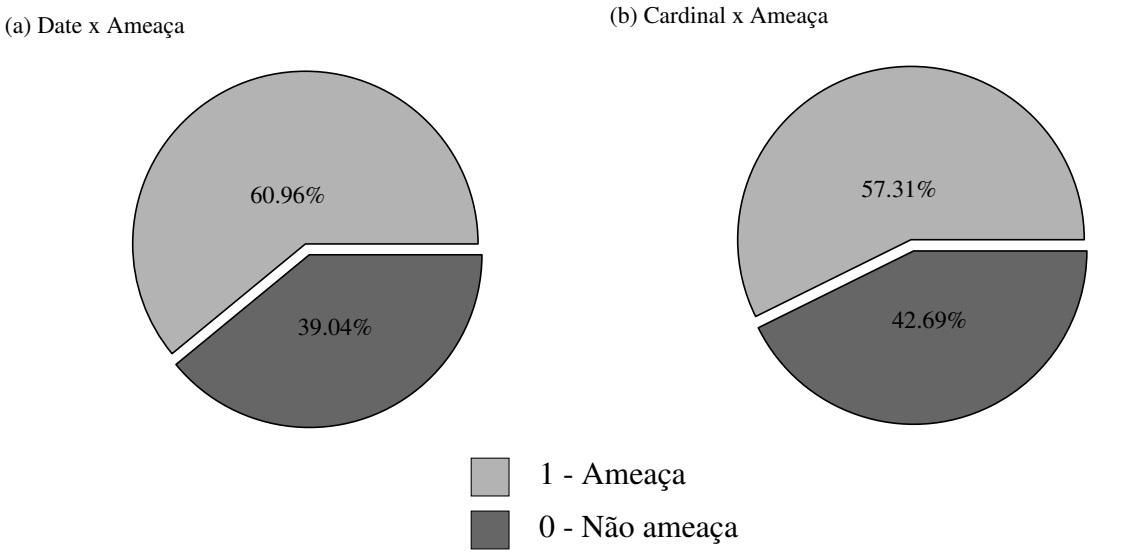
Categoria	**Ameaças (%)**	**Não ameaças (%)**
CARDINAL	32,08	23,86
DATE	38,07	24,35
EVENT	0,02	0,18
FAC	0,03	0,08
GPE	3,92	10,11
LANGUAGE	0,02	0,03
LAW	0,04	0,06
LOC	0,30	0,69
MONEY	0,14	0,23
NORP	2,54	9,76
ORDINAL	0,97	2,09
ORG	12,88	14,62
PERCENT	0,03	0,09
PERSON	6,30	10,86
PRODUCT	0,52	0,65
QUANTITY	1,33	0,50
TIME	0,79	1,78
WORK_OF_ART	0,03	0,05

Fonte: Autores, 2023.

Direcionando o foco para uma análise quantitativa da Base “Cybertweets”, ao olhar para entidades CARDINAL (números) e DATE (datas), há uma inclinação mais acentuada para a presença em *tweets* ameaçadores, com 57,3% e 60,96% respectivamente. Esta particularidade pode indicar que números e datas são mais prevalentes em comunicações que contêm ameaças, talvez devido à especificidade requerida ao discutir eventos de segurança cibernética ou ataques.

Em virtude dos padrões observados, destaca-se a complexidade e a sutileza necessárias na aplicação do processamento de linguagem natural para a identificação de ameaças cibernéticas em mídias sociais. A precisão do modelo em diferenciar entre conteúdo ameaçador e não ameaçador poderia ser aprimorada em função da análise de padrões mais específicos nas entidades e sentimentos revelados.

Figura 6 – Distribuição de entidades DATE e CARDINAL



Fonte: Autores, 2023.

5.3.2 Mapeamento de Sentimentos

Ao examinar sentimentos, percebe-se que *tweets* rotulados como ameaças possuem um equilíbrio entre sentimentos positivos (51,61%) e negativos (48,39%), sugerindo que a polaridade do sentimento, por si só, não é um indicador confiável de conteúdo malicioso.

Apesar destes resultados, a análise de sentimentos representa um atributo importante na identificação de *tweets* potencialmente ameaçadores e a intensidade do sentimento expresso pode ser um indicador útil do conteúdo ameaçador, até um certo ponto. Isso pode ser particularmente interessante para refinar algoritmos de detecção de ameaças cibernéticas, permitindo-lhes considerar a nuance emocional como um fator na avaliação de riscos potenciais.

Tabela 7 – Distribuição percentual de ameaças e não ameaças

Categoria	**Ameaças (%)**	**Não ameaças (%)**
Positivo	51,61	58,40
Negativo	48,39	41,60

Fonte: Autores, 2023.

5.3.3 Análise do Contexto

Para a abordagem proposta neste projeto de pesquisa, a similaridade de contexto pode ser interpretada como a proximidade de um texto em relação aos termos mais recorrentes na área de segurança cibernética. Textos que possuem maior proximidade com o glossário podem ser considerados como tendo um contexto mais similar ao de discussões de segurança cibernética. Neste âmbito, o "uso normal" de termos específicos do setor se refere à sua aplicação em contextos típicos ou educacionais, como discussões informativas sobre segurança de TI, sem intenções maliciosas. Por outro lado, o "uso anormal" caracteriza-se pelo emprego desses mesmos termos em contextos atípicos ou suspeitos, que podem sugerir atividades mal-intencionadas, como a discussão de vulnerabilidades com intuito de exploração. Portanto, a eficácia desta técnica de similaridade não reside apenas em identificar a presença de termos relevantes, mas em discernir cuidadosamente entre seus usos normais e anormais para detectar possíveis ameaças cibernéticas.

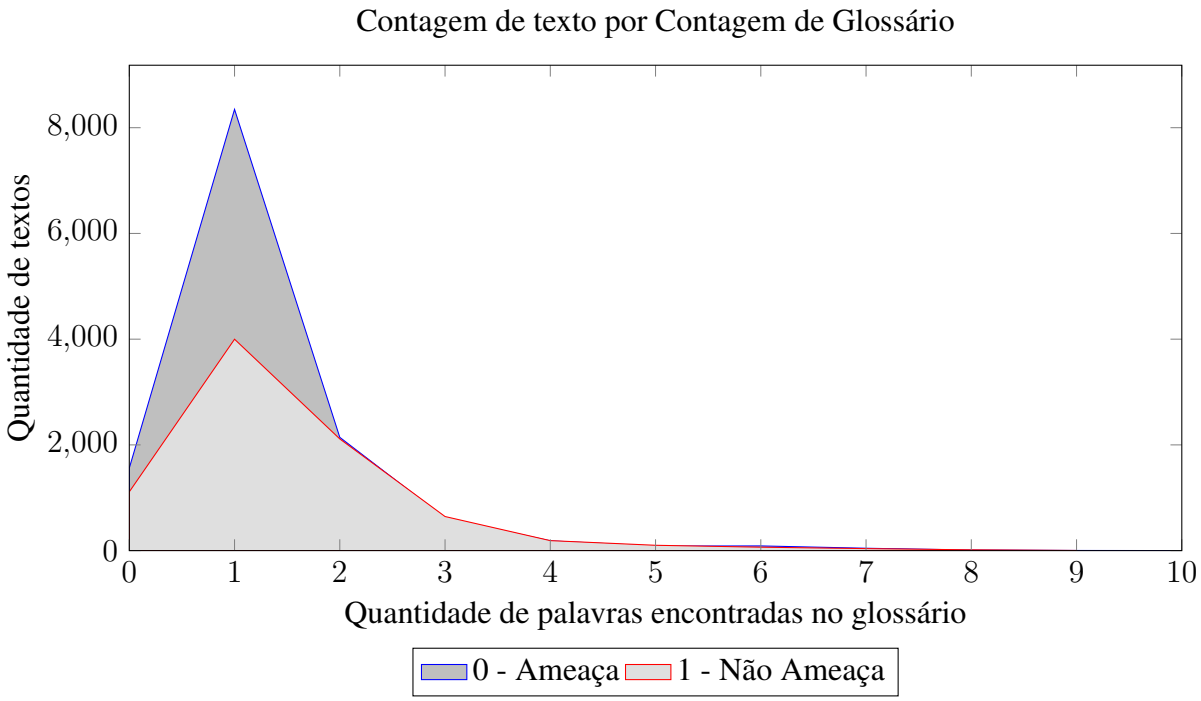
Estendendo a aplicabilidade desse projeto para outras áreas do conhecimento torna-se evidente a vantagem de utilização da técnica de análise de contexto. Isso é possível porque, uma vez que temos uma representação vetorial dos principais termos de um dado tema e a representação vetorial do texto em análise, torna-se matematicamente factível mensurar a similaridade destas representações. Em especial, se estas representações forem geradas utilizando técnicas de *Word Embedding* tem-se a vantagem de trazer a probabilidade de ocorrência de cada palavra condensado na representação vetorial, sendo assim, preservando propriedades semânticas dos textos.

No gráfico da Figura 7, observamos uma distribuição que indica a frequência com que os termos do glossário aparecem nos textos. A curva azul, que representa os textos considerados como "Ameaça", mostra um pico agudo para textos contendo uma quantidade baixa de palavras do glossário, seguido por uma diminuição rápida conforme aumenta o número de palavras do glossário. A maioria dos textos contém poucos termos do glossário, como mostrado pelo pico agudo para um baixo número de termos e o declínio rápido à medida que o número de termos aumenta. Isso sugere que, embora o glossário de segurança cibernética seja relevante, a maioria dos textos não utiliza tais termos de forma abundante. Este padrão pode indicar que a discussão sobre ameaças cibernéticas muitas vezes não emprega terminologia altamente técnica, ou que as ameaças são discutidas em um contexto mais geral ou com linguagem informal.

Essa análise ressalta a complexidade de detectar ameaças cibernéticas com base em linguagem textual. A presença de termos técnicos é apenas um dos vários fatores a serem conside-

rados, e métodos mais sofisticados de NLP podem ser necessários para auxiliar na compreensão do contexto e a intenção oculta dos textos para identificar ameaças verdadeiras eficazmente.

Figura 7 – Contagem de texto por Contagem de Glossário



Fonte: Autores, 2023.

5.4 REPRODUTIBILIDADE

Para assegurar a reprodutibilidade e transparência dos experimentos conduzidos neste trabalho, todos os códigos-fonte, conjuntos de dados e materiais relacionados foram documentados e disponibilizados publicamente em um repositório do GitHub.[1], onde os interessados podem encontrar instruções sobre como configurar o ambiente de desenvolvimento, instalar as dependências necessárias e executar os scripts para reproduzir os resultados apresentados ou comparar com futuras abordagens propostas.

[1] O repositório no GitHub pode ser acessado através do link: https://github.com/akajhon/TweetThreatDetect

6 CONCLUSÃO

Os resultados detalhados no Capítulo 5 evidenciam a eficácia do uso de técnicas de Processamento de Linguagem Natural (NLP) na identificação de padrões relacionados a ameaças cibernéticas em tweets. A distribuição percentual de entidades em tweets categorizados como ameaças e não ameaças fornece conclusões relevantes sobre as características distintas dessas comunicações. Notavelmente, entidades como datas ('DATE') foram significativamente mais prevalentes em tweets contendo ameaças, com 38,07%, comparado a 24,35% em tweets não ameaçadores. Semelhantemente, entidades numéricas ('CARDINAL') apareceram em 32,08% dos tweets ameaçadores, contra 23,86% nos não ameaçadores.

Por outro lado, categorias como "NORP" (identidades coletivas nacionais, religiosas ou políticas) e "PERSON" (pessoas) foram mais comuns em tweets benignos, com 9,76% e 10,86%, respectivamente, em comparação a 2,54% e 6,30% em tweets ameaçadores. Isso indica uma tendência de discussões benignas para focar mais em tópicos relacionados a identidades coletivas e indivíduos.

Essas descobertas, como discutido na seção 5.3.1, demonstram a importância de analisar a natureza das entidades mencionadas nos tweets para uma compreensão mais profunda dos padrões de comunicação relacionados à segurança da informação e ameaças cibernéticas.

A análise comparativa dos resultados dos modelos de classificação, que inclui SVM, K-Nearest Neighbors (KNN), Naive Bayes e Random Forest, realça a importância de experimentar com diferentes combinações de atributos e algoritmos. Nos experimentos, observou-se que as abordagens propostas superaram consistentemente os resultados do baseline de comparação. Por exemplo, utilizando o classificador Random Forest, a combinação dos atributos sentimento, contexto e entidade resultou em uma acurácia de 79,58%, uma melhoria marginal em comparação ao baseline que foi de 79,25%. Notavelmente, a maior acurácia alcançada foi de 79,59% quando focando apenas no parâmetro de sentimento, um aumento de 0,34% em relação ao baseline.

Da mesma forma, com o classificador Naive Bayes, o método que ponderava baseline, sentimento, contexto e entidade atingiu a maior acurácia de 76,63%, um aumento significativo de 2,11% em relação ao baseline de 74,52%. Já o classificador K-NN mostrou um aumento na acurácia para 78,44% ao combinar sentimentos, contexto e entidade, um incremento de 0,36% em relação ao baseline de 78,08%. Por fim, o classificador SVM teve um aumento de acurácia

para 77,89% com a combinação dos mesmos atributos, representando uma melhoria de 1,33% sobre o baseline de 76,56%.

Estes resultados numéricos evidenciam o impacto positivo da escolha e combinação de diferentes características dos textos para aprimorar a categorização. Este ponto é detalhado na Seção 5.2.

Além disso, a análise de sentimentos revelou um equilíbrio quase igual entre sentimentos positivos e negativos nos tweets identificados como ameaças, desafiando a noção de que a polaridade do sentimento é um indicador confiável de conteúdo mal-intencionado, assim como evidenciado na seção 5.3.2. Especificamente, 51,61% dos tweets classificados como ameaças exibiram sentimentos positivos, enquanto 48,39% apresentaram sentimentos negativos. Em contraste, os tweets não ameaçadores tiveram uma distribuição de 58,40% para sentimentos positivos e 41,60% para negativos.

Essas descobertas destacam a desafio de classificar ameaças cibernéticas com base em linguagem textual e a necessidade de métodos mais sofisticados como NLP para auxiliar na compreensão das características dos texto, contexto e intenções ocultas dos textos. A eficácia de utilizar a similaridade de contexto e a análise quantitativa das entidades nomeadas reforça a ideia de que a presença de termos técnicos é apenas um dos vários fatores a serem considerados na identificação de ameaças verdadeiras, conforme explorado em maior detalhe na seção 5.3.3.

6.1 IMPLICAÇÕES PRÁTICAS

As descobertas deste estudo têm implicações diretas, na prática de monitoramento de segurança cibernética. As técnicas desenvolvidas podem ser integradas em ferramentas de segurança digital para identificar automaticamente comunicações potencialmente maliciosas. Além disso, a metodologia pode ser adaptada para monitorar e analisar a opinião pública sobre questões de segurança, fornecendo às organizações observações perspicazes para aprimorar suas estratégias de comunicação e políticas de segurança.

6.2 LIMITAÇÕES DO ESTUDO

Apesar dos resultados encontrados, este estudo possui algumas limitações. A dependência de dados específicos de *tweets* pode não capturar completamente a variedade de comunicações em segurança cibernética. Além disso, a natureza dinâmica das ameaças cibernéticas exige constante atualização dos modelos para manter sua eficácia. Tais limitações destacam a neces-

sidade de um escopo mais amplo de dados e a atualização contínua dos modelos em estudos futuros.

6.3 TRABALHOS FUTUROS

Dadas as observações e conclusões deste estudo, deseja-se fazer uma continuação da pesquisa na área de análise de sentimentos e reconhecimento de entidades nomeadas para aprimorar ainda mais a detecção de ameaças cibernéticas. Explorar a aplicação de técnicas de aprendizado profundo e o desenvolvimento de modelos mais complexos, que podem captar nuances mais sutis em textos, pode ser um caminho promissor. Além disso, a expansão dos conjuntos de dados para incluir uma variedade maior de fontes online e a integração de informações de contexto mais amplas podem enriquecer a análise e a precisão dos modelos.

Futuros trabalhos também podem se concentrar em melhorar a interpretação dos resultados dos modelos, utilizando visualizações interativas e interfaces mais intuitivas. Isso não só facilitaria a compreensão dos dados por parte dos analistas de segurança, mas também ajudaria a disseminar o conhecimento para um público mais amplo, incluindo profissionais que não são especialistas em dados.

Além disso, é importante considerar a aplicabilidade das metodologias propostas em áreas além da cibersegurança. As técnicas de Processamento de Linguagem Natural e análise de sentimentos desenvolvidas neste estudo têm potencial para serem adaptadas e aplicadas em diversos outros campos. Por exemplo, no marketing digital, essas metodologias podem ser utilizadas para analisar a percepção e o sentimento do consumidor em relação a marcas e produtos, fornecendo conclusões relevantes para estratégias de marketing e desenvolvimento de produtos. Da mesma forma, no monitoramento de saúde pública, a análise de sentimentos e reconhecimento de entidades em mídias sociais pode ajudar a identificar tendências emergentes de saúde ou preocupações públicas, contribuindo para uma resposta mais rápida e eficaz a crises de saúde. Portanto, a expansão e adaptação dessas técnicas para outros domínios de pesquisa oferecem um caminho promissor para futuros projetos, ampliando o impacto e a relevância das descobertas deste estudo.

O objetivo é que este projeto não represente um ponto final, mas sim um elo numa cadeia contínua de descobertas, inspirando e capacitando outros pesquisadores a avançar ainda mais nas fronteiras do conhecimento na interseção entre Processamento de Linguagem Natural e segurança cibernética.

REFERÊNCIAS

ALSAEEDI, A.; KHAN, M. A study on sentiment analysis techniques of twitter data. **International Journal of Advanced Computer Science and Applications**, v. 10, p. 361–374, 2 2019. Disponível em: <https://www.researchgate.net/profile/Abdullah-Alsaeedi/publication/331411860_A_Study_on_Sentiment_Analysis_Techniques_of_Twitter_Data/links/5c78175ba6fdcc4715a3d664/A-Study-on-Sentiment-Analysis-Techniques-of-Twitter-Data.pdf>.

BEHZADAN, V. et al. **Corpus and Deep Learning Classifier for Collection of Cyber Threat Indicators in Twitter Stream**. Seattle, WA, USA: [s.n.], 2018.

BIAU, G.; SCORNET, E. A random forest guided tour. Paris, France, 2016. Disponível em: <https://www.researchgate.net/publication/284219299_A_Random_Forest_Guided_Tour>.

CHURCH, K. W. Word2vec. **Natural Language Engineering**, Cambridge University Press, v. 23, n. 1, p. 155–162, 2017.

CYBER Glossary. [S.l.]: Fortinet, 2023. <https://www.fortinet.com/resources/cyberglossary>.

DEB, A.; LERMAN, K.; FERRARA, E. Predicting cyber-events by leveraging hacker sentiment. **Information**, v. 9, n. 11, 2018. ISSN 2078-2489. Disponível em: <https://www.mdpi.com/2078-2489/9/11/280>.

DIONíSIO, N. et al. Cyberthreat detection from twitter using deep neural networks. p. 1–8, 2019.

GO, A.; BHAYANI, R.; HUANG, L. **Twitter sentiment classification using distant supervision**. [S.l.]: CS224N Project Report, Stanford, 2009.

GRAVES, A.; SCHMIDHUBER, J. Framewise phoneme classification with bidirectional lstm and other neural network architectures. **Neural Networks**, v. 18, n. 5, p. 602–610, 2005. ISSN 0893-6080. IJCNN 2005. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0893608005001206>.

GUNAWAN, D.; SEMBIRING, C. A.; BUDIMAN, M. A. The implementation of cosine similarity to calculate text relevance between two documents. **Journal of Physics: Conference Series**, IOP Publishing, v. 978, n. 1, p. 012120, mar 2018. Disponível em: <https://dx.doi.org/10.1088/1742-6596/978/1/012120>.

GUTTMAN, B.; ROBACK, E. **An Introduction to Computer Security: the NIST Handbook**. [S.l.]: Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD, 1995.

HERNANDEZ, A. et al. Security attack prediction based on user sentiment analysis of twitter data. In: **Proceedings - 2016 IEEE International Conference on Industrial Technology, ICIT 2016**. Estados Unidos: Institute of Electrical and Electronics Engineers Inc., 2016. (Proceedings of the IEEE International Conference on Industrial Technology), p. 610–617.

HERNANDEZ-SUAREZ, A. et al. Social sentiment sensor in twitter for predicting cyber-attacks using ℓ_1 regularization. **Sensors (Basel)**, v. 18, n. 5, p. 1380, 4 2018.

INDUSTRIALCYBER. **LockBit, Conti, SunCrypt, Alphv, Blackcat, Hive Emerge as Key RAAS Groups Targeting Healthcare and Public Health Sector**. 2020. Disponível em: <https://tinyurl.com/industrialcyber>.

JURAFSKY, D.; MARTIN, J. H. Speech and language processing. Pearson Education, Upper Saddle River, NJ, 2014.

NAYAK, J.; NAIK, B.; BEHERA, H. S. A comprehensive survey on support vector machine in data mining tasks: Applications & challenges. **International Journal of Database Theory and Application**, Odisha, India, v. 8, n. 1, p. 169–186, 2015. Disponível em: <https://www.researchgate.net/publication/291748219_A_comprehensive_survey_on_support_vector_machine_in_data_mining_tasks_Applications_challenges>.

NUNES, E. et al. Darknet and deepnet mining for proactive cybersecurity threat intelligence. p. 7–12, 2016.

SAKAKI, T.; OKAZAKI, M.; MATSUO, Y. Earthquake shakes twitter users: Real-time event detection by social sensors. In: **Proceedings of the 19th International Conference on World Wide Web**. New York, NY, USA: Association for Computing Machinery, 2010. (WWW '10), p. 851–860. Disponível em: <https://doi.org/10.1145/1772690.1772777>.

SMITH, M. Where will the security community turn, if not twitter? **Cybersecurity Dive**, 2022. Acesso em: 23 de mar. de 2023. Disponível em: <https://www.cybersecuritydive.com/news/twitter-security-community/637002/>.

STALLINGS, W. **Criptografia e segurança de redes: princípios e práticas**. [S.l.]: Pearson Education, 2017.

TRIPWIRE. **Windows Zero-Day LPE Flaw Exploited in the Wild**. 2018. Disponível em: <https://threatpost.com/windows-zero-day-lpe/144976/>.

WANG, S.; MANNING, C. D. Baselines and bigrams: Simple, good sentiment and topic classification. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers**. Stroudsburg, PA, 2012. p. 90–94. Disponível em: <https://aclanthology.org/P12-2018.pdf>.

XU, G. et al. Sentiment analysis of comment texts based on bilstm. **IEEE Access**, v. 7, p. 51522–51532, 2019.

ZHANG, Y. et al. Twitter trends manipulation: A first look inside the security of twitter trending. **IEEE Transactions on Information Forensics and Security**, v. 12, p. 144–156, 2017.