UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Um estudo de caso sobre o mapeamento entre NIST CSF 2.0 e a diretiva NIS2 baseada em análise de similaridade

Ricardo Luiz da Silva

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Um estudo de caso sobre o mapeamento entre NIST CSF 2.0 e a diretiva NIS2 baseada em análise de similaridade

Ricardo Luiz da Silva

USP - São Carlos

2024

Ricardo Luiz da Silva

Um estudo de caso sobre o mapeamento entre NIST CSF 2.0 e a diretiva NIS2 baseada em análise de similaridade

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Dr. Rafael Geraldeli Rossi

Versão Original

USP - São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

da Silva, Ricardo Luiz
d111e Um estudo de caso sobre o mapeamento entre NIST CSF 2.0 e a diretiva NIS2 baseada em análise de similaridade / Ricardo Luiz da Silva; orientador Rafael Geraldeli Rossi. -- São Carlos, 2024.
82 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Machine Learning. 2. Open-set classification. 3. Few-shot learning. 4. Transfer learning. I. Geraldeli Rossi, Rafael, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Ricardo Luiz da Silva

A case study on the mapping between NIST CSF 2.0 and the NIS2 directive based on similarity analysis.

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo – ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Dr. Rafael Geraldeli Rossi

Original version

USP - São Carlos

2024

ERRATA

Errata			
Folha	Linha	Onde se lê	Leia-se

DEDICATÓRIA

Dedico este trabalho a todos os profissionais de cibersegurança que diariamente dedicam tempo para manter seguro o ciberespaço.

AGRADECIMENTOS

Um bem-haja a todos que, direta e indiretamente, contribuem para o meu aprendizado.

Agradeço a minha família, esposa e filha por todo o suporte nesta jornada e nos desafios que nos rodeiam diariamente.

Agradeço especialmente ao meu orientador, Prof. Dr. Rafael Rossi, por todo o suporte durante o desenvolvimento deste trabalho.

Agradeço à Professora Solange Rezende pela dedicação e incentivos em garantir a minha participação neste curso e da conclusão deste trabalho.

EPÍGRAFE

Os meus alter egos vivem num enlace heurístico dinâmico de zeros e uns entrelaçados numa Web Semântica, sujeitos a uma sobreposição de estados constante.

My alter egos live in a dynamic heuristic interconnection of intertwined zeros and ones within a Semantic Web, subject to a constant superposition of states.

Ricardo Luiz da Silva (nov. 2024)

RESUMO

Da Silva, R. L. Um estudo de caso sobre o mapeamento entre NIST CSF 2.0 e a diretiva NIS2 baseada em análise de similaridade. 2024. 82 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

A conformidade regulatória e a adoção de boas práticas de governança e gestão de riscos são essenciais para empresas que operam em ambientes dinâmicos e altamente regulamentados, especialmente em setores críticos. No entanto, a necessidade de inovação rápida em ambientes ágeis frequentemente entra em conflito com processos rigorosos de conformidade e auditoria. Para evitar sanções, melhorar a governança e mitigar riscos de cibersegurança, as organizações devem assegurar que suas políticas e controles de segurança estejam alinhados com normas globais. Dada esta lacuna, o objetivo deste trabalho de curso é avaliar a aplicação de técnicas de classificação de texto para ajudar na validação de políticas de cibersegurança e mapeamento dos controles de segurança com as necessidades legais orientadas pelo padrão NIST 2.0 CSF e a diretiva NIS2. O estudo analisa o desempenho dessas técnicas no alinhamento entre as subcategorias do NIST CSF 2.0 e os requisitos da diretiva NIS2. Para isso, foram utilizadas técnicas de mineração de texto suportadas por técnicas de conjunto aberto (*open-set classification*), aprendizagem com poucos casos (*few-short learning*) e aprendizagem por transferência (*transfer learning*) para identificar se textos pertencem a classes semânticas - extraídas de uma análise comparativa dos documentos normativos e das categorias definidas por estes padrões - utilizando análise de similaridade cosseno aplicada a vetores gerados por modelos de Processamento de Linguagem Natural (PLN). São comparados métodos para a geração de representações baseadas em frequência de palavras (*Bag-of-Words* com TF-IDF LSA) e *embeddings* semânticos (como BERT, ERNIE, XLNET e E5 Large), conhecidos por oferecerem um melhor nível de detalhe devido ao extenso treinamento em pares. Os resultados procuram identificar a abordagem que apresenta a classificação mais correta, contribuindo para soluções automatizadas que ajudem organizações a gerenciar riscos, adaptar-se às regulamentações de segurança da informação e cibersegurança, e alcançar conformidade contínua de maneira ágil e sustentável.

Palavras-chave: Conformidade regulatória 1; Governança 2; Gestão de riscos 3; Cibersegurança 4; Classificação de textos 5; Similaridade cosseno 6; Processamento de Linguagem Natural (PLN) 7; *Embeddings* semânticos 8; TF-IDF 9; SBERT 10; E5 Large 11; NIST CSF 2.0 12; Diretiva NIS2 13; Automação 14; Eficiência 15; Sustentabilidade 16; Conjunto aberto 17; Aprendizagem com poucos casos 18; Aprendizagem por transferência 19.

ABSTRACT

Da Silva, R. L. A case study on the mapping between NIST CSF 2.0 and the NIS2 directive based on similarity analysis. 2024. 82 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

Regulatory compliance, good governance, and risk management practices are essential for companies operating in dynamic and highly regulated environments, especially in critical sectors. However, the need for rapid innovation in agile environments often conflicts with rigorous compliance and audit processes. Organizations must ensure that their security policies and controls are aligned with global standards to avoid penalties, improve governance, and mitigate cybersecurity risks. Given this gap, this coursework aims to evaluate the application of text classification techniques to assist in the validation of cybersecurity policies and the mapping of security controls to legal requirements guided by the NIST 2.0 CSF and the NIS2 directive. The study analyzes the performance of these techniques in aligning the subcategories of the NIST CSF 2.0 framework with the requirements of the NIS2 directive. For this purpose, text mining techniques supported by open-set classification, few-shot learning, and transfer learning were used to identify whether texts belong to semantic classes - extracted from a comparative analysis of normative documents and the categories defined by these standards - using cosine similarity analysis applied to vectors generated by Natural Language Processing (NLP) models. Methods for generating representations based on word frequency (Bag-of-Words with TF-IDF LSA) and semantic *embeddings* (such as BERT, ERNIE, XLNET, and E5 Large) are compared, the latter known for offering a better level of detail due to extensive training on pairs. The results seek to identify the most accurate classification approach, contributing to automated solutions that help organizations manage risks, adapt to information security and cybersecurity regulations, and achieve continuous compliance in an agile and sustainable manner.

Keywords: Regulatory compliance 1; Governance 2; Risk management 3; Cybersecurity 4; Text classification 5; Cosine similarity 6; Natural Language Processing (NLP) 7; Semantic *embeddings* 8; TF-IDF 9; SBERT 10; E5 Large 11; NIST CSF 2.0 12; NIS2 Directive 13; Automation 14; Efficiency 15; Sustainability 16; Open-set classification 17; Few-shot learning 18; Transfer learning 19.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Elemento opcional. É composto de uma relação alfabética das abreviaturas e siglas utilizadas no texto seguido do seu significado.

ABNT	–	Associação Brasileira de Normas Técnicas
ASTM	–	*American Society for Testing and Materials*
IA	–	Inteligência Artificial
PNL	–	Processamento de Linguagem Natural
RGPD	–	Regulamento Geral da Proteção de Dados
DORA	–	*Digital Operation Resilience Act*
NIS2	–	*Network and Information Systems Directive 2*
NIST	–	*National Institute of Science and Technology*
CSF	–	*Cyber Security Framework*
ENISA	–	*European Union Agency for Cybersecurity*

SUMÁRIO

1 INTRODUÇÃO

A transformação digital, impulsionada pela hiperconectividade (QUAN-HAASE; WELLMAN, 2005, p. 17) e pela massificação do processamento e armazenamento de dados na nuvem, tem gerado mudanças profundas nas interações econômicas, sociais e políticas. E as fronteiras, nos seus mais diversos significados e formas, foram ultrapassadas, pois num mundo cada vez mais virtual e digital, as relações espaço-tempo oferecem um universo sem concessões e limites. Estas mudanças são especialmente perceptíveis na forma como as organizações operam, num mundo cada vez mais globalizado, dinâmico e regulado.

Se, por um lado, esta revolução digital traz inúmeras oportunidades de negócio e inovação, por outro, impõe desafios significativos, sobretudo no campo da cibersegurança e conformidade regulatória. A crescente dependência de sistemas interconectados coloca a proteção de dados sensíveis e o cumprimento de normas como o RGPD, a NIS2 e a DORA como prioridades centrais para as organizações. Para sobreviver e prosperar em ambientes dinâmicos e voláteis, as organizações devem adotar abordagens proativas e contínuas de gestão de riscos e auditorias de segurança, principalmente as que atuam em mercados e indústrias altamente regulados, como por exemplo nas áreas da saúde, finanças e alta tecnologia.

Neste cenário, os avanços em inteligência artificial (IA), particularmente em aprendizado de máquina e processamento de linguagem natural, oferecem ferramentas poderosas para enfrentar esses desafios. As técnicas de classificação de textos, por exemplo, permitem mapear leis e regulamentações com boas práticas de mercado (MCINTOSH et al, 2024).

Em segurança da informação, em conformidade e auditoria são muitas as questões linguísticas referentes na formulação de textos e semântica. É um jogo constante de palavras, em que as evidências devem corroborar a eficácia e eficiência dos controles de segurança implementados para gerir determinados riscos. Neste sentido, a classificação de textos é essencial para que os mais diversos atores consigam interpretar um conjunto de textos legais, políticas de segurança da informação e cibersegurança, entre outros muitos documentos e termos linguísticos essenciais neste ramo de atuação.

Este trabalho explora a técnica de classificação de conjunto aberto (*open-set classification*), que investiga se uma amostra de teste pertence ou não a uma das classes semânticas no conjunto de treinamento de um classificador (WANG; VAZE; HAN, 2022), geralmente com base em um limiar de decisão ou métrica. Esta classificação é feita através da análise de similaridade cosseno (YIN et al, 2013), mapeando sentenças para vetores, onde as

técnicas de incorporação de frases representam frases inteiras e suas informações semânticas como vetores.

As análises são realizadas através de modelos que transformam sentenças em vetores semânticos. Além disso, é utilizada a técnica de aprendizado com poucos casos (*few-shot learning*), que permite generalizar novas classes com poucos exemplos (YAN; ZHENG.; CAO, 2017), e a transferência de aprendizado (*transfer learning*), que aproveita o conhecimento de corpora de textos dos modelos pré-treinados para gerar embeddings representativos (ORTAKCI, 2024).

Para este trabalho, são usados alguns modelos pré-treinados da categoria BERT, como o E5 Large, o GLOVE, que precisa ser treinado antes do uso, e o *Bag-of-Words* (TF-IDF com LSA), que não é um modelo pré-treinado. Enquanto BERT, ERNIE, XLNET e E5 Large possuem milhares de pares usados para treinamento, permitindo uma pontuação de similaridade semântica mais correta, TF-IDF e GLOVE baseiam-se em vetores gerados a partir de frequências e ocorrências de palavras. Todos os métodos são avaliados com base na similaridade de cosseno para encontrar as consultas mais semelhantes entre as frases dos diferentes documentos (VENKATESH SHARMA et al, 2024).

Alguns modelos apresentam um bom desempenho ao correlacionar textos com vocabulários diferentes, como os legais e formais do NIS2 e os mais técnicos do NIST CSF 2.0. Esta capacidade de identificar relações semânticas latentes demonstra como é possível alinhar padrões distintos, mesmo quando os textos não são diretamente equivalentes. Contudo, ajustes finos nos modelos são essenciais para garantir resultados mais alinhados ao domínio de aplicação nas tarefas de *downstream* (WAHBA, MDHAVJI, STEINBACHER, 2022).

Adicionalmente, a técnica de validação cruzada (*k-fold cross-validation*) para estimar a habilidade de um modelo para dados que não estão visíveis (SCHIMID, PHILIPP, 2024), combinada com *embeddings* semânticos e características baseadas em Bag-of-Words e SVM (Support Vector Machine), é utilizada para avaliar os modelos. Métricas como *precision, recall, F1-Score*, F1-Macro, F1-Micro e ROC-AUC não apenas validam o desempenho geral dos modelos, mas também a qualidade dos mapeamentos entre os textos do NIS2 e as subcategorias do NIST.

Este trabalho, ao formalizar relações entre padrões de mercado e regulamentações, contribui para orientar organizações sobre como aplicar técnicas de mineração de texto e modelos pré-treinados de forma eficiente, minimizando a necessidade de recursos computacionais intensivos ou treinamento extensivo de novos modelos.

1.1 Definição do problema

Com a transformação digital e a crescente interdependência das infraestruturas digitais, as organizações enfrentam desafios complexos em segurança da informação e conformidade regulatória. A sofisticação dos ciberataques e o aumento das exigências de conformidade, como o Regulamento Geral sobre a Proteção de Dados (RGPD), a Diretiva NIS2 e o ato DORA (*Digital Operational Resilience Act*), tornam insuficientes as abordagens tradicionais e reativas de cibersegurança. Empresas precisam de estratégias avançadas, integrando governança, gestão de risco e monitorização contínua para responder rapidamente aos incidentes. Além disso, o ambiente empresarial, marcado pelo paradigma VUCA (Volatilidade, Incerteza, Complexidade e Ambiguidade), exige que as organizações sejam ágeis e capazes de lidar com os desafios do negócio, com ameaças e regulamentações em constante mudança. Esse cenário é agravado em ambientes ágeis, onde a necessidade de inovação rápida entra em conflito com processos manuais de conformidade.

Dada a complexidade linguística e a heterogeneidade das leis, normas e diretivas, um dos principais desafios é identificar relações semânticas entre os diferentes textos, sobretudo com as políticas, práticas e controles de cibersegurança definidas pelas organizações. O desafio é maior quando as empresas são auditadas para manter a conformidade com exigências governamentais, regulatórios e de fiscalidade. Este é um trabalho que envolve muitos recursos, equipas multidisciplinares e sujeitos a erros de interpretação e por vezes de classificação.

Os modelos tradicionais de análise textual, baseados apenas em frequência de palavras ou abordagens generalistas, são insuficientes para capturar as nuances semânticas necessárias para mapear adequadamente os textos legais, políticas e práticas de mercado.

As técnicas de classificação de textos que permitem generalização, como *open-set classification*, *few-shot learning* e *transfer learning*, surgem como alternativas promissoras.

1.2 Objetivo deste trabalho

O objetivo deste trabalho é avaliar a aplicação de técnicas de aprendizado de máquina e processamento de linguagem natural (PLN) na classificação de textos no contexto de cibersegurança e conformidade regulatória. O estudo avalia a eficácia de métodos baseados em similaridade de cosseno, classificação de conjunto aberto (*open-set classification*), aprendizado com poucos casos (*few-shot learning)*, e transferência de aprendizado (*transfer learning*) na identificação de relações semânticas entre textos de regulamentações, frameworks de segurança

e boas práticas de mercado, nomeadamente entre NIST CSF 2.0 e NIS2. Por meio do uso de modelos pré-treinados como BERT, ERNIE, XLNET e E5 Large, bem como métodos clássicos como TF-IDF e GLOVE, este trabalho pretende evidenciar como o pré-treinamento e ajustes finos de alguns dos modelos podem melhorar significativamente a corretude e a utilidade das análises semânticas no mapeamento dos textos legislativos e técnicos.

No entanto, há uma lacuna no uso dessas abordagens em contextos que necessitam de classificações mais corretas, como o mapeamento entre padrões de mercado e documentos regulatórios.

Apresentados os objetivos, este trabalho aborda as seguintes questões:

1. Como garantir que as políticas, as práticas e os controles de cibersegurança e segurança da informação de uma organização estejam sempre alinhados com as regulamentações em constante mudança, de maneira automatizada, ágil, economicamente viável e eficiente?
2. Como avaliar a eficácia de diferentes modelos de programação de linguagem natural (PNL), na tarefa de mapeamento semântico de textos em cenários regulatórios específicos?
3. Qual é o impacto das técnicas de *few-shot learning* e *transfer learning* na capacidade de generalização dos modelos para novas classes de documentos?

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Conceitos-chaves relevantes

Como já referido, há um grande desafio em adaptar as políticas e controles de segurança de uma organização para que esteja em conformidade com as mais diferentes leis e regulamentações.

As auditorias em segurança da informação e cibersegurança enveredam por campos interdisciplinares devida a ampla cobertura das conformidades com questões regulatórias, certificações e atestações, em que é necessário um alinhamento semântico entre textos dos diferentes domínios.

É neste sentido que técnicas como classificação de conjunto aberto (*open-set classification)*, aprendizado com poucos casos (*few-shot learning*) e aprendizado por transferência (*transfer learning*) podem ser empregados em aprendizado de máquina, em que visam superar desafios associados a limitação dos dados, classes desconhecidas e generalização para novos cenários. Esses conceitos são utilizados em tarefas de classificação, reconhecimento de padrões e análise de dados, concretamente em contextos em que os métodos tradicionais enfrentam limitações, como a necessidade de grandes volumes de dados, recursos humanos e computacionais para classificarem os textos, e consequentemente treinarem novos modelos.

A classificação de conjunto aberto (*open-set classification*) é uma abordagem que permite ao modelo lidar com cenários em que, algumas classes de dados durante a fase de inferência, não foram vistas no treinamento. Diferentemente da classificação tradicional, que assume que todas as classes possíveis estão presentes nos dados de treinamento, a classificação de conjunto aberto considera a possibilidade de que novas classes possam surgir, exigindo que o modelo seja capaz de identificar amostras que não pertencem a nenhuma das classes conhecidas. A técnica geralmente emprega métodos que medem a distância ou a incerteza associada à representação de uma amostra para decidir se ela pertence a uma classe conhecida ou a uma nova classe (J.SHEIRER et al, 2013).

O aprendizado com poucos casos (*few-shot learning*) é uma técnica complementar projetada para treinar modelos, em que seja possível generalizar eficientemente apenas com algumas amostras rotuladas disponíveis para cada classe. Isto é útil em situações em que a coleta de dados rotulados é dispendiosa ou inviável. Modelos aprendizado com poucos casos frequentemente utilizam métodos como redes neurais ou técnicas de aprendizado contrastivo para extrair características compartilhadas entre classes. Essas características permitem que o

modelo adapte rapidamente seu conhecimento a novas classes com base em poucas amostras de treinamento. Esse conceito tem aplicações em áreas como reconhecimento de objetos, diagnóstico médico e classificação de textos em cenários de recursos escassos (GAO, FISCH, CHEN, 2021).

O aprendizado por transferência *(transfer learning*) refere-se ao uso do conhecimento adquirido por um modelo em uma tarefa específica para melhorar o desempenho em uma tarefa relacionada. Essa abordagem é fundamentada na ideia de que as características aprendidas em uma tarefa geral, como reconhecimento de padrões em texto ou imagens, podem ser transferidas para tarefas específicas que possuem conjuntos de dados menores ou diferentes sem a necessidade de treinar novos modelos (AGGARWAL, ZHAI, 2012). A ideia principal é utilizar os novos dados para treinar novamente os modelos existentes. Essa técnica reduz significativamente os requisitos de dados e computação, além de permitir que modelos menores obtenham desempenho comparável a modelos treinados para grandes volumes de dados. As questões inerentes a utilização desta técnica é o que, quando e como inferir a transferência de aprendizado (AGGARWAL, ZHAI, 2012).

Esses conceitos, embora distintos, frequentemente se complementam na resolução de problemas complexos em mineração de textos. A classificação de conjunto aberto e o aprendizado com poucos casos lidam com cenários de dados limitados ou desconhecidos, enquanto o aprendizado por transferência oferece uma base sólida para que modelos generalizem rapidamente a partir de poucos exemplos. A combinação dessas abordagens pode proporcionar maior solidez e flexibilidade aos sistemas de aprendizado de máquina, tornando-os mais adequados para aplicações de mineração de texto para os mais diferentes cenários.

2.2 Ciberataques e suas consequências

Muitas foram as vítimas e os prejuízos causados por ataques informáticos nos últimos 25 anos, como se pode observar na tabela 1, que retrata os dez maiores ataques informáticos sofridos a nível mundial.

Um bom exemplo da magnitude e dos impactos de um ataque informático, é o exemplo do *WannaCry*, um ataque do tipo *ransomware*, que em 2017 disseminou-se através do sistema operativo Windows, atingindo mais de 230.000 computadores, impactando mais 55.000 utilizadores, empresas e serviços públicos em pelo menos 100 países. Um dos principais alvos e que talvez tenha sido o que mais causou danos colaterais, foi o serviço de saúde britânico, pois além dos mais de US$ 120 milhões de prejuízo materiais, causou também o cancelamento

de mais de 19.000 consultas médicas e cirurgias e o redirecionamento das ambulâncias, deixando pessoas sem respostas aos atendimentos de urgência. O impacto deste e de outros ataques expôs vulnerabilidades não apenas tecnológicas, mas também relacionadas com a ausência de uma governação de segurança e estruturas de conformidade adequadas. Estes ataques destacaram a necessidade de as organizações irem além de medidas reativas e adotarem uma abordagem proativa e contínua à cibersegurança. Estima-se que globalmente os prejuízos estejam entre os US$ 4 a 8 bilhões.

Tabela 1 – Os 10 maiores ataques informáticos a nível mundial

Ataque	Ano	Impacto	Modus Operandi
Marriott Hotel Data Breach	**2018**	**500 milhões de contas de hóspedes expostas**	**Exploração de credenciais de login de funcionário não seguras**
WannaCry Ransomware	**2017**	**$4 a 8 bilhões em dados a nível mundial**	**Exploração de vulnerabilidade crítica no sistema operativo Windows**
Ukraine Power Grid Attack	**2015**	**Quebra no fornecimento de energia para milhares de clientes**	**Exploração de vulnerabilidades no SCADA**
2014 Yahoo Attack	**2014**	**500 milhões de contas de utilizadores hackeadas**	***E-mail phishing* direcionados para roubar credenciais dos utilizadores**
Adobe Cyber Attack	**2013**	**Exposição de código fonte do Photoshop e cerca de 38 milhões de contas de utilizadores expostas**	**Exploração de uma vulnerabilidade no software Adobe**
PlayStation Network Attack	**2011**	**77 milhões de contas de utilizadores comprometidas**	**Exploração de vulnerabilidades no software Adobe**
Estonia Cyber Attack	**2007**	**58 websites na Estônia ficaram inoperantes**	**Ataque de negação de serviço (*DDoS*)**
NASA Cyber Attack	**1999**	**1,7 milhões de partes de código fonte foram expostos**	**Acesso aos sistemas informáticos da NASA pela exploração de vulnerabilidades dos códigos fonte**
MOVEit	**2023**	**60 milhões de dados pessoais expostos**	**Exploração de vulnerabilidade *zero-day* no software *MOVEit***
Melissa Virus	**1999**	**US$80 milhões em prejuízos**	**Infecção de computadores com vírus de macro**

Fonte: https://em360tech.com/top-10/top-10-most-notorious-cyber-attacks-history.

Em resposta a este ambiente de risco, o quadro regulamentar tem evoluído significativamente, com normas como o Regulamento Geral sobre a Proteção de Dados (RGPD), a Diretiva NIS2 (*Enhancing Cyber Security Across the EU*) e o DORA (*Digital Operational Resilience Act*), juntamente com frameworks de cibersegurança como a NIST *Cybersecurity Framework* (CSF). A Diretiva NIS2, destinada a reforçar a ciber-resiliência das infraestruturas críticas na União Europeia, impõe medidas mais rigorosas de proteção contra ciberameaças. Já o DORA, focado no setor financeiro, estabelece requisitos para garantir a

resiliência operacional digital das entidades financeiras, assegurando que os serviços financeiros possam resistir, responder e recuperar rapidamente de interrupções operacionais, incluindo ciberataques.

Estas regulamentações não se limitam a garantir que as empresas protejam os seus próprios ativos, mas também asseguram que a privacidade dos dados dos clientes, parceiros e utilizadores seja uma prioridade central, enquanto protegem a continuidade dos serviços e operações essenciais em setores chave. Para cumprir estas regulamentações, as empresas são obrigadas a adotar uma monitorização contínua, implementar controles de segurança rigorosos e demonstrar a capacidade de responder rapidamente as novas exigências e ameaças.

Contudo, adaptar-se a estas exigências não é uma tarefa trivial. A implementação e a manutenção da conformidade contínua com estas normas exigem que as empresas adotem abordagens avançadas de gestão de riscos e segurança informática. A complexidade resulta do fato de tanto os regulamentos quanto as ciberameaças estarem em constante evolução. Esta realidade sublinha a necessidade de ferramentas e soluções que permitam às organizações automatizar a conformidade e garantir que os seus controles de segurança estejam sempre alinhados com os requisitos regulamentares.

Dado o crescente nível de complexidade regulamentar, com normas como a NIS2, o RGPD e o DORA, por exemplo, bem como o aumento da sofisticação dos ciberataques, a adoção de soluções de cibersegurança que integrem tecnologias avançadas tornou-se uma questão estratégica. As organizações que conseguem implementar uma governação de segurança forte, aliada a um sistema de gestão de riscos ágil e automatizado, estão mais bem preparadas para enfrentar os desafios impostos pela revolução digital.

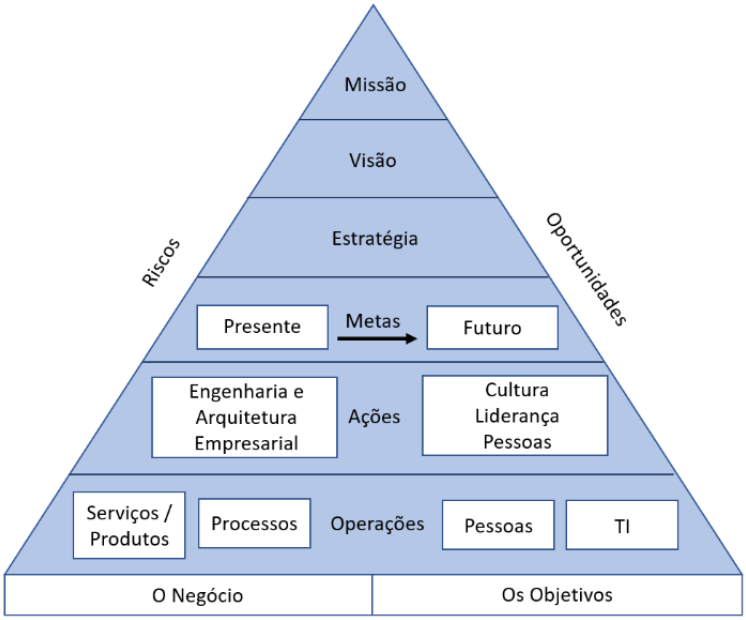
2.3 A cibersegurança como vantagem competitiva na estratégia empresarial

Existem várias teorias de vantagens competitivas, sendo algumas correntes orientadas a fatores externos (mercados, estruturas da indústria como SCP e 5 forças de Michael Porter) e outras correntes direcionadas aos fatores internos (RVB – Recursos e Competências e Capacidades Dinâmicas) (LASSERE; MONTEIRO, 2023, p. 181).

Jack Welch, (2005, p. 89) executivo estadunidense e autor de vários livros afirmou: "A estratégia de negócios não tem tanto a ver com ser capaz de prever algo, mas sim com ser capaz de responder rapidamente às mudanças reais quando estas ocorrem. Por isso é que a estratégia tem de ser dinâmica e capaz de antecipar."

A estratégia de uma empresa é definida de acordo com o mercado ou indústria em que se atua (ERNST & YOUNG, 2018), sendo as mais usuais e adotadas o modelo das 5 forças de Porter e RVB. Entende-se por vantagem competitiva a capacidade que uma empresa tem, de maneira mais eficiente possível, em agregar valor de forma sustentável em relação aos seus concorrentes. Ao delinear uma estratégia, a empresa está a definir e alinhar a missão, a visão, os objetivos e as ações das suas operações (Figura 1) através de sinergias e orquestração dos mais diferentes movimentos dos seus mais diversos atores e fatores externos e internos com a intenção primordial de obter lucro.

Figura 1 – Pirâmide estratégica – da missão às operações



Fonte: Compilação do autor [1].

Como mencionado anteriormente, a velocidade vertiginosa de como a tecnologia avança impõe-nos mudanças de comportamento e de tomada decisões. Face a este novo paradigma, é importante que uma empresa tenha alinhada em sua estratégia a cibersegurança de forma a proteger a privacidade, os dados e as interações entre todos os atores que contribuem para as dinâmicas de uma empresa.

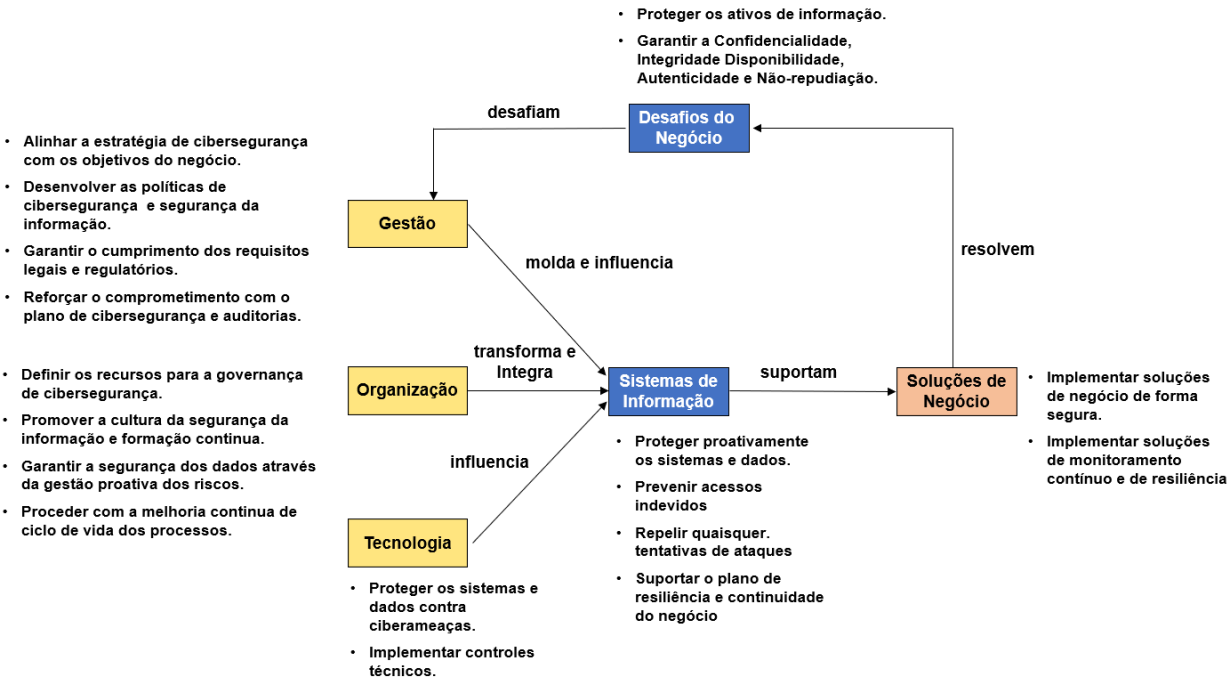
[1] Montagem a partir da pirâmide estratégica, disponível em <http://www.vaughanevansandpartners.com/consulting/for-business-clients/strategy/>. Último acesso em 17 nov. 2024.

É importante relembrar que a tecnologia dota as pessoas de ferramentas e capacidades adicionais para a execução das suas tarefas, sendo que em alguns casos até substitui os humanos por ser mais eficaz e eficiente. A simbiose entre tecnologia e humanos tem sido o grande impulsionador da transformação digital.

Com base na premissa do *framework* de Laudon, a cibersegurança é algo que tem que ser transversal e deverá ter o devido alinhamento entre Gestão, Organização e Tecnologia, permitindo assim dinamizar e proteger a simbiose dos capitais humanos e tecnológicos – não se trata mais apenas de uma questão puramente técnica da segurança informática.

Laudon & Laudon (2014, p. 72) em seu *framework* determina que os sistemas de informação deverão prover soluções de negócios para os desafios e necessidades de uma empresa através do alinhamento da gestão, organização e tecnologia (figura 2).

Figura 2 – *Framework* de Laudon



Fonte: Compilação do autor [2].

[2] Montagem a partir da matriz de Laudon & Laudon (2014, p. 73)

Como referido anteriormente, a hiperconectividade oferece um mundo sem fronteiras e concessões. A exposição de uma empresa as mais diversas tecnologias e inteligência artificial, inclusive através da simbiose entre tecnologia e humanos, expõe o negócio (e consequentemente a criação de valor) a muitos outros riscos, alguns já mensurados, mas outros ainda desconhecidos como por exemplo a gestão da experiência do cliente, seja este interno ou externo (permite a uma empresa em tempo real e de forma dinâmica perceber a experiência e sentimentos de uma pessoa na utilização de um produto ou serviço).

É importante salientar que a transformação digital requer uma maior colaboração entre todos os atores (compradores, fornecedores, clientes etc.), porque atualmente todos são vistos como parceiros de negócio, fazendo parte da mesma cadeia logística.

Como observado, as empresas, organizações e Estados dependem dos sistemas digitais, inteligentes e interconectados. Isto requer que as organizações sejam capazes de identificar qual o papel de cada ativo (função, criticidade, tipo de informação) afeta a sua exposição ao risco cibernético, e de como incorporar a cibersegurança na sua estratégia, e gerir ativamente os ciber-riscos associados, empregando inovação sem comprometer a cibersegurança, ou vice-versa.

A transversalidade da estratégia da cibersegurança deverá garantir a privacidade, a proteção dos dados e a integridade dos capitais humanos e tecnológicos (principais ativos de uma empresa), pois uma vez mais são estes que dão respostas aos desafios e necessidades da empresa na criação de valor. Neste sentido, o diferencial competitivo em proteger os principais ativos de uma empresa não assenta apenas na proteção dos dados, da privacidade e da integridade, mas também nas relações, experiências e sentimentos de todos os atores, reforçando positivamente a imagem e credibilidade da empresa.

No cenário empresarial moderno, a cibersegurança não é apenas um mecanismo de defesa contra ciberameaças, mas também um elemento estratégico que pode conferir às organizações uma vantagem competitiva significativa. A crescente dependência da tecnologia e dos dados nas operações diárias, combinada com um ambiente regulatório cada vez mais complexo, exige que as empresas integrem a cibersegurança como parte central da sua estratégia de negócio. Neste contexto, a cibersegurança é um fator determinante para garantir a confiança dos clientes, a continuidade das operações e o cumprimento das normas regulatórias.

3 TRABALHOS RELACIONADOS

A classificação automática de textos em domínios especializados, como o da segurança da informação e cibersegurança, tem sido amplamente utilizada na análise de vulnerabilidades, em correlacionar os logs, na interpretação de ameaças, no mapeamento de riscos, entre outros casos. Essa abordagem tem despertado o interesse das organizações, especialmente no que se refere à harmonização entre os termos legislativos, formais e técnicos, reduzindo os desafios impostos pela diversidade estrutural e linguística dos textos. O alinhamento das políticas de segurança da informação e cibersegurança com padrões de mercado e regulamentares, como o NIST CSF 2.0 e o NIS, permite às organizações derivarem abordagens consistentes para procedimentos e definição de controles de segurança. Isso facilita a contextualização entre diferentes domínios, assegurando o alinhamento semântico e diminuindo, por exemplo, as interpretações divergentes durante auditorias.

Embora os modelos pré-treinados como o BERT tragam benefícios para casos de uso independentes de domínio, eles enfrentam limitações em tarefas que exigem análise de textos específicos. As distribuições de palavras do corpus geral usado no treinamento do BERT diferem significativamente de corpora especializados, como os de segurança da informação e cibersegurança. Assim, aplicar diretamente esses modelos a tarefas específicas pode resultar em dificuldades para representar termos e contextos próprios do domínio especializado. Na extração de conhecimento em auditorias, no entanto, métodos baseados em BERT mostraram eficácia quando combinados com outros modelos, como BiLSTM e CRF, alcançando um F1 de 98,03%. Essa abordagem combina *embeddings* de palavras, regras de anotação de sequência e extração de entidades, demonstrando alta aplicabilidade em cenários específicos (XIANG, WEIBO, HUN, 2021).

Entre os trabalhos relacionados, destaca-se o CyBERT, um classificador baseado em *transformers* ajustados para a cibersegurança em sistemas industriais. Com ajustamento fino de hiper parâmetros e *embeddings* específicos, o CyBERT alcançou 94,4% de precisão, superando modelos tradicionais, destacando a importância de adaptar abordagens gerais a contextos especializados, como referido neste estudo (AMERI et. al, 2021).

Lawrie e Croft (2000) e Neto et al. (2000) contribuíram com abordagens metodológicas para a organização e sumarização de informações para os domínios da cibersegurança e segurança da informação. Ambos os trabalhos exploraram os processos de mineração de textos com o objetivo de gerar hierarquias ou taxonomias a partir dos documentos. Essas taxonomias foram criadas de forma automática, mas permitiram intervenções de especialistas para atender

as necessidades de compreensão. As metodologias incluíram agrupamento hierárquico de documentos, identificação de vocabulário específico do domínio, geração de palavras-chave e criação de resumos. Além disso, foram empregadas métricas estatísticas para validar e ajustar as edições realizadas, oferecendo ferramentas exploratórias que conectam a análise automatizada ao conhecimento especializado (Moura et al, 2008).

Adicionalmente, é de destacar o tema do agrupamento de textos através da aproximação semântica, com enfoque nas técnicas utilizadas na mineração de dados e aprendizagem automática, realçando as técnicas agrupamentos de textos do uso do algoritmo *K-Means*, que organiza textos em grupos com base na proximidade semântica. Foi explicado o funcionamento do algoritmo, desde a seleção inicial aleatória de centroides até a iteração que ajusta os grupos com base na média dos atributos dos pontos, bem como a dificuldade em determinar o número ótimo de grupos e a interpretação semântica dos clusters gerados (,LI; Wu, 2023).

Gururangan et. al (2020) afirma que ajustar os modelos de linguagem pré-treinados para um domínio específico melhora o desempenho em tarefas de NLP. A análise, realizada em quatro domínios (biomedicina, ciência da computação, notícias e resenhas) e oito tarefas de classificação, mostrou que uma segunda fase de pré-treinamento adaptada ao domínio (pré-treinamento adaptativo por domínio) traz ganhos de desempenho, independentemente da quantidade de dados disponíveis. Além disso, o ajuste aos dados não rotulados da tarefa específica (pré-treinamento adaptativo por tarefa) oferece melhorias adicionais. Alternativamente, estratégias de seleção simples para ampliar o corpus da tarefa mostraram-se eficazes quando recursos para pré-treinamento por domínio são limitados. O estudo conclui que o pré-treinamento adaptativo em múltiplas fases é altamente eficaz para aumentar a performance em tarefas de classificação (RANADE et. al, 2021).

Da mesma forma, este trabalho de conclusão de curso busca explorar as potencialidades de modelos baseados em transformadores, como os da família BERT, para o processamento de textos normativos e padrões de mercado. Além disso, são empregadas técnicas como validação cruzada para garantir a robustez e confiabilidade do modelo. A adaptação de abordagens gerais para o contexto da segurança da informação e cibersegurança não apenas aprimora os resultados em tarefas de classificação, mas também estabelece uma base comparativa com trabalhos previamente desenvolvidos, fornecendo uma perspectiva crítica sobre as limitações e avanços nesse campo de estudo. Assim, a trajetória delineada pelos trabalhos aqui destacados oferece uma base metodológica sólida que fundamenta e orienta a realização deste estudo.

4 METODOLOGIA DE PESQUISA

As organizações produzem muitos dados que através do processo de busca, ordenação, compreensão e contextualização produzem informação. Do ponto de vista estratégico, principalmente num mundo cada vez mais digital e globalizado, ao saber aplicar essa informação as organizações podem gerar novas ideias, aprimorar processos e tomar decisões mais assertivas, inovar e alcançar os seus objetivos (FAYYAD et. al, 1996).

A transformação digital, pela sua natureza, tem sido suportada por sistemas que se utilizam de bases dados, o que facilita a estrutura e organização dos dados. E isto tem contribuído para a produção massiva de dados. Neste sentido, a mineração de dados é considerada um processo importante em transformar grandes volumes de dados em conhecimento, algo essencial na execução estratégica de qualquer organização, exigindo para tal uma abordagem estruturada e metódica. O principal benefício da mineração de dados é a sua capacidade de identificar padrões e relações em grandes volumes de dados provenientes de múltiplas fontes. Com cada vez mais dados disponíveis – provenientes de fontes tão diversas como as redes sociais, sensores remotos e relatórios detalhados sobre o movimento de produtos e atividade de mercado –, a mineração de dados oferece as ferramentas para explorar plenamente o *Big Data* e transformá-lo em inteligência acionável. Além disso, pode atuar como um mecanismo para "pensar fora da caixa" (SAP, 2024).

Este trabalho de conclusão de curso aborda o estudo de técnicas de classificação de textos entre padrões de segurança da informação e cibersegurança com normas regulatórias, especificamente o NIST 2.0 CSF e a diretiva NIS2, o que permite por exemplo validar o alinhamento dos controles de segurança com as políticas de cibersegurança. Neste sentido a mineração de textos é a técnica a ser utilizada, porque apesar dos textos seguirem uma notação e redação formal, estes são considerados dados não estruturados, e para tal é necessário um método exploratório e estruturado que permita extrair informações, o qual é dividido em cinco fases: Identificação do Problema, Pré-Processamento, Extração de Padrões, Pós-Processamento e Uso do Conhecimento (REZENDE et al., 2023).

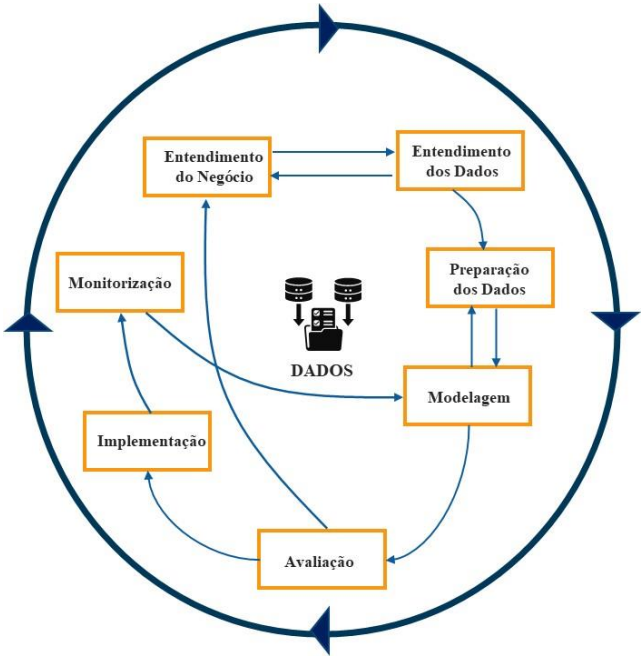
Para atender a esta necessidade, este trabalho utiliza a variante da metodologia CRISP-DM referida na formação da SAP (*SAP Machine Learning Training*), que está estruturada em sete fases, oferecendo uma abordagem sistemática e iterativa. Essas fases são adaptáveis, permitindo ajustes para se alinhar aos desafios específicos de cada projeto. Para este estudo, as cinco primeiras fases são fundamentais para organizar as atividades e orientar o

desenvolvimento das técnicas de mineração de texto aplicadas ao alinhamento dos diferentes textos. As fases seis e sete são referenciadas como observações para futuros projetos.

É de realçar que um bom projeto de ciência de dados deve ter um processo confiável e repetível para que pessoas com poucos conhecimentos ou formação em ciência de dados possam seguir e entender facilmente todas a etapas do projeto (HOTZ, Data Sicence PM).

Na figura 3 é apresentado o fluxo de etapas utilizadas para atingir os objetivos desse trabalho.

Figura 3 – Fases do CRISP DM – traduzido para português



Fonte: Imagem original disponível na comunidade SAP [3].

Esta metodologia apresenta uma abordagem estruturada para a mineração e classificação de textos, explorando as técnicas de classificação de conjunto aberto, aprendizado com poucos casos e aprendizado por transferência.

1. **Entendimento do negócio:** Esta fase inicial concentra-se na compreensão do problema e na definição de objetivos, estabelecendo para tal um âmbito dos trabalhos. Para este trabalho, o desafio principal é identificar quais categorias e subcategorias do NIST CSF 2.0 estão alinhadas semanticamente com a diretiva NIS2.

[3] Montagem a partir da imagem disponível em <https://community.sap.com/t5/technology-blogs-by-sap/sap-machine-learning-approaching-your-project/ba-p/13359323>. Último acesso em 17 nov. 2024.

2. **Entendimento dos dados:** Após a definição dos objetivos, esta fase envolve a exploração e análise preliminar da estrutura dos dados textuais disponíveis nos documentos. É importante compreender a estrutura e a semântica das funções, categorias e subcategorias descritas no NIST CSF 2.0 e da diretiva NIS2, o que permite identificar padrões, lacunas e relações semânticas entre os documentos, essenciais para a criação de modelos de classificação de texto. Para esta fase são consideras técnicas de processamento de linguagem natural (PLN), métodos para a geração de representações baseados em Bag-of-Words (com TF-IDF LSA) e *embeddings* semânticos.
3. **Preparação dos dados:** A preparação dos dados é uma etapa crítica de todo processo. Todos os textos são limpos, normalizados, transformados e lematizados em formatos adequados para análise. A criação de representações numéricas, como vetores TF-IDF e *embeddings* gerados por modelos como BERT, ERNIE e XLNET e E5 Large, é essencial para capturar características semânticas. Esta fase exige atenção a questões como remoção de ruídos e uniformização do vocabulário, garantindo a consistência necessária para a aplicação de técnicas de mineração de texto.
4. **Modelagem:** A fase de modelagem envolve a aplicação de algoritmos de mineração de texto para classificar trechos de documentos conforme as categorias definidas pelos padrões regulatórios. Técnicas de classificação de conjunto aberto, aprendizagem com poucos casos e transferência por aprendizagem são exploradas para lidar com classes emergentes ou categorias ausentes nos dados de treino. Os modelos são avaliados através análise de similaridade cosseno aplicada a vetores gerados, buscando a combinação mais eficaz para o alinhamento semântico entre as subcategorias e os textos normativos.
5. **Avaliação:** Durante a avaliação, é verificado se os modelos atendem aos objetivos definidos na fase inicial. Métricas de desempenho foram analisadas para identificar a eficácia das abordagens implementadas. A etapa inclui também a análise qualitativa da classificação do alinhamento entre os textos do NIST e NIS2.
6. **Implementação:** O principal objetivo deste estudo é identificar qual modelo consegue responder aos objetivos propostos, porque esta será a base de futuros trabalhos para implementar um sistema automatizado que permita auxiliar as organizações em minimizar os esforços e recursos empregues na manutenção das políticas e controles de segurança, respondendo de forma mais eficaz e eficiente as auditorias. Tal solução permitirá que as empresas mantenham a conformidade contínua, melhorem sua governança e mitigação de riscos, e adaptem-se rapidamente as mudanças normativas.

7. **Monitorização:** Stuart Clarke (SAP Training) explica que os modelos terão o desempenho reduzido ao longo, e que para tal, os modelos devem ter métricas de avaliação para medir o desempenho e estabilidade dos objetivos. Com os avanços constantes da IA e dos modelos pré-treinados, esta fase deve ser incluída nos futuros projetos, permitindo avaliar através de métricas definidas a abordagem e estabilidade do modelo, permitindo os ajustes necessários de acordo com os objetivos propostos.

A aplicação do CRISP-DM proporciona uma abordagem sistemática e orientada, permitindo a estruturação de atividades complexas e inter-relacionadas de forma eficiente. Além disso, a metodologia é particularmente útil na adaptação de processos tradicionais de mineração de dados às exigências específicas da segurança da informação e cibersegurança, onde a precisão e a confiabilidade são fundamentais. Ao adotar esta metodologia, o estudo oferece uma base sólida para o desenvolvimento de soluções inovadoras que auxiliem na gestão de riscos e na conformidade regulatória, promovendo um equilíbrio entre agilidade e segurança.

5 AVALIAÇÃO EXPERIMENTAL

O âmbito deste trabalho é o desenvolvimento e a avaliação de técnicas de classificação de textos para os domínios de cibersegurança, conformidade regulatória e auditoria, especificamente entre o framework NIST CSF 2.0 e a diretiva NIS2. O objetivo é explorar como métodos de processamento de linguagem natural (PLN) e aprendizado de máquina podem promover eficiência e precisão na interpretação e cruzamento de informações entre regulamentações, padrões de mercado e boas práticas de segurança da informação.

A fase de preparação dos dados envolve etapas como o carregamento dos documentos, organização dos dados textuais para facilitar sua manipulação, limpeza e normalização (conversão para minúsculas, remoção de caracteres especiais, entre outros) e lematização, que reduz palavras à sua forma raiz preservando o significado. Em seguida, ocorre a *tokenização*, dividindo o texto em unidades menores chamadas "*tokens*".

No primeiro experimento, é realizada a comparação de proximidade semântica (similaridade de cosseno) entre os termos do NIS2 e as subcategorias do NIST CSF, utilizando um limiar de similaridade de 0.7. Os valores iguais ou abaixo de 0.7 indicam ausência de similaridade, enquanto valores acima indicam similaridade. Para modelos modernos baseados em *Transformers*, são utilizados embeddings gerados por pré-treinamento, enquanto abordagens tradicionais, como *Bag-of-Words*, utilizam vetorização TF-IDF e reduções dimensionais com LSA. O modelo GLOVE emprega embeddings fixos e vetorização baseada em palavras.

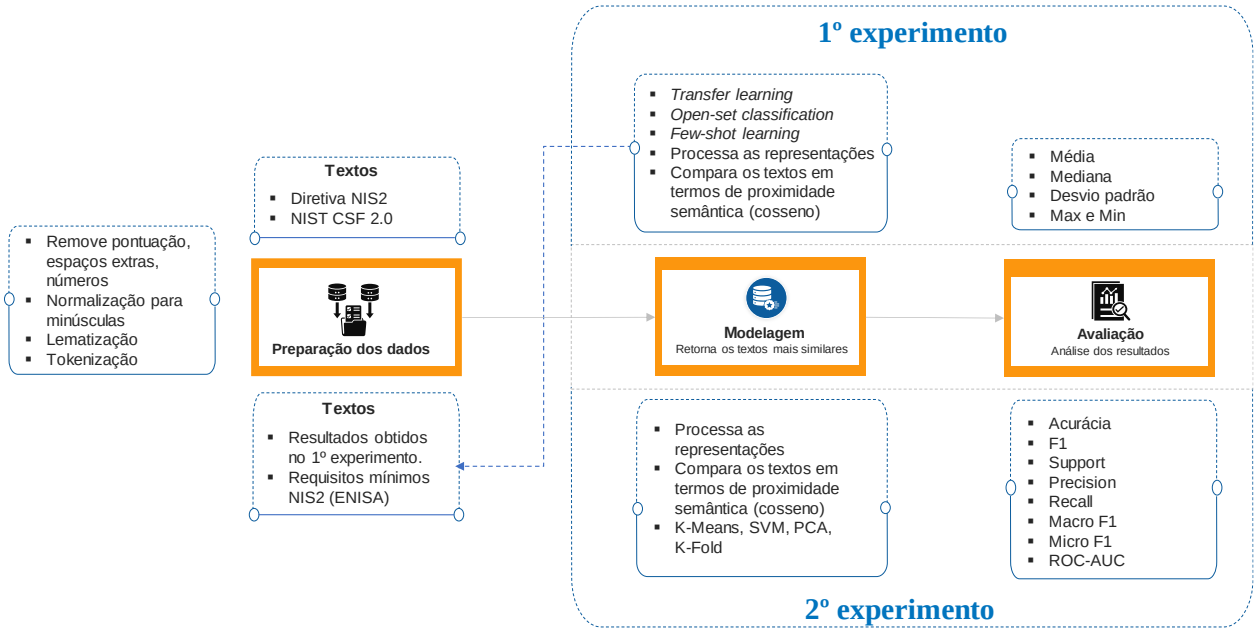
As técnicas avançadas, como transferência de aprendizado (*transfer learning*), são aplicadas para reaproveitar o pré-treinamento de modelos como BERT. Já no caso de Bag-of-Words, essa abordagem não é viável devido à ausência de algoritmos pré-treinados. O conceito de classificação de conjunto aberto (o*pen-set classification*) é explorado para identificar correspondência semântica entre os parágrafos do NIS2 e as subcategorias do NIST. Além disso, o aprendizado com poucos casos (*few-shot learning*) é utilizado para reavaliar sentenças não mapeadas com base na similaridade com sentenças previamente classificadas.

Para avaliar o primeiro experimento, são utilizadas estatísticas descritivas (média, mediana, mínimo, máximo e desvio padrão). O segundo experimento emprega a validação cruzada para comparar os resultados com referências existentes, como os requisitos mínimos do NIS2 publicados pela ENISA, incluindo mapeamentos para NIST CSF e ISO27001. Este experimento envolve:

- **Agrupamento:** *K-Means* é aplicado para agrupar textos em clusters vinculados às funções do NIST e às sentenças do NIS2, com uso de PCA para redução de dimensionalidade, facilitando a análise e visualização dos padrões semânticos.
- ***Classificação supervisionada:*** as combinações de embeddings gerados com TF-IDF são processadas com o modelo pré-treinado *SentenceTransformer* do SBERT, transformando textos em representações numéricas que capturam a semântica. O SVM é otimizado via *Grid Search* para classificação binária (relevantes e não relevantes), ajustando o parâmetro C entre 0.001 e 100 no kernel linear do SVM.
- **Ajustes dinâmicos:** um limiar de similaridade adaptativo é utilizado, baseado em estatísticas dos embeddings (média ou percentil 75). Essa estratégia permite balancear automaticamente as classes, melhorando a representatividade da classe minoritária e o desempenho geral do modelo.
- **Validação cruzada:** o método *K-Fold* (5 f*olds*) garante a representatividade das classes em cada iteração, reduzindo viés e variabilidade nos resultados.

A avaliação final é conduzida com métricas como acurácia, F1-Score e ROC-AUC, permitindo analisar o desempenho do modelo em cenários variados. Essa estratégia otimiza o desempenho geral, reduzindo o consumo de CPU e memória RAM, tornando o processo mais eficiente e escalável.

Figura 4 – Avaliação experimental dos modelos



Fonte: gráfico gerado pelo autor

5.1 Entendimento do negócio

Este estudo procura investigar como esses métodos podem ser utilizados para promover a eficiência e a classificação mais correta em processos que requeiram a interpretação e o cruzamento de informações entre regulamentações, padrões de mercado e boas práticas de segurança da informação.

Há um forte aumento global em torno das questões regulatórias de cibersegurança e segurança da informação, principalmente na União Europeia e nos Estados Unidos da América, em que as empresas têm que se adaptar as mais diversas leis, atos e regulamentações. Isto, também acarreta um aumento no número de auditorias, um esforço adicional que vai para além do ciclo de melhoria contínua, monitoramento dos controles de segurança e ajustes de políticas se segurança, boas práticas, entre outros muitos documentos e processos que estão inseridos em equipas de auditoria e conformidade.

Para dar resposta a este aumento de trabalho, permitindo que as equipes multidisciplinares mantenham o foco nas auditorias e melhoria contínua do trabalho, a resposta aos diferentes tipos de desafios por utilizar a tecnologia.

O NIST CSF 2.0 (*Cybersecurity Framework*) é uma ferramenta valiosa para empresas que buscam elevar seu nível de maturidade em cibersegurança e segurança da informação. Sua estrutura é adaptável a qualquer indústria ou mercado, e inclui referências a outros padrões amplamente utilizados, como a ISO 27001 (Gestão de Sistemas de Segurança da Informação), alinhando suas funções, categorias e subcategorias a esses *frameworks*. Além disso, devido à sua linguagem técnica e voltada para os domínios de cibersegurança, o NIST também oferece referências taxonômicas para facilitar sua aplicação prática. E, por ser agnóstica, pode ser conjugada com outros padrões de mercado, com por exemplo, o padrão ISO 27035 (Excelência em resposta a incidentes), que neste exemplo, permite uma abordagem mais aprofundada em resposta a incidentes, aumentando a segurança a resiliência e confiança do ecossistema em que uma organização atua.

As normas, atos e diretivas são essenciais para qualquer organização que queira estar em conformidade com o que exigido pelos diferentes governos e entidades públicas. Não se trata apenas de uma obrigação legal, mas da consciência em proteger todos os ativos (incluindo dados) de ataques maliciosos. Os textos têm uma estrutura e linguísticas legislativa e formal, recorrendo também a uma linguagem mais técnica, mas não tão aprofundada como no NIST.

As políticas de segurança da informação e cibersegurança são essenciais na definição da estratégia e do comprometimento da empresa em ter um ambiente seguro face as constantes

ciberameaças. Este tipo de texto utiliza muitos termos técnicos, recorrendo por vezes a diferentes *frameworks* e normas de mercado.

Utilizar padrões e recomendações oficiais como um guia orientador na definição dos controles de segurança demonstra não só preocupação, mas seriedade e compromisso com temas tão essenciais para as operações de qualquer organização inserida no mundo digital.

5.2 Entendimento dos dados

Os textos, apesar de possuírem uma organização, estrutura e características próprias, são considerados dados não estruturados. A análise de um texto requer contextualização e compreensão, e por conta da diversidade de termos técnicos e legislativos requerem esforços e recursos para que sejam interpretados e ajustados a realidade de cada organização. Caso, contrário os termos, e suas combinações, não passam de palavras soltas sem qualquer significado ao não ser o de sua etimologia.

Figura 5 – Nuvem de palavras dos documentos

Fonte: gráfico gerado pelo autor

Os dados – textos – utilizados neste trabalho são públicos e estão disponíveis em:

- Documentação sobre o framework NIST CSF 2.0, no formato PDF:
  <disponível em: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf>

- Matriz NIST CSF 2.0, no formato Excel:
  <disponível em:
  :https://csrc.nist.gov/extensions/nudp/services/json/csf/download?olirids=all>
- Diretiva NIS, no formato PDF:
  <disponível em: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02022L2555-20221227&qid=1732382984293
- Ato DORA, no formato PDF:
  <disponível em: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2554&from=FR>
- British Business Bank, política de segurança de um banco, no formato PDF:
  <disponível em: https://www.british-business-bank.co.uk/about/our-values/transparency/policies-and-procedures/information-security-policy>
- Exemplo de uma política de segurança, no formato PDF:
  <disponível em: https://www.bowiestate.edu/files/resources/information-security-public.pdf>
- Requisitos mínimos de segurança da ENISA, no formato Excel:
  <disponível em: https://www.enisa.europa.eu/topics/cybersecurity-policy/nis-directive-new/minimum-security-measures-for-operators-of-essentials-services>

O desenvolvimento deste trabalho utiliza a diretiva NIS2, a Matriz NIST CSF 2.0 e os requisitos mínimos de segurança da ENISA. Os outros documentos quando usados servem para comparação e explicação de algum tópico que assim seja necessário.

Tabela 2 – Volume dos documentos

Documento	Páginas	Palavras
Template de Política de segurança	27	9 169
Política de segurança *Bowie State University*	19	8 274
Política de segurança *British Business Bank*	14	4 651
Ato DORA	50	18 606
Diretiva NIS2	**53**	**44 618**
NIST CSF 2.0	**5**	**18 161**

5.2.1 Verificar similaridade e distância entre os textos no espaço vetorial

Esta fase consiste na leitura dos textos e do uso de uma técnica visual, através do agrupamento dos textos das subcategorias do NIST CSF 2.0 e dos termos dos NIS2, o que permite verificar neste primeiro momento o alinhamento semântico entre os textos. A visualização é feita através de técnicas de mineração de texto, transformação de dados textuais em *embeddings* e da redução de dimensionalidade para uma representação visual no espaço vetorial.

Para o agrupamento de textos, para além do documento NIS2, são também utilizados outros documentos, como o ato DORA (*Digital Operational Resilience Act*), um exemplo de uma política de segurança de uma Universidade (*Bowie State University*) e a política de segurança do British Business Bank. Estes documentos são necessários para obter uma amostra representativa dos termos.

Todos os documentos passam por um pré-processamento de texto, o que inclui a transformação das palavras em minúsculas, remoção de URLs, pontuações, números e palavras irrelevantes (stopwords), além da aplicação de lematização para reduzir os termos à sua forma base. Em paralelo, o framework NIST - no formato Excel - é estruturado para extrair as subcategorias descritivas, que são igualmente pré-processadas para normalização textual. A extração de termos relevantes dos documentos utiliza o método *Bag-Of-Words* com TF-IDF (*Term Frequency-Inverse Document Frequency*), o que permite a identificação de palavras-chave mais significativas com base em suas frequências relativas.

A semântica dos textos é capturada por um modelo de *embeddings*, o *SentenceTransformer*, que transforma os termos dos documentos em vetores de alta dimensão no espaço vetorial. Esses vetores encapsulam informações contextuais e semânticas, permitindo o cálculo de similaridades cosseno entre os elementos dos textos. A técnica de alinhamento identifica, para cada termo dos documentos, a subcategoria NIST mais similar, registrando a proximidade vetorial como uma métrica de distância semântica. Este alinhamento gera uma base de dados (*dataframe*) detalhada contendo termos, frequências, subcategorias NIST associadas e suas respectivas medidas de distância.

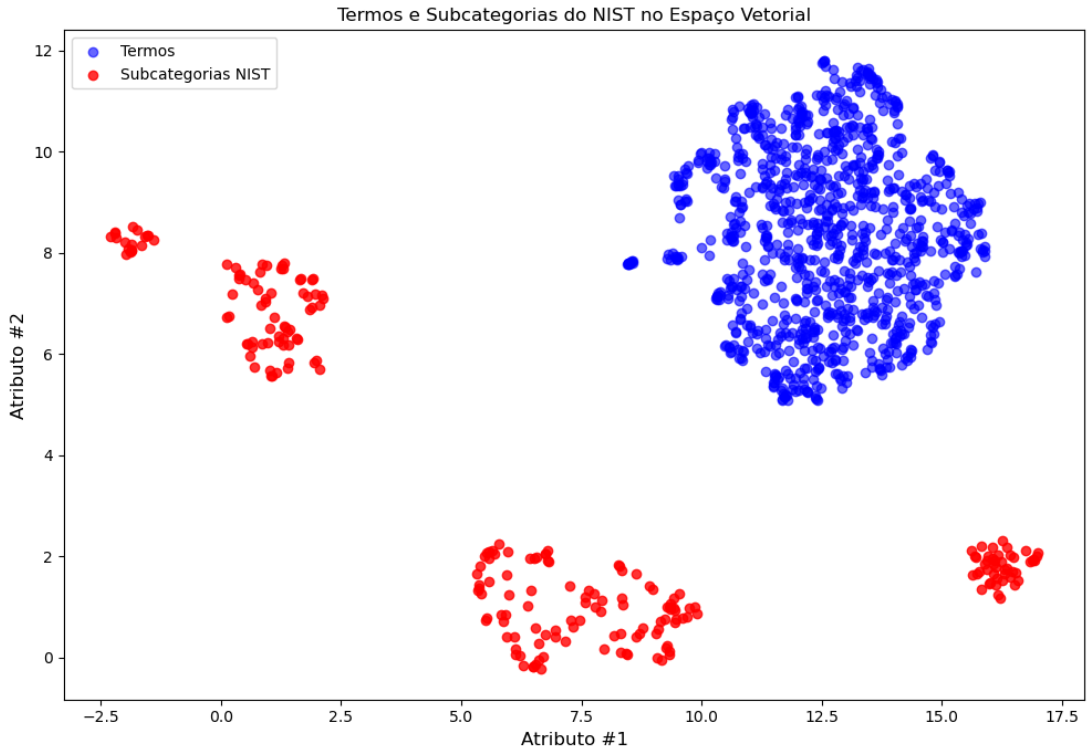
O processo de agrupamento dos documentos utiliza o algoritmo *k-Means*, uma técnica empregada em mineração de dados e aprendizado de máquina para particionar dados em grupos com base em sua similaridade. O número de clusters é definido de forma dinâmica, limitado ao número total de documentos processados – cada cluster representa um documento. Isto, assegura que o número de clusters não exceda a quantidade de dados disponíveis. Em seguida,

o algoritmo *k-Means* é aplicado aos *embeddings* dos documentos, que são previamente gerados utilizando o modelo de linguagem SentenceTransformer (*all-MiniLM-L6-v2*). Esses *embeddings* representam os documentos em um espaço vetorial de alta dimensionalidade, capturando suas características semânticas. É importante ressaltar que os *embeddings* resultantes representam os contextos gerais dos documentos.

Para garantir consistência nos resultados entre diferentes execuções, o algoritmo é configurado com uma semente fixa por meio de um parâmetro (*random_state=42*). Os *embeddings* são então particionados em clusters, com o objetivo de minimizar a soma das distâncias ao centro de cada cluster, o que resulta em agrupamentos de documentos semanticamente similares.

Esta é a representação visual do espaço vetorial (figura 6), em que se compara os termos extraídos de todos os documentos (representados pelos pontos azuis) com as subcategorias do NIST (representadas pelos pontos vermelhos). A projeção é feita utilizando o algoritmo UMAP (*Uniform Manifold Aproximation and Projection*), que reduz a dimensionalidade dos dados enquanto tenta preservar as relações de proximidade e similaridade do espaço original.

Figura 6 – Termos dos Documentos e Subcategorias do NIST no Espaço Vetorial



Fonte: gráfico gerado pelo autor

Os pontos azuis (termos) estão agrupados em uma região maior e mais compacta, o que indica que os termos dos documentos estão a compartilhar características semânticas similares no espaço vetorial. Os pontos vermelhos (subcategorias do NIST) estão distribuídos de forma distinta, formando grupos que refletem as diferenças entre os termos das subcategorias do *framework* do NIST. Quando os pontos vermelhos estão próximos aos pontos azuis, isso indica uma alta similaridade semântica entre os termos de um determinado documento e uma subcategoria específica do NIST. Quando os pontos vermelhos estão distantes dos azuis sugerem que os termos de um determinado documento têm pouca ou nenhuma relação semântica com essas subcategorias específicas do NIST.

A separação entre os agrupamentos de pontos sugere que os termos e as subcategorias possuem representações que, em sua maioria, estão bem separadas, o que pode indicar que os documentos contêm vocabulário ou conceitos que não têm relação semântica com as subcategorias do NIST, ou ainda palavras que não têm qualquer expressão na relação semântica, causando ruído na observação do conjunto de dados (textos).

Figura 7 – Frequência e Distância entre os textos dos documentos e NIST CSF

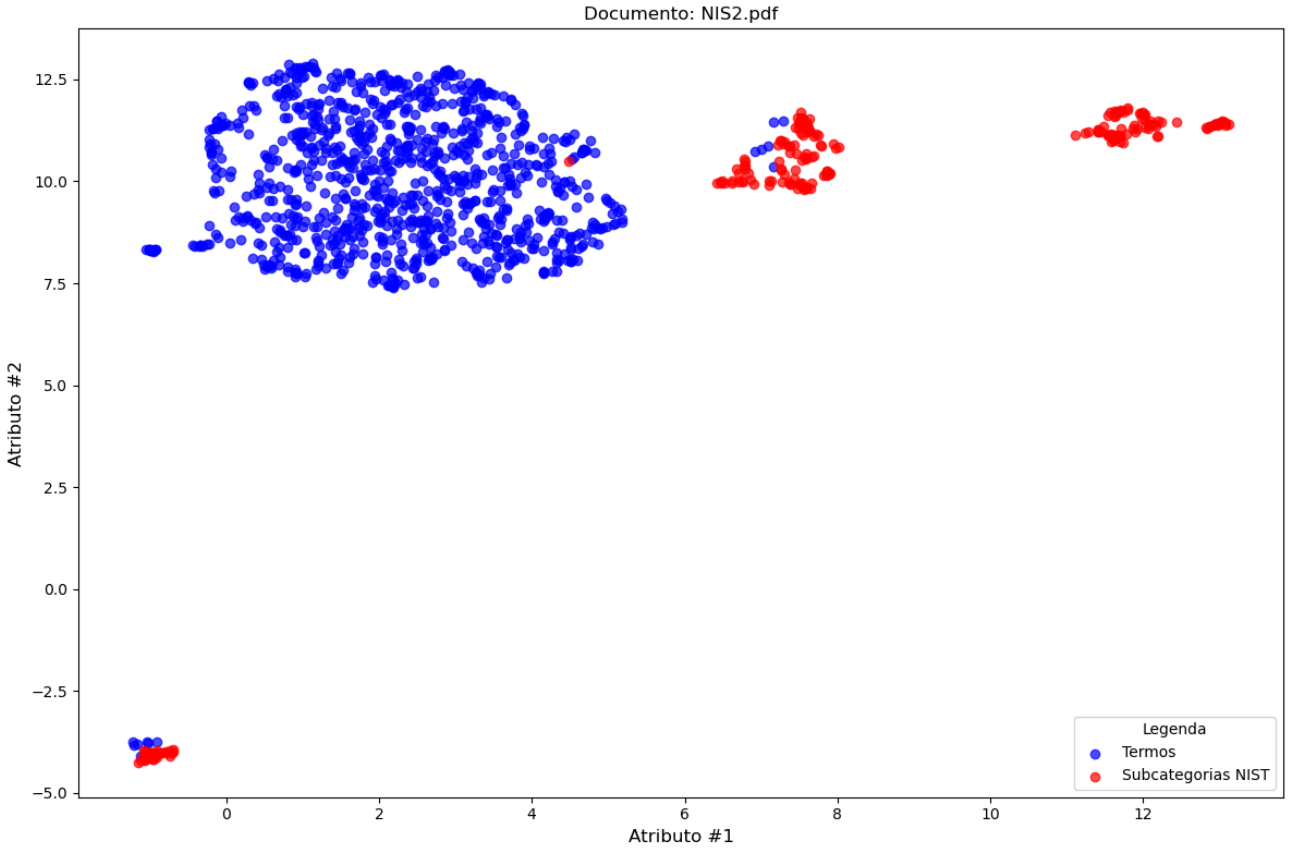
Top 20 Alinhamentos:

	Documento	Termo	Frequência	Subcategoria NIST mais próxima	Distância	Cluster
2436	2	ict	0.646910	ID.AM-03: Representations of the organization'...	0.363832	3
107	0	bsu	0.475724		0.276685	2
597	0	must	0.470772		0.399146	2
473	0	information	0.438294	RS.CO-03: Information is shared with designate...	0.372946	2
1832	1	security	0.392265	PR.AT-01: Personnel are provided with awarenes...	0.477354	0
1096	1	bank	0.336871		0.275031	0
2390	2	financial	0.312731	PR.AA-06: Physical access to assets is managed...	0.349531	3
3340	3	entity	0.310579	RS.CO-03: Information is shared with designate...	0.330467	1
1473	1	information	0.309683	RS.CO-03: Information is shared with designate...	0.372946	0
3277	3	directive	0.307054	PR.AA-05: Access permissions, entitlements, an...	0.307954	1
2340	2	entity	0.303311	RS.CO-03: Information is shared with designate...	0.330467	3
3068	3	article	0.254166		0.318866	1
3845	3	shall	0.250391		0.380484	1
1145	1	colleague	0.245290		0.281125	0
2845	2	shall	0.236192		0.380484	3
1460	1	incident	0.235359	RS.MA-02: Incident reports are triaged and val...	0.504882	0
1227	1	cyber	0.232322	GV.SC-02: Cybersecurity roles and responsibili...	0.418868	0
832	0	security	0.215723	PR.AT-01: Personnel are provided with awarenes...	0.477354	2
1913	1	team	0.207070		0.431251	0
3353	3	eu	0.205213		0.282346	1

Fonte: gráfico gerado pelo autor

O próximo gráfico (figura 8) demonstra que o documento NIS2 também contém termos e conceitos que estão a divergir ou que estão pouco alinhados com o *framework* NIST CSF. Assim como no gráfico anterior, os pontos azuis (termos) estão agrupados em uma região maior e mais compacta, o que indica que os termos dos documentos estão a compartilhar características semânticas similares no espaço vetorial. Alguns pontos estão distantes do grupo e mais próximos de alguns grupos vermelhos, o que indica correspondência semântica com algumas subcategorias do NIST. Para este exemplo os pontos vermelhos (subcategorias do NIST) também estão distribuídos de forma distinta, o que indica as diferenças entre as subcategorias do *framework* do NIST, muito provavelmente pela estrutura linguística de cada texto, o técnico do NIST e o legislativo da diretiva NIS2, por exemplo.

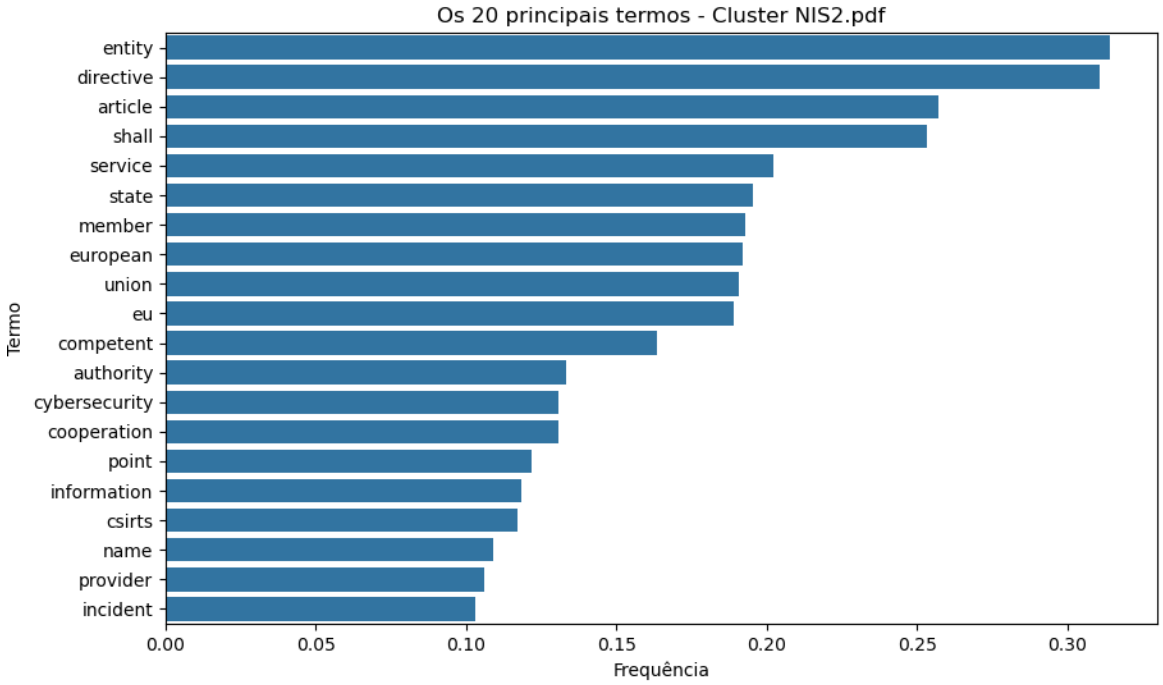
Figura 8 – Termos do NIS2 e Subcategorias do NIST no Espaço Vetorial



Fonte: gráfico gerado pelo autor

Este outro gráfico (figura 9) apresenta os 20 termos com maior frequência para o documento do NIS2, o que evidencia a divergência do âmbito ou da estrutura semântica do framework NIST, algo que também pode ser observado na Distância entre os textos apresentados na figura 10.

Figura 9 – Os 20 termos com maior frequência no documento NIS2



Fonte: gráfico gerado pelo autor

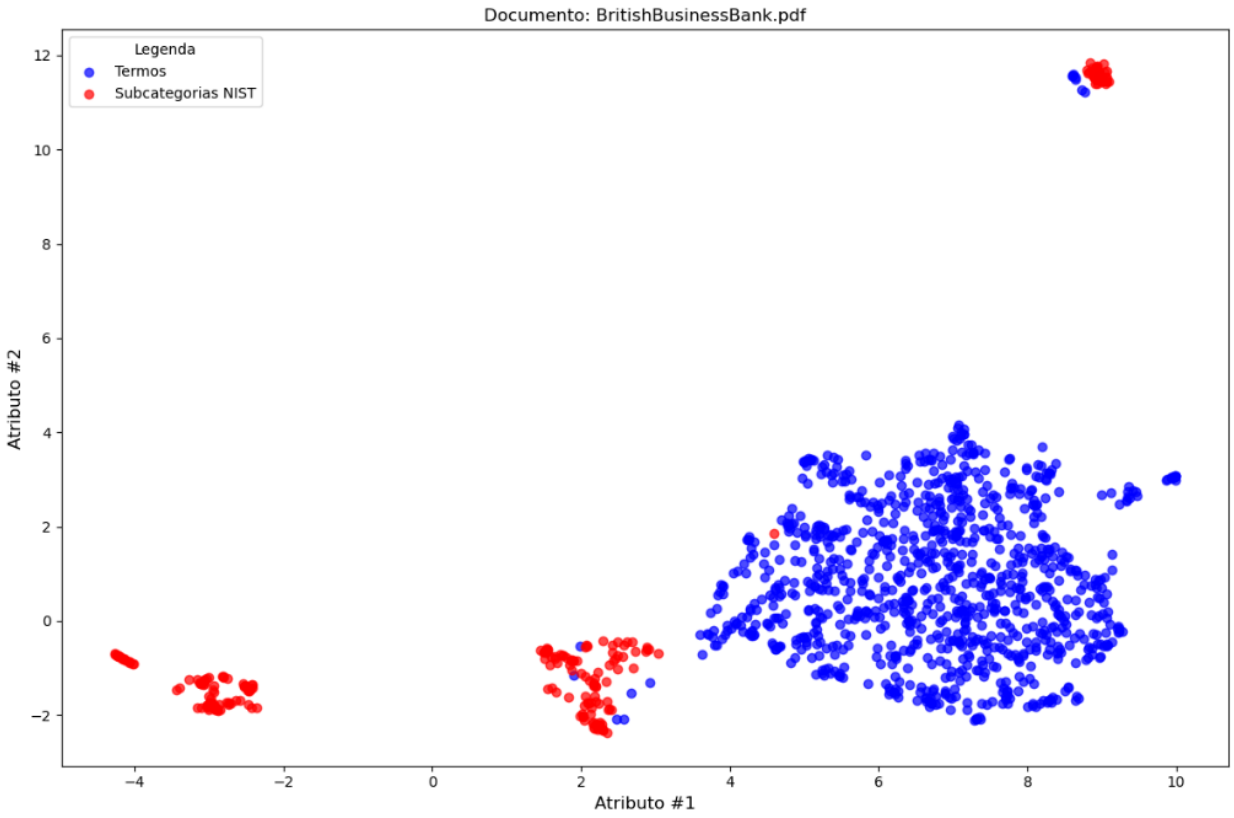
Figura 10 – Frequência e Distância entre os textos do NIS2 e NIST CSF

	Documento	Termo	Frequência	Subcategoria NIST mais próxima	Distância	Cluster
3335	3	entity	0.314238	RS.CO-03: Information is shared with designate...	0.330467	1
3271	3	directive	0.310672	PR.AA-05: Access permissions, entitlements, an...	0.307954	1
3069	3	article	0.257161		0.318866	1
3848	3	shall	0.253341		0.380484	1
3844	3	service	0.202041	ID.AM-04: Inventories of services provided by ...	0.437528	1
3879	3	state	0.195375		0.306668	1
3576	3	member	0.192811		0.398166	1
3350	3	european	0.192136		0.256996	1
3954	3	union	0.190587		0.297226	1
3348	3	eu	0.189037		0.282346	1
3152	3	competent	0.163471	PR.AT-02: Individuals in specialized roles are...	0.348846	1
3087	3	authority	0.133326	PR.AA-05: Access permissions, entitlements, an...	0.439359	1
3225	3	cybersecurity	0.130762	GV.RM-06: A standardized method for calculatin...	0.629316	1
3194	3	cooperation	0.130694	GV.OC-04: Critical objectives, capabilities, a...	0.332430	1
3680	3	point	0.121681		0.399187	1
3469	3	information	0.118455	RS.CO-03: Information is shared with designate...	0.372946	1
3219	3	csirts	0.116937	GV.SC-02: Cybersecurity roles and responsibili...	0.269576	1
3595	3	name	0.109076		0.427273	1
3739	3	provider	0.106000	DE.CM-06: External service provider activities...	0.452748	1
3453	3	incident	0.103072	RS.MA-02: Incident reports are triaged and val...	0.504882	1

Fonte: gráfico gerado pelo autor

Neste outro exemplo (figura 11), faz-se a comparação entre os termos da política de segurança do *British Business Bank* e as subcategorias do NIST. Nota-se que há uma leve aproximação entre os termos dos documentos, com os alguns dos pontos azuis mais próximos dos pontos vermelhos, o que indica para extes casos uma maior relação semântica entre o termo da política de segurança da informação da entidade e a subcategoria NIST correspondente.

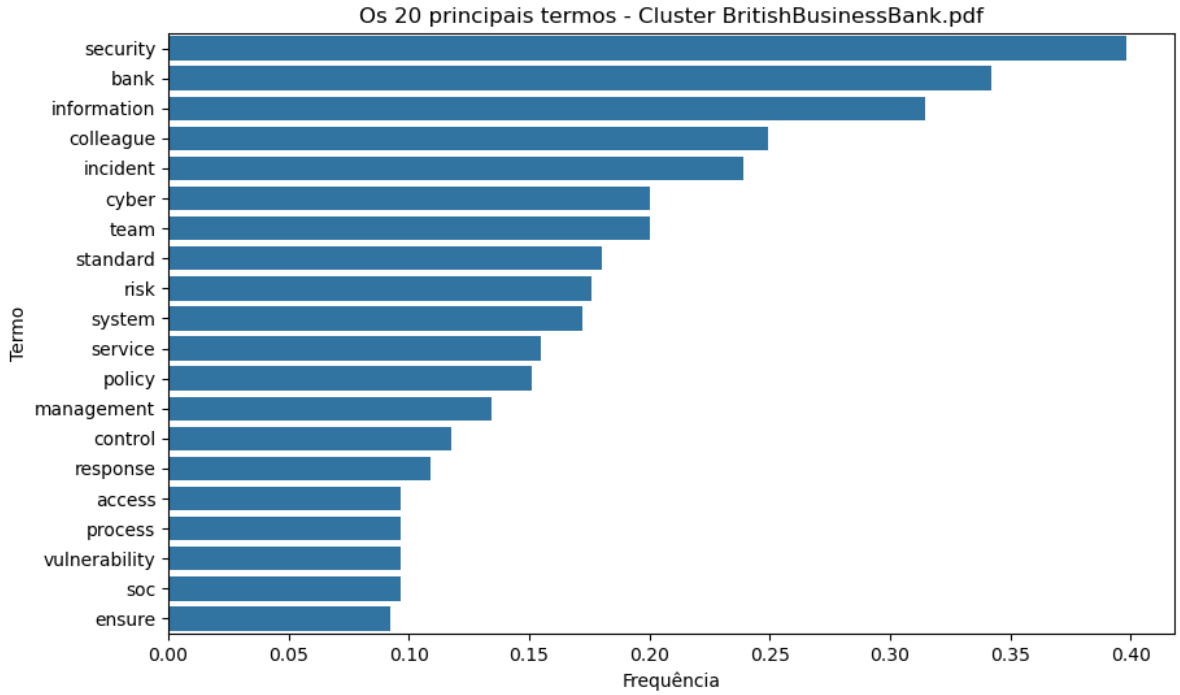
Figura 11 – Termos do documento *B.B.Bank* e Subcategorias do NIST no Espaço Vetorial



Fonte: gráfico gerado pelo autor

Na figura 12 é possível observar mais termos técnicos do que termos legislativos, realçando que o uso de termos contextualizados, e combinações entre eles, pode eventualmente melhorar a aproximação semântica entre os documentos, como também pode ser observado na métrica de Distância entre os textos apresentados na figura 13.

Figura 12 – Os 20 termos com maior frequência no documento *B.B. Bank*



Fonte: gráfico gerado pelo autor

Figura 13 – Frequência e Distância entre os textos do *B.B Bank* e NIST

	Documento	Termo	Frequência	Subcategoria NIST mais próxima	Distância	Cluster
1835	1	security	0.398195	PR.AT-01: Personnel are provided with awarenes...	0.477354	0
1098	1	bank	0.341964		0.275031	0
1469	1	information	0.314364	RS.CO-03: Information is shared with designate...	0.372946	0
1140	1	colleague	0.248998		0.281125	0
1453	1	incident	0.238917	RS.MA-02: Incident reports are triaged and val...	0.504882	0
1223	1	cyber	0.199947	GV.SC-02: Cybersecurity roles and responsibili...	0.418868	0
1914	1	team	0.199947		0.431251	0
1878	1	standard	0.180236		0.283570	0
1815	1	risk	0.176044	GV.RM-04: Strategic direction that describes a...	0.553308	0
1908	1	system	0.171852	ID.AM-08: Systems, hardware, software, service...	0.379619	0
1844	1	service	0.155086	ID.AM-04: Inventories of services provided by ...	0.437528	0
1681	1	policy	0.150895	PR.AA-06: Physical access to assets is managed...	0.366004	0
1553	1	management	0.134129	ID.AM-02: Inventories of software, services, a...	0.435000	0
1192	1	control	0.117363		0.297324	0
1804	1	response	0.108980		0.366707	0
1713	1	process	0.096405	ID.IM-03: Improvements are identified from exe...	0.399831	0
1004	1	access	0.096405	PR.AA-05: Access permissions, entitlements, an...	0.511276	0
1982	1	vulnerability	0.096405	ID.RA-04: Potential impacts and likelihoods of...	0.583192	0
1861	1	soc	0.096386	ID.AM-02: Inventories of software, services, a...	0.317095	0
1330	1	ensure	0.092214	PR.AA-04: Identity assertions are protected, c...	0.310201	0

Fonte: gráfico gerado pelo autor

O principal objetivo desta primeira etapa é entender a estrutura e os termos linguísticos dos diferentes documentos, permitindo ter uma base de comparação para poder estruturar as próximas etapas do estudo, principalmente no que toca os termos utilizados pela diretiva NIS2.

Conclui-se, portanto, que o entendimento da estrutura linguística de cada texto evidencia a necessidade de uma preparação adequada dos dados para evitar ruído, ou seja, há uma necessidade de melhorar a contextualização técnica (segurança da informação e cibersegurança) dos documentos legislativos para uma melhor relação semântica entre os textos das subcategorias do NIST.

5.3 Preparação dos dados

Para maximizar o potencial das arquiteturas e aplicações de cada modelo, é estabelecido um padrão inicial de pré-processamento comum a todos os documentos, com adaptações específicas conforme a arquitetura e os objetivos do modelo, da arquitetura ou técnica testada. Essa etapa abrange o carregamento dos documentos, a estruturação dos dados textuais de forma a facilitar sua manipulação e a aplicação de técnicas fundamentais de limpeza e normalização. Essas técnicas incluem a conversão para letras minúsculas, remoção de caracteres especiais, URLs, pontuações, números e palavras irrelevantes (stopwords), além de lematização, que reduz as palavras à sua forma raiz, preservando seu significado original.

Após essa limpeza, é realizada a *tokenização*, o processo de segmentar o texto em unidades menores chamadas "tokens", que podem ser palavras, sub palavras ou caracteres, dependendo do modelo. Para modelos baseados em *Transformers*, como os BERT ou XLNet, os *tokens* são frequentemente convertidos em embeddings contextuais gerados a partir de pré-treinamento, refletindo o significado semântico ajustado ao contexto da palavra na frase. Já abordagens tradicionais, como Bag-of-Words, utilizam técnicas como vetorização com TF-IDF, que pondera a importância de termos em relação ao corpus, e reduções dimensionais com LSA (*Latent Semantic Analysis*) para capturar relações semânticas de alto nível. O modelo GLOVE, por sua vez, emprega embeddings estáticos baseados em coocorrência de palavras, representando cada termo com vetores fixos, independentemente do contexto em que aparecem.

Essa padronização de pré-processamento garante que as diferenças na geração de embeddings, vetorização e análise semântica sejam adequadamente tratadas, possibilitando comparações justas entre os modelos e assegurando que cada arquitetura funcione dentro de suas capacidades ideais para as tarefas específicas deste estudo.

5.4 Modelagem

Como referido anteriormente, para este trabalho, diferentes algoritmos de extração de padrões são avaliados no contexto do PLN. A modelagem abrange diversas abordagens, cada uma direcionada para os diferentes objetivos de análise de texto, processamento semântico e alinhamento.

Os algoritmos de extração de padrões utilizados são aplicados para identificar estruturas relevantes nos dados textuais. O objetivo principal é comparar termos (palavras, frases, parágrafos, sentenças ou até mesmo o documento como um todo) entre os documentos e o NIST CSF 2.0, mais precisamente entre o NIS2 e este *framework*.

Para a etapa de análise com frequências de termos e redução de dimensionalidade, técnicas como TF-IDF, LSA e *clustering* são empregadas. A matriz TF-IDF é utilizada para capturar a relevância dos termos e reduzir a dimensionalidade, permitindo identificar padrões semânticos. Essas abordagens possibilitam a análise exploratória de relações estatísticas e semânticas nos dados textuais.

Além disso, é realizado um teste com *Bag-of-Words* (TF-IDF/LSA) combinado com algoritmos de *clustering*, como *KMeans*. Este processo visa agrupar os termos dos documentos com base em suas similaridades estatísticas ou semânticas, extraindo padrões que possam ser utilizados na análise comparativa entre os conjuntos de dados.

Por outro lado, para explorar os padrões semânticos mais densos, são utilizadas representações pré-treinadas por meio de embeddings semânticos fornecidos por modelos como SBERT, RoBERTa, ALBERT, DeBERTa, XLNet, ERNIE, E5 Large e SIMCSE. Embora esses modelos sejam utilizados em seu estado de pré-treinamento e não tenham sido ajustados especificamente para os domínios de segurança da informação e cibersegurança, eles auxiliam na identificação de padrões contextuais e semânticos entre NIS2 e NIST CSF.

Dessa forma, a etapa de modelagem neste trabalho centra-se nos algoritmos de extração de padrões, assegurando uma análise robusta e abrangente, tanto em termos estatísticos quanto semânticos, para os objetivos definidos.

5.5 Avaliação

Para todos os modelos, a avaliação de acurácia para os pares de textos é feita através da similaridade de cosseno. No caso do TF-IDF a similaridade de cosseno é feita entre os vetores LSA, enquanto para os demais modelos a similaridade de cosseno é aplicada entre os vetores dos *embeddings*.

A pontuação de similaridade do cosseno é estabelecida para todos os casos em 0.7, limiar de semelhança (*similarity_threshold*). Esse limiar estabelece que apenas correspondências com um escore de similaridade maior ou igual a 0.7 são consideradas suficientemente relevantes para análise qualitativa mais detalhada.

A categorização dos níveis de similaridade depende desse limiar e é definida como:

- **Alta**: maior ou igual a 0.8 (≥ 0.8);
- **Moderada**: entre 0.5 e 0.8 ($0.5 < x < 0.8$);
- **Baixa**: menor ou igual a 0.5 (≤ 5).
- **Rótulo binário (*binary label*)**: Baseado em um limiar de 0.8:
    - **1** (similaridade relevante): pontuação maior ou igual 0.8 (≥ 0.8);
    - **0** (não relevante): pontuação menor do que 0.8 (< 0.8).

Apenas os pares de termos que atinjam ou superam o limiar de similaridade definido são considerados candidatos relevantes. Os mais semelhantes de acordo com o limiar são então classificados para a análise mais detalhada.

Os que não atingem o limiar são marcadas como não semelhantes, influenciando o cálculo da percentagem de cobertura, o que ajuda a identificar lacunas no mapeamento.

Esta avaliação permite calcular a percentagem para a distribuição das categorias de similaridade (pares de termos):

- **% de Similaridade Alta:** percentual de pares com similaridade Alta;
- **% de Similaridade Moderada:** percentual de pares com similaridade Moderada;
- **% de Similaridade Baixa:** percentual de pares de termos com similaridade Baixa.

Também, são criadas métricas de estatísticas gerais de similaridade para avaliar a correspondência relevante e proximidade dos termos do NIS2 com as subcategorias do NIST:

- Média (*mean*) da pontuação de similaridade;
- Desvio padrão (*std*);
- Valores máximo (*max*) e mínimo (*min*) do limiar de similaridade.

Este projeto inclui ainda uma combinação de PLN e aprendizado automático para realizar validação cruzada (*k-fold cross-validation*) com cinco divisões, dividindo os dados em subconjuntos para garantir uma avaliação abrangente do modelo em diferentes combinações de treino e teste, reduzindo o viés e a variabilidade dos resultados. Neste sentido, os resultados obtidos durante as atividades de mapeamento entre os pares dos textos são reutilizados. Para além do pré-processamento dos textos já referido na preparação dos dados, a geração de *embeddings* é realizada através do modelo pré-treinado "*all-mpnet-base-v2*" da biblioteca *SentenceTransformers*, que converte os textos em vetores numéricos. Estes vetores capturam características semânticas dos textos, permitindo cálculos de similaridade. Em complemento aos *embeddings*, *Bag-of-Words* (TF-IDF) é utilizada para representar os textos com base na frequência de termos ajustada pela importância inversa em relação aos documentos. Esta combinação cria um conjunto de características capazes de capturar nuances textuais nos diferentes níveis. Com bases nestas características combinadas, a classificação dos textos é realizada através do SVM (Support *Vector Machine*) com kernel linear para balancear as classes de forma mais eficiente. As métricas de avaliação utilizadas são *precision, recall, f1-score e ROC-AUC*, o que permite obter uma visão detalhada sobre o equilíbrio entre os falsos positivos e falsos negativos, bem como a capacidade do modelo em distinguir corretamente as classes. A análise estatística das similaridades inclui o cálculo de valores mínimos, máximos, médios e percentis, permitindo a definição de limiares dinâmicos para a categorização.

5.6 Experimentos – testes e avaliações

Neste capítulo estão descritas as características técnicas de cada modelo e as técnicas utilizadas nos testes destes modelos. O sumário com as características técnicas de cada um dos modelos e técnicas utilizadas está disponível na tabela 3. É importante observar que para os modelos baseados em *transformers*, da família BERT (como ALBERT e ROBERTA), e outros abordados neste trabalho, há uma limitação padrão de 512 *tokens* no processamento de texto. Para textos maiores, a abordagem comum é dividir o conteúdo em blocos menores com no máximo 512 tokens antes de realizar os cálculos de *embeddings*, garantindo que nenhuma informação relevante seja perdida. Em modelos como DEBERTA, que suportam até 1280 tokens, a análise de textos mais extensos pode ser realizada diretamente, representando uma vantagem em relação aos modelos com o limite padrão.

Tabela 3 – Resumo técnico dos modelos e técnicas utilizadas nos testes

INDICADORES	ALBERT	DEBERTA	E5LARGE	ERNIE	ROBERTA	SBERT	SIMCSE	BoW (TF-IDF LSA)	GLOVE	XLNET
Representação textual	Média sobre *tokens*	Média sobre *tokens*	Média sobre *tokens*	Média sobre *tokens*	Média sobre *tokens*	*Pooling* sobre saída	*Pooler* Output (CLS)	TF-IDF para pesos, redução com LSA	Vetorização baseada em palavras	Média sobre *tokens*
Número máximo de *Tokens*	512	1280	512	512	512	512	512	Não aplicável	Baseados nas palavras	512
Dimensões (vetor de saída)	768	1024	1024	768	768	768	1024	1000	300	768
Cálculo das dimensões	Nativo	Nativo	Nativo	Nativo	Nativo	Nativo	Nativo	Vetorização TF-IDF Redução dimensões LSA	Nativo	Nativo
Comparação Direcional	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Não	Sim	Sim

5.6.1 1º experimento – testes e resultados

No primeiro experimento, é realizada a comparação entre os termos de proximidade semântica (similaridade por cosseno) do NIS2 e as subcategorias do NIST, utilizando um limiar de similaridade de 0,7 para todos os modelos. Os valores iguais ou inferiores a 0,7 indicam ausência de similaridade, enquanto os valores acima do limiar apontam para a existência de similaridade.

Nos modelos modernos baseados em *Transformers*, é utilizado o pré-treinamento já disponibilizado por essas arquiteturas. Para o modelo tradicional *Bag-of-Words*, a abordagem consisti na vetorização TF-IDF combinada com reduções dimensionais por LSA. Já no caso do GLOVE, empregam-se embeddings fixos de palavras e vetorização baseada em palavras.

A técnica de *transfer learning* é aplicada nos modelos com suporte a pré-treinamento, como aqueles baseados no BERT, permitindo o reaproveitamento de conhecimento previamente aprendido. No entanto, essa técnica não é utilizada no *Bag-of-Words*, porque não possui algoritmos pré-treinados.

O *open-set classification* busca determinar se há ou não correspondência semântica entre os parágrafos do NIS2 e as subcategorias do NIST. Complementarmente, o *few-shot learning* é aplicado para reavaliar sentenças que não apresentaram mapeamento inicial. Essa técnica consiste em analisar a similaridade entre sentenças mapeadas e não mapeadas, ajustando

a classificação com base na maior similaridade encontrada. Para isso, também é adotado o limiar de 0,7.

Os detalhes sobre cada modelo seguem abaixo:

- **Modelo ALBERT (*A Lite BERT*):** Este é um modelo da família BERT. O modelo ALBERT utiliza a variante "*albert-base-v2*" e sua arquitetura compacta foi projetada para eficiência, usando a fatoração da matriz de *embeddings* e partilha de parâmetros entre camadas, o que reduz o número de parâmetros sem comprometer o desempenho (Lan et. al, 2020). Os testes para este modelo utilizam a mesma abordagem do SBERT, com a diferença de que a geração de *embeddings* é feita através do vetor "*pooler_output*" quando disponível ou pela média dos estados ocultos dos *tokens*. Este é um modelo mais leve e consome menos recursos computacionais (CPU e RAM), contudo, não foi projetado para tarefas de similaridade de sentenças. As técnicas de *Open-set classification*, *few-shot learning* e *transfer-learning* seguem as abordagens já descritas para o modelo SBERT.
- **Modelos DEBERTA** (*Decoding-enhanced BERT with disentangled attention*): Este modelo é uma evolução do BERT e do RoBERTa, projetada para melhorar o desempenho em tarefas de PLN. O modelo base é o "*microsoft/deberta-v3-base*". A abordagem utilizada para testar este modelo foi a mesma do SBERT, com a nuance em aproveitar as características próprias deste modelo. O DeBERTa utiliza o *Disentangled Attention Mechanism*, separando as representações de conteúdos e posição, o que permite uma melhor compreensão do contexto semântico. Através do *Enhanced Mask Decoder*, melhora também a forma como o modelo prevê palavras mascaradas durante o pré-treinamento. Assim como no ALBERT, o modelo utiliza o atributo "*pooler_output*" quando disponível, ou realiza uma média dos estados ocultos (*last_hidden_state*) para gerar os *embeddings*. Outra vantagem deste modelo é o suporte até 1280 tokens, o que é maior do que o limite padrão de 512 tokens dos modelos anteriores (He et.al, 2021). As técnicas de *Open-set classification*, *few-shot learning* e *transfer-learning* seguem as abordagens já descritas para o modelo SBERT.
- **Modelo E5LARGE**: Projetado para tarefas de busca semântica e recuperação de informações, gera embeddings especializados que refletem a relevância dos textos para determinadas tarefas. Essa abordagem permite priorizar o alinhamento semântico das sentenças extraídas dos diferentes documentos, o que facilita destacar quais são os mais impactantes em classificação e similaridade (Wang et.al, 2024). Os testes deste modelo utilizam a mesma abordagem descrita no modelo ERNIE. Sua arquitetura está ajustada

para busca semântica, tornando-o eficiente para encontrar relações contextuais em grandes conjuntos de dados, sendo projetado para tarefas de correspondência semântica, como classificação de documentos e busca em texto. Por ser otimizado para capturar relações semânticas em textos curtos e moderados, é necessário também fazer a divisão do texto em blocos menores até 512 *tokens* antes de processá-lo, evitando assim a perda de informação. A desvantagem deste modelo reside na necessidade de recursos computacionais para o processamento dos textos e geração de *embeddings*. As técnicas de *Open-set classification*, *few-shot learning* e *transfer-learning* seguem as abordagens já descritas para o modelo SBERT. A variante utilizada é "*intfloat/e5-large-v2*".

- **Modelo ERNIE (*Enhanced Representation through Knowledge Integration*)**: Combina o aprendizado contextual com integração de conhecimento externo, permitindo capturar tanto as relações semânticas locais quanto as informações globais específicas para um domínio. Isto permite explorar a capacidade de análises mais densas das frases dos documentos, principalmente no que toca a diferença entre as linguagens técnicas e legislativas. Este modelo foi desenvolvido pela Baidu para análise de similaridade semântica. Também é uma alternativa aos modelos BERT. O ERNIE é projetado para integrar informações externas de conhecimento no treinamento, tornando-o particularmente eficaz em tarefas que envolvem significados contextuais e semânticos mais profundos. A variante utilizada é "*ernie-2.0-large-en*". A base dos testes é a mesma dos modelos anteriores. A geração de *embeddings* é feita a partir dos estados ocultos *(last_hidden_state*) do modelo. O modelo também suporta até 512 tokens, uma característica intrínseca à maioria dos *transformers*, o que requer novamente a divisão do texto em blocos menores antes de processá-lo, evitando assim a perda de informação. A vantagem deste modelo é que ele incorpora as informações semânticas externas, como fatos, entidades e relacionamentos, em sua estrutura de pré-treinamento, sendo eficaz em capturar significados implícitos e dependências de contexto (Sun et.al, 2019). Contudo, é importante ressaltar que, em tarefas altamente específicas como as definidas no âmbito deste trabalho, pode requerer um ajuste mais fino no que respeita os domínios específicos como legislação, segurança da informação e cibersegurança. Uma desvantagem deste modelo é a necessidade de mais recursos computacionais (CPU e RAM) para processar informações semânticas mais complexas quando comparado aos outros modelos. As técnicas *de Open-set classification* e *transfer-learning* seguem as abordagens já descritas para o modelo SBERT. Como o

modelo ERNIE é utilizado diretamente sem treinamento adicional com poucos exemplos, a técnica de *few-shot learning* não foi abrangida para este modelo.

- **Modelo ROBERTA:** Este é outro modelo pertencente à família BERT. Muito do que é feito para a arquitetura SBERT é reaproveitado durante os testes com este modelo. Uma das diferenças é que o modelo ROBERTA utiliza a variante "*roberta-base*". Outra característica deste modelo é que este não está otimizado diretamente para tarefas de similaridade de sentenças, pois gera os *embeddings* token por token, a partir do estado oculto deste modelo. A média dos estados dos tokens da sequência (*mean pooling*) é usada para representar o texto como um vetor único e criar uma única representação vetorial por sentença. Ainda, é de realçar a utilização do framework PyTorch para os cálculos de *embeddings*, permitindo a manipulação explícita dos tensores. Os textos mais longos que 512 tokens são truncados automaticamente pelo *tokenizer* do ROBERTA, outra importante característica a realçar (Liu et.al, 2019). Neste caso, quando o texto for maior do que 512 tokens, o mesmo é dividido em blocos menores antes de processá-lo, evitando assim a perda de informação. As técnicas de o*pen-set classification*, *few-shot learning* e *transfer-learning* seguem as abordagens já descritas para o modelo SBERT.
- **Arquitetura SBERT (Sentence-BERT):** Utiliza embeddings gerados com base no modelo BERT (*Bidirectional Encoder Representations from Transformers*), ajustado para calcular similaridades semânticas entre sentenças de forma eficiente. Esse modelo projeta os textos em um espaço vetorial onde vetores de textos similares são próximos, independentemente do tamanho ou complexidade do texto (REIMERS, GUREVYCH, 2019). A métrica de comparação mais comum é a similaridade de cosseno, que mede o ângulo entre os vetores. É importante observar que a magnitude do vetor não afeta diretamente o cálculo do cosseno, pois o denominador na fórmula da similaridade já normaliza os vetores, garantindo que o resultado varie entre -1 e 1. O parâmetro "*normalize_embeddings=True*" utilizado na biblioteca *SentenceTransformer*, configurada para a variante "*all-mpnet-base-v2*", é uma opção técnica para garantir que os vetores gerados sejam projetados em uma esfera unitária no espaço vetorial. Essa normalização é particularmente útil quando as embeddings precisam ser submetidas a outras etapas da rede ou análises adicionais. Isso evita que diferenças de magnitude entre os vetores, geradas por fatores como o tamanho do texto ou características do modelo, influenciem processos subsequentes que não utilizem diretamente a métrica de

cosseno. No entanto, para a similaridade de cosseno, a normalização intrínseca do denominador já é suficiente para neutralizar a influência da magnitude.

- **Modelo SimCSE (*Supervised Contrastive Sentence Embedding*):** Desenvolvido para capturar similaridade contextual, é utiliza para gerar *embeddings* semânticos refinados. O seu uso permite identificar as relações semânticas e contextuais entre os textos, priorizando a classificação mais correta em na classificação de similaridade. Este modelo é conhecido por gerar *embeddings* eficientes para comparações de sentenças, sendo particularmente eficaz em tarefas de correspondência semântica e recuperação de informações. Para os testes é utilizada a variante "*princeton-nlp/sup-simcse-roberta-large*", uma versão supervisionada baseada no modelo RoBERTa Large, ajustado para aprendizado contrastivo supervisionado, utilizando pares de sentenças positivas e negativas. Os *embeddings* são extraídos diretamente do "*pooler_output*", que representa o *embedding* da sentença completa, eliminando a necessidade de cálculos adicionais, como a média ou a soma dos vetores. Como descrito anteriormente para outros modelos, para dirimir a limitação dos 512 tokens, a divisão do texto em blocos menores é necessária antes de processá-lo, evitando assim a perda de informação (GAO, YAO, CHEN, 2022). Por ser um modelo que captura bem os *embeddings*, por ter treinamento supervisionado e pela eficiência para lidar com frases longas, o SIMCSE pode ser uma boa escolha para responder aos objetivos propostos para este trabalho. As técnicas de *Open-set classification*, *few-shot learning* e *transfer-learning* seguem as abordagens já descritas para o modelo SBERT.
- **Bag-of-Words (TF-IDF com LSA)**: Permite extrair os termos relevantes dos documentos através do método Bag-of-Words com TF-IDF (*Term Frequency-Inverse Document Frequency*), para realizar a análise de frequência dos termos, onde transforma os textos em vetores ponderados que refletem sua relevância no corpo do texto. A aplicação do LSA (*Latent Semantic Analysis*) permite a redução dimensional e captura de relações semânticas latentes. Esta é uma técnica que utiliza *Bag-of-Words* com TF-IDF (*Term Frequency-Inverse Document Frequency*) combinada com LSA (*Latent Semantic Analysis*) para identificar similaridades entre documentos baseados em representações de tópicos. Não requer o uso de modelos pré-treinados, sendo adequada para tarefas com grandes volumes de textos (Zhen, 2015). O uso do SVD (*Singular Value Decomposition*) torna a análise mais gerenciável, ágil e veloz, destacando as relações semânticas mais importantes. Essa abordagem é uma alternativa aos modelos anteriores e é eficaz para identificar as relações semânticas com base na frequência de

palavras e na decomposição de matrizes (Kadhim et. al, 2017). O TF-IDF, através da configuração do vetor, gera uma matriz de frequência ponderada que reflete a relevância de cada termo em relação ao documento. O LSA reduz a dimensionalidade da matriz TF-IDF usando SVD, identificando as relações latentes entre os termos e os documentos, melhorando a robustez da análise semântica e consequentemente a velocidade de processamento. Os testes para este modelo usam algumas das abordagens anteriormente descritas, como pré-processamento, cálculo de similaridade e classificação de similaridade (BALA, KUMARI, 2020). A técnica de TF-IDF com LSA não adapta ou ajusta modelos supervisionados com poucos exemplos; neste sentido, a técnica de *few-shot learning* não foi abrangida.

- **Modelo GLOVE** (*Global Vectors for Word Representation*): É um modelo de *embeddings* pré-treinados que converte palavras em vetores densos e estáticos, capturando relações semânticas entre termos. Para representar frases, utiliza a média dos vetores das palavras, facilitando análises de similaridade e tarefas de comparação semântica. Por ser pré-treinado, este modelo permite cálculos baseados em *embeddings* independentes de contexto. Isso significa que palavras com a mesma grafia, mas significados diferentes dependendo do contexto como "*token*" usado em autenticação em cibersegurança ou como unidade de texto em PLN), são representadas pelo mesmo vetor. Em contrapartida, modelos como as LLMs apresentadas anteriormente produzem representações dependentes de contexto, ajustadas ao significado contextual de cada palavra. (PENNINGTON, SOCHER, MANNING, 2014). O GLOVE destaca-se como uma alternativa eficiente, demandando menos recursos computacionais (CPU e RAM) devido à simplicidade de sua arquitetura. Essa característica permite cálculos rápidos e diretos, especialmente em tarefas que não exigem modelagem contextual aprofundada. A variante utilizada nos testes foi a "*glove.6B.300d*". Cada frase é representada pela média dos vetores das palavras que a compõem. No entanto, os *embeddings* estáticos apresentam limitações significativas, como a incapacidade de levar em conta o contexto. Isso pode diluir significados importantes em frases mais longas, tornando desafiador encontrar correspondências precisas, como por exemplo entre as subcategorias do NIST e os parágrafos do NIS2. Por outro lado, o modelo não é adaptável nem suporta aprendizado com poucos exemplos, de forma que a técnica de *few-shot learning* não foi aplicada. Apesar dessas limitações, as técnicas de *open-set classification* e *transfer-learning* seguem as abordagens já descritas para o modelo SBERT. Nos testes, são

utilizados os mesmos processos de pré-processamento, cálculo de similaridade e classificação de similaridade mencionados anteriormente.

- **Modelo XLNET:** Baseado em aprendizado autorregressivo bidirecional, o modelo XLNet destaca as relações semânticas considerando a posição das palavras no texto. Essa abordagem valoriza o impacto do contexto e a ordem das frases, enfatizando aquelas mais relevantes para a análise. A variante utilizada, "*xlnet-base-cased*" pertence à família de *transformers* autorregressivos e apresenta uma arquitetura projetada para capturar relações semânticas de forma bidirecional e autorregressiva, o que é particularmente vantajoso em análises que exigem atenção ao contexto sequencial, como as realizadas neste trabalho (Yang et.al, 2020).

  Assim como no BERT, o pré-processamento, a geração de embeddings, o cálculo de similaridade e a classificação de similaridade seguem as mesmas abordagens. A geração de *embeddings* utiliza os estados ocultos (*last_hidden_state*) do modelo. No entanto, a principal diferença do XLNet está em explorar sua capacidade autorregressiva (predição com base em tokens anteriores) e bidirecional (captura de relações semânticas em ambas as direções do texto). O modelo também apresenta um limite padrão de 512 *tokens*, exigindo a divisão do texto em blocos menores para evitar perda de informações.

  Outro diferencial significativo do XLNet é o mecanismo de ordenação permutada, que aprimora o aprendizado do contexto. As técnicas de *open-set classification* e *transfer-learning* utilizadas seguem as mesmas abordagens aplicadas ao modelo SBERT. Como o XLNet é empregado diretamente, sem ajustes ou treinamento adicional com exemplos limitados, a técnica de *few-shot learning* não é explorada neste caso.

Por fim, a avaliação deste primeiro experimento baseia-se em estatísticas descritivas, incluindo média, mediana, valores mínimo e máximo, e desvio padrão.

O modelo GLOVE apresenta o melhor desempenho global, alcançando 100% de alta similaridade. Isso indica que todas as correspondências avaliadas são consideradas relevantes. Para complementar este resultado, o modelo registra uma média alta de similaridade de 0,97475 e uma mediana igualmente alta de 0,97933. Além disso, o baixo desvio padrão (0,00189) reflete uma consistência nos resultados, demonstrando que o modelo oferece melhores correspondências. No entanto, é importante relembrar que, por ser baseado em *embeddings* estáticos, este modelo não considera o contexto de palavras em frases mais longas, o que pode limitar sua aplicabilidade em cenários mais complexos.

Tabela 4 – Resultados dos testes de acurácia de similaridade

MÉTRICA	ALBERT	DEBERTA	E5 LARGE	ERNIE	ROBERTA	SBERT	SIMCSE	BoW (TF-IDF)	GLOVE	XLNET
Total de palavras processadas	44 618	44 618	44 618	44 618	44 618	44 618	44 618	44 618	44 618	44 618
Percentagem de mapeamentos	92,97	100,00	100,00	60,00	100,00	100,00	100,00	100,00	**100,00**	100,00
Categorias não mapeadas	13	0	0	74	0	0	0	0	0	0
CPU (%) para cálculos das representações	17,40	13,90	10,90	92,20	11,60	13,50	26,60	12,70	**4,33**	19,70
RAM (%) para cálculos das representações	53	26	27	49	24	25	50	49	**24**	50
Média da Similaridade	0,84617	0,82627	0,78493	0,79160	0,83545	0,25431	0,44693	0,05346	**0,97475**	0,95779
Mediana da Similaridade	0,85374	0,82351	0,78115	0,78830	0,83987	0,22821	0,41113	0,00000	**0,97933**	0,96322
Std Similaridade	0,06820	0,04588	0,02214	0,02886	0,03794	0,12985	0,13269	0,10160	**0,00189**	0,01926
Max Similaridade	0,96754	0,96563	0,87250	0,86502	0,94908	0,76615	0,79315	0,89394	**0,99885**	0,99144
Min Similaridade	0,70005	0,75001	0,71747	0,75034	0,75010	0,46360	0,21361	0,00000	**0,87151**	0,86481

O XLNET também se destaca, alcançando 100% de alta similaridade, o que o coloca no mesmo nível do Glove em termos de relevância das correspondências. Além disso, registra uma média alta de similaridade de 0,95779 e uma mediana igualmente elevada de 0,96322. O baixo desvio padrão (0,01926) reflete uma boa consistência nos resultados. Esses números indicam que o modelo é eficaz para capturar relações semânticas de forma precisa, mesmo em textos complexos.

Os modelos baseados em *transformers*, como ROBERTA, ALBERT e DEBERTA, também apresentam bom desempenho. O ROBERTA destaca-se com 77,69% de alta similaridade, uma média de 0,83545, e um desvio padrão baixo (0,03794), o que reflete resultados consistentes. O ALBERT tem desempenho similar, com 73,12% de alta similaridade, uma média de 0,84617 e uma mediana de 0,85374, além de uma leve dispersão nos resultados (desvio padrão de 0,06820). O DEBERTA, embora registre uma proporção um pouco menor de alta similaridade (65,75%), também tem uma média boa de 0,82627 e resultados consistentes (desvio padrão de 0,04588).

Os modelos E5 LARGE e ERNIE apresentam desempenhos mais moderados. O E5 LARGE tem 22,27% de alta similaridade, mas a maior parte de suas correspondências fica na faixa de similaridade moderada (77,73%). Sua média de 0,78493 e o desvio padrão baixo (0,02214) sugere que, embora o modelo não atinja frequentemente a alta similaridade, ele

fornece resultados consistentes em níveis moderados. O ERNIE, por outro lado, alcança apenas 36,76% de alta similaridade e 63,24% de similaridade moderada, conseguindo mapear apenas 60% das categorias, o que o torna limitante na sua cobertura. Sua média de 0,79160 e o desvio padrão de 0,02886 indicam um desempenho aceitável, embora inferior aos demais modelos.

Os modelos como SBERT, SIMCSE e *Bag-of-Words* (BoW com TF-IDF) têm desempenhos inferiores. O SBERT apresenta 0% de alta similaridade e a maior parte de suas correspondências é classificada como baixa (94,13%), com uma média de apenas 0,25431 e um desvio padrão alto (0,12985), o que indica alta variabilidade nos resultados. Isto reflete a premissa de que este modelo é sensível aos hiper parâmetros e treinamento do modelo. O SIMCSE, com 65,39% de similaridade baixa e uma média de 0,44693, apresenta um desempenho melhor do que o SBERT, mas ainda muito inferior aos outros modelos. O método BoW (TF-IDF) tem 0,18% de alta similaridade, uma média de 0,05346, e uma grande proporção de correspondências com baixa similaridade (88,49%), reforçando as limitações dessa abordagem em capturar relações semânticas.

Os resultados destacam o GLOVE, XLNET e os modelos *transformers* avançados (ROBERTA, ALBERT e DEBERTA) como as melhores abordagens para identificar similaridades semânticas. Esses modelos demonstram uma combinação de alta similaridade, médias elevadas e consistência nos resultados, tornando-os ideais para cenários complexos que exigem maior sensibilidade semântica. Por outro lado, as abordagens mais simples, como SBERT, SIMCSE e BoW, mostram limitações significativas, especialmente em contextos em que é necessário capturar nuances semânticas e contextuais, como nos textos do NIST CSF e NIS2.

5.6.2 2º experimento - testes e resultados

A validação cruzada é uma técnica utilizada na área de aprendizado de máquina para avaliar a capacidade de generalização dos modelos. A finalidade é assegurar que o modelo de aprendizado de máquina não está super ajustado aos dados de treinamento e possui boa capacidade de generalização para dados não observados.

Para tal, é utilizada a técnica *K-Fold Cross-Validation*, em que se divide o conjunto de dados em múltiplos subconjuntos (*folds*) e avalia iterativamente o modelo, garantindo que todos os dados sejam usados tanto para o treinamento quanto para o teste. O método "*StratifiedKFold*", em particular, preserva a proporção das classes em cada divisão, o que é essencial em casos de desequilíbrio de classes.

Para os testes de validação cruzadas são utilizados os resultados dos testes anteriores, onde os mapeamentos já estão feitos, comparando assim com os requisitos mínimos exigidos pela ENISA (*European Union Agency for Cybersecurity*).

O pré-processamento dos textos segue a mesma abordagem dos testes anteriores. Para a representação dos textos, utiliza-se o modelo *Sentence-BERT* (SBERT), variante "*all-mpnet-base-v2*", que gera os *embeddings* de alta dimensionalidade capazes de capturar o significado semântico dos textos. Esses *embeddings* são utilizados tanto para as tarefas de classificação quanto para o agrupamento. Este modelo foi escolhido porque permite uma melhor generalização das comparações.

A técnica de agrupamento por *K-Means* permite agrupar os textos com base em seus *embeddings*. Este método identifica padrões e similaridades latentes nos dados, dividindo-os em sete *clusters*, considerando seis categorias associadas as funções NIST e uma categoria adicional para os textos não mapeados. Após a identificação dos clusters, os textos são associados às categorias mais frequentes ou classificados como "Não mapeado" quando nenhuma correspondência clara é encontrada. Para facilitar a interpretação dos agrupamentos, é ainda realizada a redução de dimensionalidade utilizando PCA (*Principal Component Analysis*), projetando os dados em duas dimensões. A visualização mostra a distribuição dos textos em relação aos clusters, destacando a coerência interna de cada grupo.

Para a classificação utiliza-se SVM (*Support Vector Machines*) com *kernel* linear, otimizando o hiper parâmetro C por meio de *Grid Search* com validação cruzada. Este processo testa múltiplos valores de C para encontrar o equilíbrio ideal entre o erro de classificação e a margem de separação.

Diversas métricas são calculadas para avaliar o desempenho do modelo em cada *fold* da validação cruzada, tais como acurácia, *precision*, *recall*, *F1-Score*, *F1-Macro*, *F1-Micro* e ROC-AUC para os treinamentos e testes do modelo.

A técnica de *open-set classification* é explorada através da utilização de técnicas de deteção de classes na SVM, suportada também pelo agrupamento do *K-Means* e da classificação binária (relevante e não relevante) através do limiar de similaridade adaptativo baseado nas estatísticas dos *embeddings* (mediana ou percentil 75). A exploração dos *embeddings* de alta qualidade ajudam em cenários com poucos dados representativos, realçando a utilização da técnica de *few-shot learning*. O *transfer learning* é explorado através da transferência de aprendizado do modelo pré-treinado do SBERT, e dos ajustes finos feitos durante os testes e treinamento do modelo.

Os modelos baseados em *transformers* ALBERT, DEBERTA, ROBERTA e XLNET apresentam desempenho perfeito durante os treinamentos e testes, com todas as métricas avaliadas atingindo o valor máximo de 1.0. Isso inclui precisão (*precision*), *recall, F1-score* (macro e micro), acurácia e ROC-AUC. Esses resultados indicam que esses modelos são capazes de mapear com extrema eficácia as similaridades semânticas entre os textos avaliados.

O desempenho perfeito, entretanto, pode ser interpretado como um possível indicativo de *overfitting*, dado o pequeno volume de dados textuais utilizado. Embora esses modelos tenham demonstrado uma capacidade notável de generalização no conjunto de teste, seu desempenho deve ser validado em cenários mais complexos ou com conjuntos de dados mais variados, o que permitiria avaliar melhor a qualidade dos resultados.

O modelo E5 LARGE apresenta métricas ligeiramente abaixo da perfeição nos testes. A acurácia média foi de 0.9984, enquanto o *recall* atingiu 1.0 e o *F1-score* ficou em 0.9984. Esses números indicam um excelente desempenho, com uma pequena variação em relação aos dados de treino, o que pode ser visto como uma generalização mais realista em comparação aos modelos com resultados perfeitos.

Já o ERNIE tem um desempenho um pouco inferior, com uma acurácia de 0.9906, *precision* de 0.9891, *recall* de 0.9882 e *F1-score* de 0.9884. Esses resultados mostram que o modelo, embora com bastante qualidade, apresenta maior sensibilidade às variações nos dados de teste, especialmente em relação aos *transformers* como DEBERTA e ROBERTA.

O modelo SBERT apresenta desempenho muito próximo do E5 LARGE, com acurácia média de 0.9969, F1-score de 0.9968, e um ROC-AUC de 0.9939 nos testes. Apesar de não alcançar a perfeição, o modelo demonstra ser altamente confiável para a tarefa de identificação de similaridades. Os resultados sugerem que, embora o SBERT não esteja no mesmo nível dos *transformers* mais modernos, ele ainda é uma escolha eficiente em termos de classificações mais corretas e consistência.

O SIMCSE surpreende ao apresentar métricas perfeitas em todas as avaliações, tanto no treino quanto no teste. Com 1.0 em todas as métricas avaliadas, o modelo demonstra uma capacidade muito boa de identificar similaridades semânticas. No entanto, assim como os outros modelos com resultados perfeitos, a ausência de erros de classificação levanta preocupações sobre *overfitting*.

A abordagem baseada em Bag-of-Words com TF-IDF apresenta resultados perfeitos em todas as métricas avaliadas. Embora surpreendente para um método tradicional, é provável que o desempenho perfeito seja um reflexo da simplicidade dos dados avaliados, indicando que o modelo pode não ser o mais indicado para os cenários mais complexos.

O GLOVE apresenta resultados muito altos, com acurácia média de 0.9984, *precision* de 1.0, e *F1-score* de 0.9984 nos testes. Apesar de não atingir métricas tão boas como alguns dos modelos *transformers*, o modelo demonstra uma grande eficácia e consistência, sendo uma escolha sólida em cenários onde a simplicidade é desejável.

Tabela 5 – Resultados dos testes validação cruzada (*K-Fold Cross Validation*)

Modelo	Acurácia	Precisão	Recall	F1-Score	F1-Macro	F1-Micro	ROC-AUC
ALBERT	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
DEBERTA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
E5 LARGE	0,9984	0,9969	1,0000	0,9985	0,9984	0,9984	1,0000
ERNIE	0,9906	0,9891	0,9882	0,9884	0,9902	0,9906	1,0000
ROBERTA	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
SBERT	0,9969	1,0000	0,9937	0,9968	0,9969	0,9969	0,9940
SIMCSE	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
BoW (TF-IDF LSA)	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
GLOVE	0,9984	1,0000	0,9968	0,9984	0,9984	0,9984	1,0000
XLNET	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Os resultados excelentes, ou quase perfeitos, apresentados por modelos como ALBERT, DEBERTA, ROBERTA, XLNET, SIMCSE, e BoW sugerem que os modelos podem estar super ajustados aos dados. O pequeno volume do conjunto de dados avaliados é um fator crucial nesse contexto, pois limita a variabilidade dos padrões semânticos presentes, facilitando o aprendizado completo pelos modelos. Embora isso não invalide os resultados, destaca a importância de validar esses modelos em conjuntos de dados mais diversos e extensos.

Os modelos *transformers* mais modernos, como ALBERT, DEBERTA, ROBERTA e XLNET, destacam-se por sua capacidade de mapear similaridades complexas com classificações muito extremas. Por outro lado, os métodos baseados em *embeddings* estáticos, como GLOVE, e métodos tradicionais, como BoW, também apresentam desempenho impressionante, mas podem ser mais limitados em cenários que demandem maior sensibilidade ao contexto semântico.

O SBERT e o E5 LARGE mostram desempenhos e consistentes e de qualidade, com métricas que, embora não perfeitas, refletem uma generalização mais realista. O ERNIE, embora eficaz, demonstra maior sensibilidade às variações nos dados.

Os resultados da validação cruzada indicam que os modelos *transformers* mais recentes, como DEBERTA, ROBERTA, e ALBERT, oferecem o melhor desempenho para a tarefa de identificação de similaridades semânticas, especialmente em conjuntos de dados pequenos. No

entanto, o desempenho perfeito registrado por muitos dos modelos levanta preocupações sobre *overfitting*, destacando a necessidade de validações adicionais em cenários mais diversificados.

Modelos como SBERT, E5 LARGE, e GLOVE também demonstram ser alternativas viáveis, equilibrando precisão e qualidade. Abordagens tradicionais, como o BoW, apresentam resultados inesperadamente bons, mas podem não ser tão eficazes em cenários mais complexos.

Os trabalhos futuros devem considerar o uso de conjuntos de dados mais extensos e diversificados para validar os resultados obtidos e explorar as capacidades de generalização dos modelos em diferentes domínios semânticos. Além disso, a implementação de métodos avançados de regularização e a adaptação dos modelos a contextos específicos, como os de cibersegurança e segurança da informação, podem ampliar ainda mais sua aplicabilidade prática.

6 CONCLUSÃO

A motivação deste trabalho de conclusão de curso assenta na necessidade crescente de lidar com os desafios linguísticos associados à conformidade com normas, diretivas, regulamentações, auditorias e padrões técnicos, particularmente nos domínios da cibersegurança e segurança da informação. A complexidade do alinhamento semântico entre textos normativos e técnicos afeta transversalmente todos os departamentos das organizações, que enfrentam dificuldades em interpretar e integrar diferentes tipos de textos de forma eficaz.

Este estudo tem como principal objetivo explorar e avaliar técnicas de aprendizado de máquina aplicadas à classificação automática de textos normativos, exemplificados pela diretiva NIS2, e padrões de mercado, representados pelo NIST CSF 2.0. Busca-se identificar abordagens práticas, eficazes e reutilizáveis em múltiplos setores da economia, destacando as mais promissoras para o alinhamento semântico.

Apesar do volume limitado de dados textuais, os resultados obtidos destacam o GLOVE, o XLNET e os *transformers* avançados, como ROBERTA, ALBERT e DEBERTA, como as melhores abordagens para identificar similaridades semânticas. Modelos como SBERT, SIMCSE e BoW (TF-IDF), por outro lado, apresentam limitações significativas, especialmente em contextos que requerem a captura de nuances semânticas e contextuais presentes nos textos normativos e técnicos.

Entre os destaques, o GLOVE e o XLNET atingiram 100% de alta similaridade, demonstrando elevada precisão e consistência. O GLOVE apresenta uma média de similaridade de 0,97475 e um desvio padrão de 0,00189, enquanto o XLNET alcança uma média de 0,95779 e um desvio padrão de 0,01926. Estes resultados indicam que o GLOVE, apesar de eficaz, tem limitações em contextos mais complexos devido à sua natureza de *embeddings* estáticos, enquanto o XLNET se sobressai na captura de relações semânticas precisas em textos mais desafiadores.

Os *transformers* mais modernos, como ROBERTA, ALBERT e DEBERTA, apresentam resultados sólidos, com percentuais de alta similaridade de 77,69%, 73,12% e 65,75%, respetivamente. Estes modelos mostram-se consistentes, como demonstrado pelas métricas de média e desvio padrão, mas evidenciam sensibilidade ao pequeno volume de dados disponível, o que requer ajustes finos e atenção ao *overfitting*.

Os modelos como SBERT e E5 LARGE, embora apresentem métricas ligeiramente inferiores, demonstram qualidade e potencial para generalização. Por exemplo, o SBERT atinge uma acurácia média de 0,9969 e um F1-Score de 0,9968, mas é mais sensível a variações nos

dados. Já as abordagens tradicionais como o BoW (TF-IDF), embora tenham alcançado métricas perfeitas nos testes, podem ser menos eficazes em cenários mais complexos devido à sua simplicidade e dependência direta do alinhamento de classes.

A validação cruzada (*K-Fold Cross-Validation*), combinada com as técnicas como a redução de dimensionalidade via PCA e a classificação utilizando SVM com otimização por Grid Search, revelou-se essencial para a avaliação detalhada dos modelos. Estas ferramentas permitem equilibrar erros de classificação e margens de separação, aumentando a confiabilidade dos resultados. Adicionalmente, a utilização de métricas abrangentes, como acurácia, *precision*, *recall*, *F1-Score* (macro e micro) e ROC-AUC, permitiu uma análise completa e robusta do desempenho, tanto em treino como em teste.

Os resultados do estudo destacam a viabilidade de aplicação de técnicas de aprendizado de máquina na classificação automática de textos normativos e técnicos, com modelos como o GLOVE, o XLNET e os *transformes* mais recentes a representarem avanços significativos na área de cibersegurança e segurança da informação. Estes modelos proporcionam maior eficiência no alinhamento semântico entre as regulamentações normativas e os padrões de mercados focados em termos tecnológicos, permitindo sua adaptação a diferentes setores e domínios.

As técnicas como *Open-set classification*, *Few-shot learning* e *Transfer learning* ajudam na classificação proposta para os textos, mas não são essenciais porque é possível utilizar abordagens combinadas como as descritas na avaliação dos textos (capítulo 5.3, secção 5.3.1) ou nas utilizadas para a validação cruzada (*K-Fold cross validation*).

Os testes revelados não demonstraram a necessidade de grande poder computacional, o que torna esta abordagem mais acessível para ambientes em que existam limitações de infraestruturas. Isto permite manter a eficiência sem comprometer a qualidade dos resultados.

Vale lembrar, contudo, que os avanças tecnológicos em inteligência artificial e processamento massivo de dados será algo trivial e muito mais acessível no futuro. Neste sentido, há que ter em consideração algumas das observações destacadas no capítulo 3 (Trabalhos Relacionados), como por exemplo, treinar um modelo específico para os domínios da cibersegurança e segurança da informação em que sejam abordados os alinhamentos semânticos entre os textos normativos e puramente tecnológicos, bem como ter taxonomias criadas e o suporte de conhecimento especializado para esta tarefa, aumentando assim o leque de oferta de ferramentas exploratórias, de análise e classificação textual automatizadas.

As limitações para este trabalho são a quantidade de exemplos utilizados e a falta de tempo para explorar os ajustes finos dos modelos. Em função dos resultados obtidos, pode-se

considerar para trabalhos futuros o treinamento de um modelo abrangente, como o XLNET, para os domínios da cibersegurança e segurança da informação, utilizando para tal um conjunto de dados mais diversificado.

REFERÊNCIAS

Aggarwal, Charu C; Zhai, ChengXiang, Mining Text Data, Chapter 1, p1-10, Springer, 2012.

Aggarwal, Charu C; Zhai, ChengXiang, Mining Text Data, Chapter 7, p223-258, Springer, 2012.

Ameri, Kimia; Hempel, Michael; Sharif, Hamid; Lopez Jr, Juan; Perumalla, Kalyan, CyBERT: Cybersecurity Claim Classification by Fine-Tuning the BERT Language Model, MDPI, 2021. Available at: <https://www.mdpi.com/2624-800X/1/4/31>.

Bala, S. Sai Manasa; Kumari, Santoshi, Comprehensive Analysis of Variants of TF-IDF Applied on LDA and LSA Topic Modelling, IJEAT, 2020. Available at: <https://www.ijeat.org/wp-content/uploads/papers/v9i6/D7669049420.pdf>.

Ernst & Young, Cybersecurity for competitive advantages, 2018 < https://www.ey.com/en_nl/news/2018/10/cybersecurity-in-organizations-must-enable-competitive-advantage-while-they-continue-to-protect-and-optimize-security-ey-report-reveals >. 2018. Último acesso em 13 de novembro de 2024.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, 1996. Available at: <https://onlinelibrary.wiley.com/doi/10.1609/aimag.v17i3.1230>.
SAP, SAP Business Technology Platform / SAP HANA Cloud, what is data mining (2014), Available at: <https://www.sap.com/products/technology-platform/hana/what-is-data-mining.html>, Último acesso a 19 Nov. 2024.

Gao, Tianyu, Fisch, Adam; Chen, Danqi, Making Pre-trained Language Models Better Few-shot Learners, Priceton, 2021. Available at: <https://arxiv.org/abs/2012.15723>.

Gao, Tianyu; Yao, Xingcheng; Chen, Danqi, SimCSE: Simple Contrastive Learning of Sentence Embeddings, ArXiv, 2022. Available at: <https://arxiv.org/pdf/2104.08821>.

Gururangan, Suchin; Marasovic, Ana; Swayamdipta, Swabha; Lo, Kyle; Beltagy, Iz; Downey, Doug; Smith, Noah A. Don't Stop Pretraining: Adapt Language Models to Domain and Tasks, ARXIV, 2020. Available at: <https://arxiv.org/pdf/2004.10964>.

He, Pengcheng; Liu, Xiaodong; Gao, Jianfeng; Chen, Weizhu, DeBERTa: Decoding-enhanced BERT with Disentangled Attention, ArXiv, 2021. Avaiable at: <https://arxiv.org/pdf/2006.03654>.

Hotz, Nick, What is CRISP DM, Data Science PM Training, Available at: <https://www.datascience-pm.com/crisp-dm-2/>, Último acesso a 19 Nov 2024.

Kadhim, Ammar Ismael; Cheah, Yu-N; Hieder, Inaam Abbas; Ali, Rawaa Ahmed, Improving TF-IDF with Singular Value Decomposition (SVD) for Feature Extraction on Twitter, ReasearchGate, 2017. Available at: <https://www.researchgate.net/publication/323546295_Improving_TF-IDF_with_Singular_Value_Decomposition_SVD_for_Feature_Extraction_on_Twitter>.

Lan, Zhenzhong; Chen, Mingda; Goodman, Sebastian; Gimpel, Kevin; Sharma, Piyush; Soricut, Radu, ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, ArXiv, 2020. Available at: <https://arxiv.org/pdf/1909.11942>.

Lasserre, Philippe; Monteiro, Felipe, Global Strategic Management, Bloomsbury Publishing, p. 37-96, 2022.

Laudon, C Kenneth - Laudon, P Jane, Management Information Systems, Managing the Digital Firm, Pearson, p72-96, 2014.

Lawrie, D. e W. B. Croft, Discovering and comparing topic hierarchies. p. 314–330, 2000.

Li, Youguo; Wu, Haiyan, A Clustering Method Based on K-Means Algorithm, Elsevier, 2012. Available at: <https://www.sciencedirect.com/science/article/pii/S1875389212006220>.

Liu, Yinhan; Ott, Myle; Goyal, Naman; Du, Jingfei; Joshi, Mandar; Chen, Danqi; Levy, Omer; Lewis, Mike; Zettlemoyer, Luke; Stoyanov, Veselin, RoBERTa: A Robustly Optimized BERT Pretraining Approach, ArXiv, 2019. Available at: <https://arxiv.org/pdf/1907.11692>.

McIntosh, Timothy R. et al. From COBIT to ISO 42001: Evaluating cybersecurity frameworks for opportunities, risks, and regulatory compliance in commercializing large language models, Computers & Security, Volume 144, 2024. Available at: <https://www.sciencedirect.com/science/article/pii/S0167404824002694?ref=pdf_download&fr=RR-2&rr=8dfe8bf71f186936>.

Moura, Maria Fernanda; Marcacini, Ricardo Marcondes; Nogueira, Bruno Magalhães; Conrado, Merley da Silva; Rezende, Solange Oliveria, Uma Abordagem Completa para a Construção de Taxonomia de Tópicos em um Domínio, USP, 2008. Available at: <https://repositorio.usp.br/bitstreams/49895a89-7629-4cc6-898f-8c6b40bc0e44>.

Neto, J. L., A. D. Santos, C. A. A. Kaestner, e A. A. Freitas, Document clustering and text summarization. In L. T. P. A. Company (Ed.), p. 41–55, 2000.

Ortakci, Yasin, Revolutionary text clustering: Investigating transfer learning capacity of SBERT models through pooling techniques, ELSEVIER, 2024. Available at: <https://www.sciencedirect.com/science/article/pii/S2215098624001162>.

Pennington, Jeffrey; Socher, Richard; Mannin, Christopher, GloVe: Global Vectors for Word Representation, Stanford, 2014. Available at: <https://nlp.stanford.edu/pubs/glove.pdf>.

Philipp, Schimid, K-Fold as Cross-Validation with a BERT Text-Classification Example, PHILSCHMID Blog, April 2020, Available at: <https://www.philschmid.de/k-fold-as-cross-validation-with-a-bert-text-classification-example>.

Ranade, Priyanka; Piplai, Aritran; Joshi, Joshi, Anupam; Finin, Tim, CyBERT: Contextualized Embeddings for the Cybersecurity Domain, IEEE, 2021. Available at: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9671824>.

Reimers, Nils; Gurevych, Iryna, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, ArXiv, 2019. Available at: <https://arxiv.org/pdf/1908.10084>.

REZENDE, S. Sistemas inteligentes: fundamentos e aplicações. Manole, 2003. ISBN 9788520416839. Available at: https://books.google.com.br/books?id=UsJe_PlbnWcC

SAP Machine Learning Training, available at: <https://community.sap.com/t5/technology-blogs-by-sap/machine-learning-in-a-box-week-2-project-methodologies/ba-p/13367317>, Último acesso a 19.nov.

Scheires, Walter J; Rocha, Anderson de Rezende; Sapkota, Archana; Boult, Terrance E. Toward Open Set Recognition, IEEE, 2013. Available at: <https://ieeexplore.ieee.org/document/6365193>.

Sun, Yu; Wang, Shuohuan; Li, Yukun; Feng, Shikun; Chen, Xuyi; Zhang, Han; Tian, Xin; Zhu, Danxiang; Tian, Hao; Wh, Hua, ERNIE: Enhanced Representation through Knowledge Integration, ArXiv, 2019. Available at: <https://arxiv.org/pdf/1904.09223>.

Venkatesh Sharma, K.; Ayiluri, Pramod Reddy; Betala, Rakesh; Jagdish Kumar, P.; Shirisha Reddy, K., Enhancing query relevance: leveraging SBERT and cosine similarity for optimal information retrieval, Springer Nature, s10772-024-10133-5, 2024, Available at: <https://link.springer.com/article/10.1007/s10772-024-10133-5>.

Wahba, Yasmen; Madhavji, Nazim.; Steinbacher, John. A Comparison of SVM against Pre-trained Language Models (PLMs) for Text Classification Tasks, Cornell University, 2022. Available at: <https://arxiv.org/abs/2211.02563>.

Wang, Hongjun; Vaze, Sagar; Han, Kai, Dissecting Out-of-Distribution Detection and Open-Set Recognition: A Critical Analysis of Methods and Benchmarks, International Journal of Computer Vision, 2024. Available at: <https://link.springer.com/article/10.1007/s11263-024-02222-4#citeas>.

Wang, Liang; Yang, Nan; Huang, Xiaolong; Yang, Linjun; Majumder, Rangan; Wei, Furu, Multilingual E5 Text Embeddings: A Technical Report, ANIXV, 2024. Available at: <https://arxiv.org/pdf/2402.05672>.

Welch, Jack, The Real-Life MBA: Your No-BS Guide to Winning the Game, Building a Team, HarperCollins, p. 89, 2015.
Wellman, Barry Physical Place and Cyber Place: The Rise of Networked Individualism, International Journal of Urban and Regional Research, June 2001. Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-2427.00309>.

Yan, L.; Zheng, Y.; Cao, J. Few-shot learning for short text classification. Multimed Tools Appl 77, 29799–29810, 2018. Available at: <https://link.springer.com/article/10.1007/s11042-018-5772-4#citeas>.
Yang, Zhilin; Dai, Zihang; Yan, Yiming; Caronell, Jaime; Salakhutdinov, Ruslan; Le, Quoc V, XLNet: Generalized Autoregressive Pretraining for Language Understanding, ArXiv, 2020. Available at: <https://arxiv.org/pdf/1906.08237>.

Yin, Hujun; Tang, Ke; Gao, Yang; Klawonn, Frank; Lee, Minho; Li, Bin, Intelligent Data Engineering and Automated Learning – IDEAL, Springer,2013. Available at: <https://link.springer.com/book/10.1007/978-3-030-03496-2>.

Xiang, Rui; Weibo, Li; Hun, Yan, Research on BERT-Based Audit Entity Extraction Method, IEEE, 2021, Available at: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9638834>.

Zheng, Alice. Evaluating Machine Learning Models, Chapter 4. Sebastopol, CA: O'Reilly Media, 2015. Available at: <https://www.oreilly.com/library/view/feature-engineering-for/9781491953235/ch04.html>.