

O processamento de informações de ações militares da MONUSCO: uma proposta de utilização de inteligência artificial para a Prospecção Tecnológica.

Vinícius Melquíades Scaldaferrri da Cunha

Aderson Campos Passos

Ronaldo Moreira Salles

¹Programa de Pós-graduação em Engenharia de Defesa – Instituto Militar de Engenharia
(IME)

22.290-270 Praça General Tibúrcio, 80 — Rio de Janeiro – RJ – Brasil

cunha.vinicius@ime.eb.br

aderson@ime.eb.br

salles@ime.eb.br

Abstract. *The present study was conducted as a Knowledge Discovery in Databases (KDD) approach with the objective of carrying out Text Mining on MONUSCO reports to identify information for the benefit of Technology Foresight in that mission. The method used followed the KDD operational stages of Pre-Processing, Data Mining (Text Mining) and Post-Processing, using a Recurrent Neural Network and BERT transformer, to fulfill the textual classification tasks. As main results, the feasibility of using artificial intelligence for this purpose was observed, having obtained a satisfactory classification in the Combat Functions: Command and Control, Movement and Maneuver, Intelligence, Fires, Logistics and Protection. Furthermore, the usefulness of this aspect of processing textual data, obtained from a corpus made up of reports, was verified as a rich source of useful information for associating the results obtained with technological branches defined in a specific taxonomy. Finally, the present work proved to be a viable modality for carrying out Technology Foresight in areas of conflict with Text Mining, adopting Artificial Intelligence resources to meet the demands of Natural Language Processing.*

Keywords: *KDD, Text Mining, Technology Foresight, Neural Network and transformer.*

Resumo. *O presente estudo foi conduzido como uma abordagem de Knowledge Discovery in Databases (KDD) com o objetivo de realizar a Mineração Textual (Text Mining) em relatórios da MONUSCO para identificar informações em proveito da Prospecção Tecnológica naquela missão. O método utilizado seguiu as etapas operacionais KDD de Pré-Processamento, Mineração de Dados (Mineração Textual) e de Pós-Processamento, com a utilização de Rede Neural Recorrente e transformer BERT, para cumprir tarefa de classificação textual. Como principais resultados, observou-se a viabilidade da utilização da inteligência artificial para essa finalidade, tendo-se obtido uma classificação satisfatória nas Funções de Combate: Comando e Controle, Movimento e Manobra, Inteligência, Fogos, Logística e Proteção. Ainda, verificou-se a*

utilidade dessa vertente de processamento de dados textuais, obtidos de um Corpus formado por relatórios, como uma farta fonte de informações úteis para a associação dos resultados obtidos com ramos tecnológicos definidos em taxonomia específica. Por fim, o presente trabalho mostrou-se como uma modalidade viável para a realização de Prospecção Tecnológica em áreas de conflito com Mineração Textual, adotando-se recursos de Inteligência Artificial para atender às demandas de Processamento de Linguagem Natural.

Palavras-chave: KDD, Mineração Textual, Prospecção Tecnológica, Rede Neural e transformer.

1. Introdução

A condução de operações militares é uma dinâmica que, por meio do registro das atividades realizadas, pode oferecer diversas informações de interesse em nível tático e estratégico. O presente trabalho traz as técnicas da Mineração Textual, como etapa do processo *Knowledge Discovery in Database* – KDD (SAFHI; FRIKH; OUHBI, 2019) como instrumento de análise de relatórios de operações militares extraídos da Missão das Nações Unidas na República Democrática do Congo -MONUSCO, com o propósito de identificar oportunidades de Prospecção Tecnológica. Entre todas as missões conduzidas pela ONU¹, atualmente, a MONUSCO é a de maior vulto, pela quantidade de tropa empregada, e de maior complexibilidade, pela natureza do conflito com várias forças irregulares em território de difícil acesso, o que a torna seus relatórios extremamente atrativos para a Mineração de Dados (ALVAREZ-ESPADA; FUENTES-BARGUES; GONZÁLEZ-GAYA, 2022). Além disso, a MONUSCO mostra-se um ambiente de pesquisa relevante pela representatividade da reação da ONU para a estabilização de uma região, Leste da República Democrática do Congo, que foi palco do massacre de Ruanda, em 1994, que chocou o mundo pela brutalidade e número de mortos (KAHN; DENOV, 2022). Todo esse conjunto de fatores demonstra a importância dessa missão na geopolítica mundial, bem como no aproveitamento da análise de informações produzidas em seus relatórios operativos, uma vez que apresenta um ambiente de operações desafiador e com enorme carência de recursos, o que é propício para a identificação de inovações e de soluções tecnológicas (CHO; JU, 2023). Dessa forma, foi proposto o seguinte problema: em que medida a Mineração Textual de relatórios de operações militares pode contribuir para a Prospecção Tecnológica?

¹ <https://peacekeeping.un.org/es/mission/monusco>

Foi definido o uso das Funções de Combate como elementos de classificação das ações militares constantes nos relatórios analisados da MONUSCO, para ordenar os seguimentos operativos identificados e auxiliar na posterior associação com ramos de tecnologia na área de atuação da MONUSCO. Essas Funções podem ser definidas como um conjunto de fatores relacionados às tarefas, atividades e sistemas interconectados aos diversos seguimentos estruturados na gestão operativa em uma área de combate. Possuem seis divisões: Comando e Controle, Movimento e Manobra, Inteligência, Fogos, Logística e Proteção (“MINISTÉRIO DA DEFESA EXÉRCITO BRASILEIRO ESTADO-MAIOR DO EXÉRCITO Manual de Fundamentos DOCTRINA MILITAR TERRESTRE 3 a Edição 2022”, [s.d.]).

O Comando e Controle é a função relacionada à capacidade comunicativa, por diversos meios, em níveis hierárquicos verticais e horizontais. O Movimento e Manobra referem-se à disposição de tropas no terreno para a adoção de ações táticas e estratégicas, bem como a utilização de meios de transporte de diversas modalidades. A função Inteligência liga-se à captura e processamento de dados, que podem resultar na geração de informações e de conhecimento sobre as operações. A função Fogos se atém à observância da utilização de fogo cinético, como o de Artilharia e de atuadores não cinéticos, como a Guerra Eletrônica, para neutralizar ou dificultar a ação do inimigo e favorecer a progressão da tropa amiga. A Logística compreende a gestão e transporte de material, em diversas classes, para suprir a permanência das tropas em combate. Por último, a função de combate Proteção visa preservar a integridade da tropa e da população civil por meio de ações e recursos que as preservem dos fogos inimigos (“MINISTÉRIO DA DEFESA EXÉRCITO BRASILEIRO ESTADO-MAIOR DO EXÉRCITO Manual de Fundamentos DOCTRINA MILITAR TERRESTRE 3 a Edição 2022”, [s.d.]).

O objetivo geral proposto é realizar a Mineração Textual (*Text Mining*) de relatórios da MONUSCO para identificar informações em proveito da Prospecção Tecnológica naquela missão. Nessa abordagem, foram aproveitados diversos relatórios disponibilizados no próprio site da missão, de forma aleatória a partir de 2010, para compor um *Corpus* para análise. Para atingir esse objetivo, foram definidos os seguintes objetivos específicos:

- a) Aplicar as etapas KDD, com a utilização de Inteligência Artificial (IA), em um *Corpus* formado por relatórios da MONUSCO;

- b) Realizar o Pré-processamento dos dados do Corpus de análise utilizando o software *WordSmith Tools* (WST);
- c) Realizar a Mineração Textual do Corpus de análise com uso de Rede Neural; e
- d) Realizar Prospecção Tecnológica, na fase de Pós-Processamento (com um *transformer* em BERT e com o WST), nos resultados de classificação obtidos por Funções de Combate, analisando as conexões de palavras-chave encontradas com uma Taxonomia Tecnológica.

A utilização de IA nesse estudo, para o processamento de informações extraídas de relatórios da MONUSCO, foi concebida por meio de uma Rede Neural com classificador binário (SIM/NÃO) para a identificação da relação dos textos com as Funções de Combate. Além disso, foram seguidas as 3 etapas operacionais KDD (GOLDSCHIMIDT; PASSOS, 2005): Pré-processamento, Mineração de Dados e Pós-Processamento. A organização deste artigo segue-se com os itens 2. Desenvolvimento, 3. Metodologia, 4. Discussão de Resultados e 5. Conclusão.

2. Revisão Teórica

2.1. Trabalhos Relacionados

O presente estudo difere-se de modelos mais consagrados na realização de Prospecção Tecnológica, como no uso de Inteligência Artificial para prospectar patentes, com análise de banco de dados textual (CHO; JU, 2023). A associação com outros trabalhos foi segmentada nos campos de estudo presentes nesse artigo, que apresenta uma formatação peculiar, uma vez que oferece um viés alternativo e inovador em sua concepção de aproveitamento de relatórios de missões militares para a identificação de áreas tecnológicas associadas à um teatro de operações.

A seguir, cabe mencionar alguns artigos que foram importantes na formulação desse estudo, com temática sobre uso de BERT e Rede Neural para Processamento de Linguagem Natural com funções classificatórias (BI et al., 2023). Esse *transformer*, igualmente, foi observado para a clusterização de um banco de dados textual e obtenção de melhor funcionamento da Rede Neural (GUAN et al., 2022).

Na Mineração Textual, foi observado o Processamento de Linguagem Natural para análise estratégica multipropósito, que pode incluir tarefas de prospecção (KUZMINOV et al., 2021). Igualmente, adicionou-se uma revisão de literatura sobre o suporte dado pela Prospecção Tecnológica ao gerenciamento de novas tecnologias (MINGHUI et al., 2021). Foi relevante considerar a extração de palavras-chave de banco de dados formado por documentos de dimensão distinta, aspecto relevante, pois pode dificultar o funcionamento da IA utilizada (SHIN; LEE; CHO, 2023).

Com a observação do *Corpus* de análise, coube mencionar um artigo sobre o impacto de novas tecnologias no trabalho com o uso de IA na otimização de funções laborais (FERNÁNDEZ-MACÍAS; BISELLO, 2022). Observaram-se pesquisas correlatas com aplicação da Mineração textual em Big Data para a descoberta de padrões e correlações escondidas em banco de dados (HASSANI et al., 2020). Na área da Linguística de Corpus, dentre os estudos observados, vale mencionar o que tratou sobre o aproveitamento de Banco de Dados Textual para o uso científico de *Corpus* Linguístico, processamento de Linguagem Natural (NLP) e *Information Retrieval*, todos muito conectados com a Mineração Textual (GERLACH; FONT-CLOS, 2020).

2.2 O processamento de informações

O processamento de informações textuais representou um grande desafio na utilização de aprendizado de máquina. Como já mencionado, os textos de entrada possuem tamanhos diferentes, o que implicou na normalização dos dados de forma vetorial, processo denominado *embedding* (LEE; LEE; MIN, 2023). Esse ajuste contribui no funcionamento da Rede Neural para o processamento de Linguagem Natural, uma vez que as matrizes normalizadas podem ser classificadas por padrões definidos em treinamento (BI et al., 2023). O *transformer* BERT, arquitetado em Python, foi utilizado na etapa de Pós-Processamento das informações extraídas (SOUZA; FILHO, 2023). Forneceu a descoberta de palavras-chave mais recorrentes nas classificações de extratos do *Corpus* de análise em Funções de Combate. A partir das descobertas dessas palavras, foram feitas as associações com a taxonomia de tecnologias de Defesa para a realização de Prospecção Tecnológica (EDA, 2021).

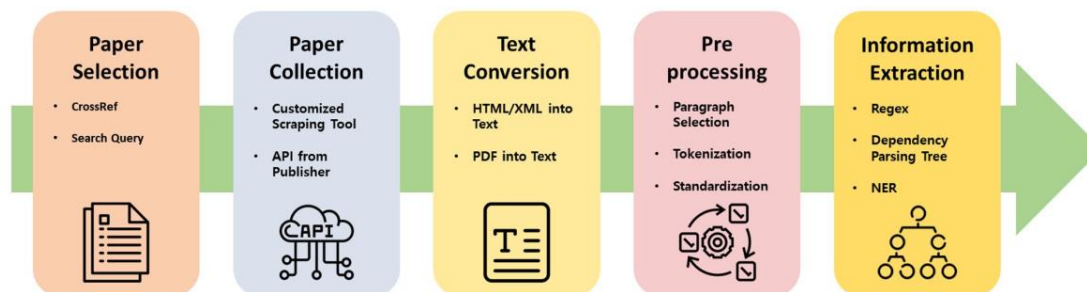


Figura 1. Sequência do processamento de informações de extratos do Corpus formado por relatórios da Monusco. Fonte:(LEE; LEE; MIN, 2023).

A sequência ilustrada acima orienta a aplicação das etapas KDD nesse estudo. O processo pode ser compreendido pela própria tradução do acrônimo em inglês como a Descoberta de Conhecimento em Bases de Dados, envolvendo várias etapas para a identificação de padrões compreensíveis, válidos, novos e potencialmente úteis de um banco de dados, nesse caso um *Corpus* composto por relatórios da MONUSCO (SAFHI; FRIKH; OUHBI, 2019). As 3 etapas operacionais são compostas pelo Pré-processamento das informações, onde ocorre a seleção, organização e limpeza dos dados, a Mineração de Dados, feita através da Mineração Textual nesse estudo e o Pós-Processamento, onde as informações obtidas são triadas e analisadas (GOLDSCHIMIDT; PASSOS, 2005).

Observando-se a referência da Figura 1, as fases *Paper Selection* e *Paper Collection* foram executadas com auxílio do software WordSmith Tools² para a montagem do *Corpus* e seleção textual para classificação, análise e Mineração de Dados em proveito da Prospecção Tecnológica. Nas etapas de *Text Conversion* e *Pre Processing* os textos selecionados foram convertidos em arquivos CSV e TXT em ZIP para processamento na Rede Neural Recorrente e no *transformer* BERT, para a classificação em Funções de Combate e para a identificação de palavras-chave mais recorrentes em cada uma delas. Por fim, na fase de *Information Extraction* foram feitas associações das palavras-chave mais

² <https://www.lexically.net/wordsmith/>

recorrentes por cada classificação com uma taxonomia de tecnologias de Defesa, para cumprir a finalidade de prospecção tecnológica no contexto da MONUSCO (EDA, 2021).

3. Metodologia

3.1 Tipo de Pesquisa

O presente trabalho abrange características qualitativas, pelo papel da função classificatória estabelecida para os textos seleccionados do Corpus, e aspectos quantitativos, na mensuração do desempenho da Rede Neural e para a análise dos resultados obtidos. A coleta de dados foi feita em campo, materializada pelos relatórios das ações da MONUSCO, extraídos, aleatoriamente, no próprio site da missão³.

A população é definida pela dimensão do corpus montado com uma compilação de relatórios da MONUSCO, somando um total de 202.545 palavras. A amostragem foi formada por um conjunto de 9 extratos textuais seleccionados do Corpus, somando um total de 411.710 palavras, visto que havia intersecções entre as partes analisadas, resultando um total maior que o do *Corpus*.

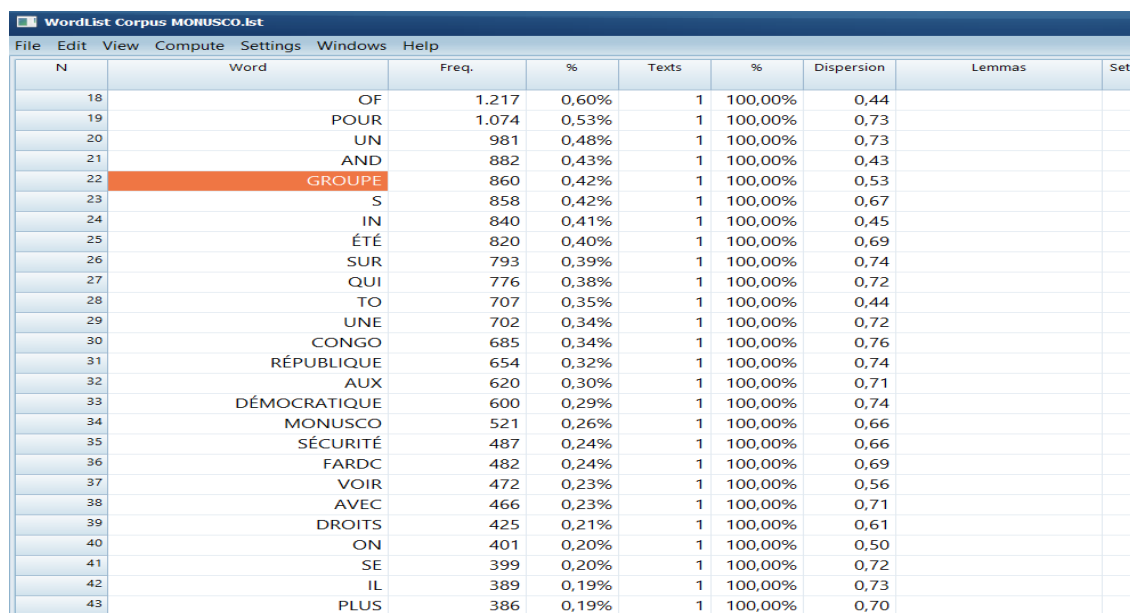
AMOSTRAGEM POR PALAVRA CHAVE	TOTAL DE PALAVRAS
ARMÉS	74601
CONTRE	21325
FORCES	121648
GROUPE	11314
MEMBRES	29776
MISSION	18577
OPÉRATIONS	109051
SÉCURITÉ	16225
TERRITOIRE	9193
TOTAL DA AMOSTRA	411710

Figura 2. Total de palavras analisadas nos contextos de cada palavra-chave extraída do *corpus*. (Fonte: o autor.)

³ <https://monusco.unmissions.org/en/un-reports>

3.2 A etapa de Pré-Processamento

Cada um dos textos escolhidos para análise representava o contexto de uma das 9 palavras mais recorrentes no *corpus*, tendo sido também considerado o valor semântico para a classificação em Funções de Combate. Com o auxílio do software WST⁴, foram identificadas essas palavras e foram extraídos contextos selecionados do uso delas no corpus. A partir desses contextos selecionados foram geradas as classificações em Funções de Combate: Comando e Controle, Movimento e Manobra, Inteligência, Fogos, Logística e Proteção. As 9 palavras selecionadas para a extração desses contextos foram, em versão original: groupe, sécurité, forces, mission, membres, contre, armés, territoire e opérations. A figura 3 demonstra o Wordlist do Corpus gerado no WST, com o destaque do vocábulo ‘groupe’, por frequência e semântica (análise do especialista executando o processo KDD). Na figura 6 foram geradas listas de concordância, de onde possibilitou-se a formação dos contextos selecionados ao redor de ‘groupe’. Na figura 4, segue uma imagem do contexto selecionado.



N	Word	Freq.	%	Texts	%	Dispersion	Lemmas	Set
18	OF	1.217	0,60%	1	100,00%	0,44		
19	POUR	1.074	0,53%	1	100,00%	0,73		
20	UN	981	0,48%	1	100,00%	0,73		
21	AND	882	0,43%	1	100,00%	0,43		
22	GROUPE	860	0,42%	1	100,00%	0,53		
23	S	858	0,42%	1	100,00%	0,67		
24	IN	840	0,41%	1	100,00%	0,45		
25	ÉTÉ	820	0,40%	1	100,00%	0,69		
26	SUR	793	0,39%	1	100,00%	0,74		
27	QUI	776	0,38%	1	100,00%	0,72		
28	TO	707	0,35%	1	100,00%	0,44		
29	UNE	702	0,34%	1	100,00%	0,72		
30	CONGO	685	0,34%	1	100,00%	0,76		
31	RÉPUBLIQUE	654	0,32%	1	100,00%	0,74		
32	AUX	620	0,30%	1	100,00%	0,71		
33	DÉMOCRATIQUE	600	0,29%	1	100,00%	0,74		
34	MONUSCO	521	0,26%	1	100,00%	0,66		
35	SÉCURITÉ	487	0,24%	1	100,00%	0,66		
36	FARDC	482	0,24%	1	100,00%	0,69		
37	VOIR	472	0,23%	1	100,00%	0,56		
38	AVEC	466	0,23%	1	100,00%	0,71		
39	DROITS	425	0,21%	1	100,00%	0,61		
40	ON	401	0,20%	1	100,00%	0,50		
41	SE	399	0,20%	1	100,00%	0,72		
42	IL	389	0,19%	1	100,00%	0,73		
43	PLUS	386	0,19%	1	100,00%	0,70		

Figura 3. Wordlist para identificação de palavras mais recorrentes no Corpus.

(Fonte: o autor)

⁴ <https://www.lexically.net/wordsmith/>

Concord									
File Edit View Compute Settings Windows Help									
N	Concordance					Set	Tag	Sent. #	Para. #
34	pendant 15 jours, les corps des deux membres du	Groupe	d'experts ont été retrouvés dans une tombe					0	0
35	que l'enquête sur le meurtre des deux membres du	Groupe	d'experts était close et l'affaire renvoyée au					0	0
36	octobre 2016 de Gédéon Kyungu Mutanga, le chef du	Groupe	armé Mai-Mai Kata Katanga. 1>22. Au cours de					1.727	20
37	national pour le renouveau de la démocratie	(CNRD),	dissident des FDLR, et des éléments des					0	0
38	mai, 50 pêcheurs ont été enlevés par des	membres du	Mai-Mai Charles. Dans le territoire de Masisi,					1.750	11
39	pour la sécurité des civils. Pour le seul mois de	mai, ce	a été impliqué dans 31 cas de violations des					0	0
40	de Rutshuru, dans une cache appartenant à	l'ancien	armé Mouvement du 23 mars. Il s'agissait de					1.784	9
41	En outre, la	MONUSCO s'emploie à déployer	d'évaluation de la menace militaire, qui					0	0
42	notamment la Légion nationale	d'intervention et le	mobile d'intervention, ce qui pourrait					0	0
43	Sharp and Zaida Catalán, nos deux collègues	du	d'experts sur la République démocratique du					0	0
44	siens, Zaida Catalán et Michael Sharp, nos collègues	du	d'experts sur la République démocratique du					0	0
45	Président du Rassemblement et Pierre	Lumbi	des sept, coalition de l'opposition, a pris la tête					0	0
46	le renouveau et	la démocratie-Ubwiyunge (CNRD),	dissident, les Forces armées de la					0	0
47	unilatérales pour maintenir une	pression militaire sur le	armé en ciblant ses dirigeants et ses principales					0	0
48	Le 27 janvier 2017, suite à des rapports selon lesquels	un	d'exélements du M23 étaient entrés dans le					0	0
49	novembre. Député, M. Badibanga est le président	du	parlementaire					0	0
50	poste de Premier Ministre dévolu à Samy Badibanga	du	«UDPS et Alliés », une coalition					0	0
51	appuyé	les opérations menées par les FARDC contre le	groupe,					0	0
52	de 10 de leurs membres, notamment 2 responsables	du	L'opération a également permis de récupérer					2.267	65
53	ont été identifiées comme étant des enfants associés	au	et en ont été séparées; 58 autres ont accepté					0	0
54	les FARDC. Au cours de la dernière décennie,	le	a perpétré					0	0
55	de défense et de sécurité déployées pour éliminer	le	Lors de 17 incidents différents, pas moins de					2.342	78
56	Kata-Katanga, mais n'ont pas encore été séparés	de	armé. De plus, 28 garçons ont été recensés					2.360	66
57	joué un rôle déterminant dans la défaite militaire	du	armé					0	0
58	MONUSCO n'est toujours pas parvenue à déployer	un	d'analyse des menaces militaires, qui pourrait					0	0
59	à créer une équipe	d'intervention immédiate,	un					0	0
60	Après avoir aidé les autorités nationales à venir à bout	du	armé					0	0
61	forces ont porté un sévère	coup aux activités de	ce					0	0
62	de vous transmettre ci-joint	le rapport final	du					2.463	20
63	la République démocratique du Congo par	le	d'experts sur la République démocratique du					0	0
64	République démocratique du Congo	Les membres du	d'experts sur la République démocratique du					0	0

Figura 4. Concordance List da palavra *groupe*. (Fonte: o autor)

Corpus UN MONUSCO.txt

GROUPE

419

↑

↓

4

line 5.001 to 10.000 (segment 2/4)

5. Le **Groupe** d'experts pense que les ADF ont toujours la capacité de se réorganiser et de se renforcer, comme elles l'ont déjà fait à la suite d'opérations similaires en 2005 et en 2010 (voir annexe 5 et S/2014/428, par. 7). Les commandants du **groupe** ne participant généralement pas aux combats, ils ont pour la plupart survécu à ceux de 2014. En outre, les réseaux de recrutement, de soutien et de financement des ADF n'ont pas été beaucoup touchés.

6. Les informations figurant dans la présente section du rapport proviennent des entretiens que les membres du **Groupe** d'experts ont eus avec 42 anciens membres des ADF, dont 27 ex-combattants, 11 civils et 4 bazana (voir annexe 6). Le mot bazana (pluriel de muzana), qui vient du Luganda et signifie « esclaves capturés durant la guerre », est utilisé pour désigner les personnes enlevées et réduites en esclavage par les ADF (voir annexe 7). Le **Groupe** a aussi interrogé deux personnes recrutées par les ADF en 2014, mais qui ont été arrêtées avant d'avoir pu rallier le **groupe** armé. Il s'est rendu dans des camps de ce dernier à trois reprises en avril et en mai 2014, et a analysé des centaines de pages de documents qu'il avait saisis avec le concours de l'armée congolaise (voir S/2014/428, par. 29 à 34). Il a obtenu des informations supplémentaires de sources locales à Beni et à Oicha, d'officiers de l'armée congolaise, de membres du personnel de la MONUSCO et de hauts fonctionnaires de la République démocratique du Congo, de l'Ouganda et du Royaume-Uni de Grande-Bretagne et d'Irlande du Nord. Il a aussi consulté le **Groupe** de contrôle pour la Somalie et l'Érythrée et l'Équipe d'appui analytique et de surveillance des sanctions créée par la résolution 1526 (2004) du Conseil de sécurité.

Add

Clear hits

Figura 5. Contexto selecionado. (Fonte: o autor)

A Figura 7 representa o produto final, que foi inserido na Rede Neural e no *Transformer* BERT. O tratamento do *Corpus* no WST, representado na figura 8, foi executado da maneira

descrita para que fossem aproveitadas as informações mais relevantes dentre um vasto número de palavras disponíveis. O reflexo desse procedimento é percebido no funcionamento da Rede Neural Recorrente, no *transformer* BERT e no tratamento das informações obtidas para a Prospecção Tecnológica na etapa de Pós-processamento.



Figura 6. WordSmith Tools (Fonte: <https://www.lexically.net/wordsmith/>)

3.3 A etapa de Mineração de Dados (Mineração Textual) com Rede Neural Classificatória

A Mineração de Dados foi realizada baseando-se em Rede Neural Recorrente, para a classificação textual em Funções de Combate. Além disso, na etapa de Pós-Processamento, a associação às taxonomias tecnológicas foi facilitada devido a essa classificação. Inicialmente, os textos foram ‘tokenizados’ com o *transformer* BERT. Essa medida permitiu a normalização dos dados de entrada. A arquitetura da rede foi configurada com textos de treinamentos afins a cada Função de Combate, para a classificação propriamente dita. Algumas palavras-chave foram selecionadas, de cada uma das seis Funções de Combate,

para provocar a classificação através da semelhança de elementos textuais ‘tokenizados’ nas matrizes.

```
# Run the main function
if __name__ == '__main__':
    # Define the classification parameters
    fogos = ["explosivo", "munição", "mina", "projétil", "tiros", "morteiros", "obuses", "artilharia", "granadas", "foguetes", "projéteis", "canhões", "artilharia", "barragem"]
    logistica = ["território", "cache", "munição", "Norte", "Sul", "Oeste", "Leste", "armas", "combustível", "transporte", "carro", "barco", "estrada"]
    monobra = ["ataque", "coordenada", "assalto", "emboscada", "marcha", "marcha de combate", "flanco", "penetração", "infiltração", "cerco", "infantaria", "cavalaria", "fogo"]
    inteligencia = ["informação", "dados", "contra-espionagem", "pesquisa", "reconhecimento", "inimigo", "posição do inimigo", "dispositivo", "terreno", "vegetação", "clima"]
    CC = ["rádio", "comunicação", "transmissão", "guerra eletrônica", "satélite", "Internet", "AM", "FM", "telefone", "linhas", "rede", "diagrama"]
    protecao = ["proteção", "armadura", "camuflagem", "ocultação", "antiaérea", "preservação", "preventiva", "cobertura"]

    main(fogos, logistica, monobra, inteligencia, CC, protecao)
```

Figura 7. Palavras-chave para treinamento da Rede Neural. (Fonte: o Autor)

Foi adotada a biblioteca *Natural Language Toolkit* - NLTK para a remoção de *stopwords*, o que contribuiu para a eficiência da classificação. Os textos da amostra selecionada foram usados no treinamento da rede, que se mostrou eficiente na função classificatória.

Texto	Predict - Fogos	Predict - logistica	Predict - monobra	Predict - inteligencia	Predict - CC	Predict - protecao
versão origin rádio "", "" comunicação ", "	0	0	1	1	1	0
5 est projeto lei devolvido parlamento preside...	0	0	1	1	1	0
19 commentair l ' artici premier du code condu...	0	0	0	1	1	1
31 . procurar todo custo impor restriçõ direit...	0	0	1	1	1	0
41a/hrc/31/66 par . 53. , 42 acordo test propo...	0	0	1	1	1	0
...	...	...	...	...	...	...
rdc colocado sob comando subchef departamento ...	0	0	1	1	1	0
estado necesidad decretado chefe estado . est...	1	0	1	1	1	0
ada yumbi bolobo pequena , possuem trabalho bã...	1	0	1	1	1	0
uadra territori único local poder instruir pro...	1	0	1	1	1	0

Figura 8. Resultado de classificação textual (Fonte: o Autor)

A Rede Neural concebida para o presente estudo foi arquitetada com duas camadas de 16 e 8 neurônios para cada função classificatória binária, sendo ‘1’ para sim e ‘0’ para não, em cada Função de Combate.

```

Classificador Fogos
Epoch 1/3
49/49 [=====] - 157s 3s/step - loss: 0.6995 - accuracy: 0.5672
Epoch 2/3
49/49 [=====] - 155s 3s/step - loss: 0.4668 - accuracy: 0.8649
Epoch 3/3
49/49 [=====] - 151s 3s/step - loss: 0.3071 - accuracy: 0.9169
1/1 [=====] - 1s 651ms/step - loss: 1.1017 - accuracy: 0.5196
5/5 [=====] - 2s 240ms/step
Classificador logistica
Epoch 1/3
8/8 [=====] - 26s 3s/step - loss: 0.6967 - accuracy: 0.5056
Epoch 2/3
8/8 [=====] - 23s 3s/step - loss: 0.6751 - accuracy: 0.6136
Epoch 3/3
8/8 [=====] - 23s 3s/step - loss: 0.6668 - accuracy: 0.6248
1/1 [=====] - 1s 625ms/step - loss: 1.0880 - accuracy: 0.0125
5/5 [=====] - 2s 276ms/step
Classificador monobra
Epoch 1/3
9/9 [=====] - 32s 3s/step - loss: 0.7541 - accuracy: 0.4799
Epoch 2/3
9/9 [=====] - 28s 3s/step - loss: 0.7001 - accuracy: 0.4668
Epoch 3/3
9/9 [=====] - 29s 3s/step - loss: 0.6510 - accuracy: 0.6546
1/1 [=====] - 1s 614ms/step - loss: 0.6684 - accuracy: 0.3853
5/5 [=====] - 1s 235ms/step

```

Figura 9. Característica da Rede Neural (Fonte: o Autor)

A imagem acima ilustra que o treinamento da Rede Neural foi feito em 3 épocas, com três ciclos de treinamento e dois de teste. A acurácia oscilou significativamente, em virtude do número de dados de treinamento utilizados. Entretanto, a variação dos resultados obtidos indicou baixa possibilidade de *overfitting*, aspecto não desejado. Do total do conjunto de dados, 10,6% foram utilizados como conjunto de treinamento e 2,6% para teste. As características das matrizes, formadas pelos textos ‘tokenizados’, expostas abaixo ilustram a proporção. A matriz de entrada foi dividida, mantendo-se a quantidade de colunas, com cortes na quantidade de linhas.

```

Matriz: [[ 101 8085 6240 ... 0 0 0]
[ 101 1976 22283 ... 0 0 0]
[ 101 8047 119 ... 0 0 0]
...
[ 101 602 22278 ... 0 0 0]
[ 101 169 142 ... 0 0 0]
[ 101 245 3094 ... 0 0 0]]
Matriz treino: [[ 101 119 17514 ... 0 0 0]
[ 101 10140 119 ... 0 0 0]
[ 101 15912 22280 ... 0 0 0]
...
[ 101 12664 1167 ... 0 0 0]
[ 101 11365 6636 ... 0 0 0]
[ 101 11662 22278 ... 0 0 0]]
Matriz test: [[ 101 732 117 ... 0 0 0]
[ 101 18227 119 ... 0 0 0]
[ 101 275 119 ... 0 0 0]
[ 101 1592 425 ... 0 0 0]]

```

Figura 10. Matrizes de entrada, treino e teste. (Fonte: o Autor)

```

Dimensão Matriz: 414193
Dimensão Matriz: (151, 2743)

Dimensão Matriz Treino: 43888
Dimensão Matriz Treino: (16, 2743)

Dimensão Matriz Test: 10972
Dimensão Matriz Test: (4, 2743)

```

Figura 11. Dimensão das matrizes. (Fonte: o Autor)

As matrizes obtidas demonstraram ter as linhas Linearmente Independentes, $m \times n$, onde $n > m$, e o Posto era sempre m . Esse fator contribuiu para que fosse demonstrado matematicamente que a ‘*tokenização*’ estava coerente com a característica dos textos inseridos na rede. Caso houvesse linhas Linearmente Dependentes, seria um indicador de linhas repetidas, ou de problema na tokenização.

```

O posto da matriz: 150
O posto da matriz train: 16
O posto da matriz test: 4

```

Figura 12. Posto das Matrizes. (Fonte: o Autor)

3.4 A etapa de pós-processamento

No Pós-Processamento, após a conclusão da Mineração de Textual, onde foram executadas as classificações dos textos por Funções de Combate, foram identificadas as palavras-chave mais recorrentes dentre as 6 funções. Essas palavras foram as responsáveis por ativar as classificações e em seguida foram compiladas para que fossem ordenadas por frequência, por meio do WST. As mais frequentes foram utilizadas na Prospeção Tecnológica, pela associação, feita por análise comparativa, com a taxonomia de referência. O classificador, além da identificação das Funções de Combate, forneceu as palavras/tokens que ativaram as classificações.

PalavrasChaves	IdPalavrasChaves	Classificador
['munição']	[[20614]]	Fogos
['território', 'munição']	[[2410], [20614]]	Logística
['marcha', 'marcha de combate']	[[10803], [10803]]	MovimentoManobra
['informação', 'contra-espionagem']	[[4205], [598]]	Inteligência
[]	[]	Não Identificado - Comandocontrole
[]	[]	Não Identificado - Proteção

Figura 13. Palavras-chave. (Fonte: o Autor)

4. Discussão de Resultados

Os resultados foram obtidos seguindo-se as etapas KDD para a análise de um *Corpus* linguístico formado por relatórios de ações militares da MONUSCO, e mostraram que a Mineração Textual pode contribuir, produtivamente, para a realização de Prospecção Tecnológica.

O objetivo geral de realizar a Mineração Textual de relatórios da MONUSCO para identificar informações em proveito da Prospecção Tecnológica foi atingido utilizando-se uma Rede Neural, desenvolvida para Processamento de Linguagem Natural, com destaque para o uso das bibliotecas NLTK, BERT e *Tensorflow*, do *Python*. O classificador desenvolvido foi suficiente para a classificação em Funções de Combate, bem como na identificação de palavras-chave mais recorrentes em cada uma dessas funções, informação crucial para o estabelecimento da Prospecção Tecnológica com a taxonomia de tecnologias de Defesa.

Foi realizado um teste de verificação desta capacidade do classificador com a inserção, proposital, de um texto com características simétricas entre as linhas, como ilustrado abaixo:

recorrentes por cada Função de Combate e, em seguida, foram geradas as associações com as tecnologias disponíveis na taxonomia de referência, por análise do autor no processo KDD. Essas conexões são subjetivas, e indicam apenas uma sugestão de tecnologia que pode ser útil em relação a cada palavra-chave identificada, levando em conta o cenário da MONUSCO. A palavra ‘ALFA’ foi usada para ativar a formação do Wordlist, devendo ser desconsiderada. Seguem abaixo os resultados obtidos:

WORDLIST CC.lst								
File Edit View Compute Settings Windows Help								
N	Word	Freq.	%	Texts	%	Dispersion	Lemmas	Set
1	ALFA	4.480	99,09%	1	100,00%	0,77		
2	ELETRÔNICA	9	0,20%	1	100,00%	0,00		
3	GUERRA	9	0,20%	1	100,00%	0,00		
4	RÁDIO	7	0,15%	1	100,00%	0,00		
5	REDE	6	0,13%	1	100,00%	0,00		
6	COMUNICAÇÃO	5	0,11%	1	100,00%	0,00		
7	TRANSMISSÃO	5	0,11%	1	100,00%	0,00		

Figura 16. Comando e Controle. (Fonte: o autor)

5 maiores frequências de palavras-chave identificadas na Função Comando e Controle vs. Possibilidades tecnológicas	
Palavras-chave para Comando e Controle	TAXONOMIA TECNOLÓGICA
GUERRA ELETRÔNICA	SPEECH RECOGNITION
RÁDIO	HUMAN LANGUAGE TECHNOLOGIES (HLT)
REDE	DATA FUSION
COMUNICAÇÃO	SYSTEM ON A CHIP (SoC)
TRANSMISSÃO	ELECTROMAGNETIC MODELLING OPTIMIZATION (EMO)

Figura 17. Prospecção Tecnológica em Comando e Controle. (Fonte: o autor)

WL RESULTADO Fogos.lst								
File Edit View Compute Settings Windows Help								
N	Word	Freq.	%	Texts	%	Dispersion	Lemmas	Set
1	ALFA	4.480	97,84%	1	100,00%	0,77		
2	ARTILHARIA	12	0,26%	1	100,00%	0,00		
3	DE	12	0,26%	1	100,00%	0,00		
4	EXPLOSIVO	8	0,17%	1	100,00%	0,00		
5	GRANADAS	8	0,17%	1	100,00%	0,00		
6	MUNIÇÃO	8	0,17%	1	100,00%	0,00		
7	MINA	7	0,15%	1	100,00%	0,00		
8	MORTEIROS	7	0,15%	1	100,00%	0,00		
9	FOGO	6	0,13%	1	100,00%	0,00		
10	PROJÉTEIS	5	0,11%	1	100,00%	0,00		
11	PROJÉTIL	5	0,11%	1	100,00%	0,00		
12	Ã	4	0,09%	1	100,00%	0,00		
13	CONCENTRAÇÃO	4	0,09%	1	100,00%	0,00		
14	O	4	0,09%	1	100,00%	0,00		
15	CONCENTRAÃ	2	0,04%	1	100,00%	0,00		
16	MUNIÃ	2	0,04%	1	100,00%	0,00		
17	PROJÃ	2	0,04%	1	100,00%	0,00		
18	OBUSES	1	0,02%	1	100,00%	0,00		
19	TEIS	1	0,02%	1	100,00%	0,00		
20	TIL	1	0,02%	1	100,00%	0,00		

Figura 18. Fogos. (Fonte:o autor)

5 maiores frequências de palavras-chave identificadas na Função Fogos vs. Possibilidades tecnológicas	
Palavras-chave para Fogos	TAXONOMIA TECNOLÓGICA
ARTILHARIA	DETECTION - SEGMENTATION
EXPLOSIVO/GRANADAS	PATTERN RECOGNITION
MINAS	CLUSTERING
MORTEIROS	PROBABILISTIC LOGIC PROGRAMMING
CONCENTRAÇÃO (DE FOGOS)	REAL TIME TECHNOLOGIES

Figura 19. Prospecção Tecnológica em Fogos. (Fonte: o autor)

WORDLIST INTEL.lst								
File Edit View Compute Settings Windows Help								
N	Word	Freq.	%	Texts	%	Dispersion	Lemmas	Set
1	ALFA	4.480	98,22%	1	100,00%	0,77		
2	INIMIGO	11	0,24%	1	100,00%	0,00		
3	CONTRA	9	0,20%	1	100,00%	0,00		
4	ESPIONAGEM	9	0,20%	1	100,00%	0,00		
5	DISPOSITIVO	8	0,18%	1	100,00%	0,00		
6	PESQUISA	8	0,18%	1	100,00%	0,00		
7	CLIMA	7	0,15%	1	100,00%	0,00		
8	DO	7	0,15%	1	100,00%	0,00		
9	INFORMAÇÃO	7	0,15%	1	100,00%	0,00		
10	POSIÇÃO	7	0,15%	1	100,00%	0,00		
11	TERRENO	5	0,11%	1	100,00%	0,00		
12	RECONHECIMENTO	3	0,07%	1	100,00%	0,00		

Figura 20. Inteligência. (Fonte: o autor)

5 maiores frequências de palavras-chave identificadas na Função Inteligência vs. Possibilidades tecnológicas	
Palavras-chave para Inteligência	TAXONOMIA TECNOLÓGICA
INIMIGO	DATA WAREHOUSES
CONTRA ESPIONAGEM	REAL TIME TECHNOLOGIES
DISPOSITIVO (POSICIONAMENTO DO INIMIGO)	IMAGE DATA
PESQUISA	BIG DATA ANALYTICS
CLIMA	PREDICTIVE ANALYTICS

Figura 21. Prospecção Tecnológica em Inteligência. (Fonte: o autor)

WORDLIST LOG.lst								
File Edit View Compute Settings Windows Help								
N	Word	Freq.	%	Texts	%	Dispersion	Lemmas	Set
1	ALFA	4.480	99,09%	1	100,00%	0,77		
2	TERRITÓRIO	9	0,20%	1	100,00%	0,00		
3	MUNIÇÃO	8	0,18%	1	100,00%	0,00		
4	ESTRADA	7	0,15%	1	100,00%	0,00		
5	COMBUSTÍVEL	6	0,13%	1	100,00%	0,00		
6	BARCO	5	0,11%	1	100,00%	0,00		
7	CARRO	4	0,09%	1	100,00%	0,00		
8	CACHE	2	0,04%	1	100,00%	0,00		

Figura 22. Logística. (Fonte: o autor)

5 maiores frequências de palavras-chave identificadas na Função Logística vs. Possibilidades tecnológicas	
Palavras-chave para Logística	TAXONOMIA TECNOLÓGICA
TERRITÓRIO	PATHFINDING OR PATH PLANNING
MUNIÇÃO	PATTERN RECOGNITION
ESTRADA	IMAGE ACQUISITION
COMBUSTÍVEL	LOGISTIC REGRESSION
BARCO	PATHFINDING ALGORITHMS

Figura 23. Prospecção Tecnológica em Logística (Fonte: o autor)

WORDLIST MOV MANOBRA.lst								
File Edit View Compute Settings Windows Help								
N	Word	Freq.	%	Texts	%	Dispersion	Lemmas	Set
1	ALFA	4.480	98,70%	1	100,00%	0,77		
2	FOGO	9	0,20%	1	100,00%	0,00		
3	PREPARATÓRIO	9	0,20%	1	100,00%	0,00		
4	MARCHA	6	0,13%	1	100,00%	0,00		
5	ASSALTO	5	0,11%	1	100,00%	0,00		
6	COORDENADA	5	0,11%	1	100,00%	0,00		
7	EMBOSCADA	5	0,11%	1	100,00%	0,00		
8	INFANTARIA	5	0,11%	1	100,00%	0,00		
9	PENETRAÇÃO	5	0,11%	1	100,00%	0,00		
10	COMBATE	3	0,07%	1	100,00%	0,00		
11	DE	3	0,07%	1	100,00%	0,00		
12	INFILTRAÇÃO	3	0,07%	1	100,00%	0,00		
13	CAVALARIA	1	0,02%	1	100,00%	0,00		

Figura 24. Movimento e Manobra. (Fonte: o autor)

5 maiores frequências de palavras-chave identificadas na Função Movimento e Manobra vs. Possibilidades tecnológicas	
Palavras-chave para Movimento e Manobra	TAXONOMIA TECNOLÓGICA
FOGO PREPARATÓRIO	FUZZY LOGIC
MARCHA	METAHEURISTICS
ASSALTO	GEOTAGGING
EMBOSCADA	IMAGE ACQUISITION
INFANTARIA	MILITARY DOCTRINE ANALYSIS

Figura 25. Prospeção Tecnológica em Movimento e Manobra. (Fonte: o autor)

WORDLIST PROTEÇÃO.lst								
File Edit View Compute Settings Windows Help								
N	Word	Freq.	%	Texts	%	Dispersion	Lemmas	Set
1	ALFA	4.480	99,07%	1	100,00%	0,77		
2	PROTEÇÃO	9	0,20%	1	100,00%	0,00		
3	PREVENTIVA	8	0,18%	1	100,00%	0,00		
4	ANTIAÉREA	6	0,13%	1	100,00%	0,00		
5	OCULTAÇÃO	6	0,13%	1	100,00%	0,00		
6	PRESERVAÇÃO	6	0,13%	1	100,00%	0,00		
7	CAMUFLAGEM	5	0,11%	1	100,00%	0,00		
8	COBERTURA	2	0,04%	1	100,00%	0,00		

Figura 26. Proteção. (Fonte: o autor)

5 maiores frequências de palavras-chave identificadas na Função Proteção vs. Possibilidades tecnológicas	
Palavras-chave para Proteção	TAXONOMIA TECNOLÓGICA
PROTEÇÃO (HR E ESTRUTURAS)	DATA FUSION
PREVENTIVA	TEXT ANALYTICS
ANTIAÉREA	DETECTION
OCULTAÇÃO	IMAGE - PROCESSING
PRESERVAÇÃO (HR E ESTRUTURAS)	PREDICTIVE ANALYTICS

Figura 27. Prospecção Tecnológica em Proteção. (Fonte: o autor)

5. Conclusão

Conclui-se que o presente trabalho obteve resultados suficientes para responder ao problema proposto, visto que a execução completa do Processo KDD, com as três etapas operacionais, garantiu a eficiência da Mineração Textual no Corpus formado por relatórios da MONUSCO.

Os recursos fornecidos pela Rede Neural Recorrente e pelo classificador vetorizado no BERT permitiram uma adequada classificação por Funções de Combate, que se mostrou relevante para segmentar as informações dos relatórios e posteriormente relacioná-las aos campos tecnológicos elencados pela taxonomia adotada. O modelo desenvolvido neste estudo permitiu uma associação particularizada das ações militares da MONUSCO a setores tecnológicos presentes naquele teatro de operações, tendo fornecido informações que possibilitassem uma maior consciência situacional em relação àquela missão específica. Conclui-se que essa modalidade de Prospecção Tecnológica por Mineração Textual complementa outras formas clássicas como, por exemplo, a análise de patentes. O uso periódico do recurso proposto nesta pesquisa pode contribuir, igualmente, para a melhoria em processos de apoio à decisão.

Por fim, a presente pesquisa apresenta futuras possibilidades para o aproveitamento do Processamento de Linguagem Natural como ferramenta eficiente para a execução de Prospecção Tecnológica, utilizando-se a Mineração Textual, com recursos de Inteligência Artificial na classificação textual de fontes oriundas de outras áreas de conflito.

6. Referências Bibliográficas

ALVAREZ-ESPADA, J. M.; FUENTES-BARGUES, J. L.; GONZÁLEZ-GAYA, C. Approach and Success in the Management of Peacekeeping Operations (PKOs): Application to Two Case Studies, The UNMISS and MONUSCO Missions of the UN. **Sustainability (Switzerland)**, v. 14, n. 10, 1 maio 2022.

BI, H. et al. EnvText: A Chinese text mining tool for environmental domain with advanced BERT model[Formula presented]. **Software Impacts**, v. 17, 1 set. 2023.

CHO, I.; JU, Y. Text mining method to identify artificial intelligence technologies for the semiconductor industry in Korea. **World Patent Information**, v. 74, p. 102212, set. 2023.

EDA. **OSRA Defence Technology Taxonomy v2.0**. [s.l: s.n.]. Disponível em: <www.eda.europa.eu>.

FERNÁNDEZ-MACÍAS, E.; BISELLO, M. A Comprehensive Taxonomy of Tasks for Assessing the Impact of New Technologies on Work. **Social Indicators Research**, v. 159, n. 2, p. 821–841, 1 jan. 2022.

GERLACH, M.; FONT-CLOS, F. A standardized project gutenber corpus for statistical analysis of natural language and quantitative linguistics. **Entropy**, v. 22, n. 1, p. 126, 1 jan. 2020.

GUAN, R. et al. Deep Feature-Based Text Clustering and its Explanation. **IEEE Transactions on Knowledge and Data Engineering**, v. 34, n. 8, p. 3669–3680, 1 ago. 2022.

HASSANI, H. et al. Text Mining in Big Data Analytics. **Big Data and Cognitive Computing**, v. 4, n. 1, p. 1, 16 jan. 2020.

KAHN, S.; DENOV, M. Transgenerational trauma in Rwandan genocidal rape survivors and their children: A culturally enhanced bioecological approach. **Transcultural Psychiatry**, v. 59, n. 6, p. 727–739, 1 dez. 2022.

KUZMINOV, I. F. et al. Modern Natural Language Processing Technologies for Strategic Analytics. **Scientific and Technical Information Processing**, v. 48, n. 6, p. 467–475, 1 dez. 2021.

LEE, J. H.; LEE, M.; MIN, K. Natural Language Processing Techniques for Advancing Materials Discovery: A Short Review. **International Journal of Precision Engineering**

and Manufacturing - Green Technology Korean Society for Precision Engineering, , 1 set. 2023.

MINGHUI, Z. et al. **Literature Review and Practice Comparison of Technology Foresight**. Procedia Computer Science. **Anais...**Elsevier B.V., 2021.

MINISTÉRIO DA DEFESA EXÉRCITO BRASILEIRO ESTADO-MAIOR DO EXÉRCITO Manual de Fundamentos DOCTRINA MILITAR TERRESTRE 3 a Edição 2022. . [s.l: s.n.].

RYAN, C. S. et al. Sustainable strategies for Ebola virus disease outbreak preparedness in Africa: a case study on lessons learnt in countries neighbouring the Democratic Republic of the Congo. **Infectious Diseases of Poverty**, v. 11, n. 1, 1 dez. 2022.

SAFHI, H. M.; FRIKH, B.; OUHBI, B. **Assessing reliability of Big Data Knowledge Discovery process**. Procedia Computer Science. **Anais...**Elsevier B.V., 2019.

SHIN, H.; LEE, H. J.; CHO, S. General-use unsupervised keyword extraction model for keyword analysis. **Expert Systems with Applications**, p. 120889, dez. 2023.

SOUZA, F. D.; FILHO, J. B. DE O. E. S. Embedding generation for text classification of Brazilian Portuguese user reviews: from bag-of-words to transformers. **Neural Computing and Applications**, v. 35, n. 13, p. 9393–9406, 1 maio 2023.

ZHU, H.; LEI, L. The Research Trends of Text Classification Studies (2000–2020): A Bibliometric Analysis. **SAGE Open**, v. 12, n. 2, 1 abr. 2022.