Em Questão
ISSN: 1807-8893
ISSN: 1808-5245
emquestao@ufrgs.br
Universidade Federal do Rio Grande do Sul
Brasil

Absorção das tarefas de processamento de Linguagem Natural (NLP) pela Ciência da Informação (CI): uma revisão da literatura para tangibilização do uso de NLP pela CI

Falcão, Luander Cipriano de Jesus; Lopes, Brenner; Souza, Renato Rocha
Absorção das tarefas de processamento de Linguagem Natural (NLP) pela Ciência da Informação (CI): uma revisão da literatura para tangibilização do uso de NLP pela CI
Em Questão, vol. 28, núm. 1, 2022
Universidade Federal do Rio Grande do Sul, Brasil
Disponible en: https://www.redalyc.org/articulo.oa?id=465669358002

Absorção das tarefas de processamento de Linguagem Natural (NLP) pela Ciência da Informação (CI): uma revisão da literatura para tangibilização do uso de NLP pela CI

Absorption of Natural Language Processing (NLP) Tasks by Information Science (IS): a literature review for tangibilizing the use of NLP by IS

Luander Cipriano de Jesus Falcão 1
Universidade Federal de Minas Gerais, Brasil
luanderfalcao@yahoo.com.br

Brenner Lopes 2
Universidade Federal de Minas Gerais, Brasil
brenner.lopes@gmail.com

Renato Rocha Souza 3
Universidade Federal de Minas Gerais, Brasil
rsouzaufmg@gmail.com

Redalyc: https://www.redalyc.org/articulo.oa?id=465669358002

Recepción: 02 Febrero 2021
Aprobación: 30 Abril 2021

Resumo:

Um dos recentes desafios da abordagem denominada *Big Data* tem sido extrair informações relevantes de grandes quantidades de dados não estruturados, como por exemplo de textos escritos em diversos idiomas. A principal abordagem de análise de texto e linguagem por meio computacional é chamada de Processamento de Linguagem Natural (*Natural Language Processing* - NLP, na sigla em inglês). Identificar como as áreas do conhecimento estão utilizando as evoluções em seus domínios, no que tange à NLP, especialmente a Ciência da Informação, por fornecer os principais conceitos de tratamento de dados, informações e conhecimento, é o cerne desse estudo. Este foi estruturado tendo como base uma Revisão Sistemática da Literatura, entendendo ser essa uma abordagem capaz de fundamentar noções iniciais e, ao mesmo tempo, consistentes para a análise da questão central que motivou esse trabalho. Dentre os resultados encontrados está a pouca utilização dos recursos de NLP pela Ciência da Informação.

Palavras-chave: Ciência da Informação, Processamento de Linguagem Natural (NLP), Rede Neural, Revisão Sistemática da Literatura.

Abstract:

One of the recent challenges of the approach called Big Data has been to extract relevant information from large amounts of unstructured data, such as texts written in different languages. The main approach to text and language analysis by computational means is given the name Natural Language Processing (NLP). Identifying how the areas of knowledge are using developments in their domains, with regard to NLP, especially Information Science, for providing the main concepts of data processing, information and knowledge, are at the heart of this study. To find answers to this relevant question, this study was structured based on a Systematic Literature Review, understanding that this is an approach capable of supporting initial notions that is at the same time consistent for the analysis of the central question that motivated this work. Among the results found is the little use of NLP resources by Information Science.

Keywords: Information Science, Natural Language Processing (NLP), Network Neural, Systematic Literature Review.

Notas de autor

1 Doutorando, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil; luanderfalcao@yahoo.com.br; ORCID: https://orcid.org/0000-0003-2417-6345

2 Doutorando; Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil; brenner.lopes@gmail.com; ORCID: https://orcid.org/0000-0002-5807-0437

3 Doutor; Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil; rsouzaufmg@gmail.com; ORCID: https://orcid.org/0000-0002-1895-3905

Abstract:

One of the recent challenges of the approach called Big Data has been to extract relevant information from large amounts of unstructured data, such as texts written in different languages. The main approach to text and language analysis by computational means is given the name Natural Language Processing (NLP). Identifying how the areas of knowledge are using developments in their domains, with regard to NLP, especially Information Science, for providing the main concepts of data processing, information and knowledge, are at the heart of this study. To find answers to this relevant question, this study was structured based on a Systematic Literature Review, understanding that this is an approach capable of supporting initial notions that is at the same time consistent for the analysis of the central question that motivated this work. Among the results found is the little use of NLP resources by Information Science.

Keywords: Information Science, Natural Language Processing (NLP), Network Neural, Systematic Literature Review.

1 Introdução

Os avanços ocorridos nos últimos anos no campo da Tecnologia da Informação e Comunicação (TIC) proporcionaram uma maior capacidade de processamento e armazenamento de dados e informações. A internet possui um papel preponderante nesse contexto, pois impulsionou uma maior geração de dados e informações. Segundo Jin *et al*. (2015, p. 60), "[...] o uso extensivo da Internet, Internet das Coisas, Computação em Nuvem e outras tecnologias emergentes de TI fez com que várias fontes de dados aumentassem a uma taxa sem precedentes.".

Mesmo com o aumento das redes e a capacidade crescente da computação, apenas uma fração dos dados é usada para obter insights (LAUSCH; SCHMIDT; TISCHENDORF, 2015). Como resultado, "[...] a extração de informações e a produção de conhecimentos que poderiam ser úteis para a sociedade, não acontecem com a agilidade e a eficácia necessárias.". (ISOTANI; BITTENCOURT, 2016, p. 16).

Diante desse cenário surge o conceito de Data Mining. Enquanto estatística, matemática, computação e outas áreas do conhecimento focam em suas respectivas competências, *Data Mining* faz a confluência de diferentes e múltiplas disciplinas no processo de mineração de dados (LAUSCH; SCHMIDT; TISCHENDORF, 2015) com o objetivo de encontra valor ou insights nesses dados.

Com a perspectiva de encontrar insights em grandes massas de dados, o *Data Mining* encontrou no *Big Data* um pragmatismo focado em propósitos e entregas tangíveis. Apesar de haver várias definições para *Big Data*, ele pode ser entendido basicamente:

> [...] como uma abordagem holística para gerenciar, processar e analisar 5 V's (volume, variedade, velocidade, veracidade e valor), a fim de criar *insights* acionáveis para a entrega sustentada de valor, medir o desempenho e estabelecer vantagens competitivas. (FOSSO WAMBA et al., 2015, p. 235).

Esse conceito explica em parte a adoção em larga escala do *Big Data* pelas empresas, pois passa a ser o direcionador da estratégia, da tática e da operação das "[...] áreas da empresa cuja gerência pretenda conservar ou ampliar uma vantagem competitiva." (MILLER, 2002, p. 36).

Uma das iniciativas internacionais de *Big Data* tem sido preparar a próxima geração de cientistas e engenheiros com habilidades para extrair e analisar informações de textos em qualquer linguagem (JIN et al., 2015). À análise de texto e linguagem por meio computacional é dado o nome de Processamento de Linguagem Natural (*Natural Language Processing* – NLP). O NLP é um tópico da Ciência da Computação.

Os recentes avanços em NLP por meio de redes neurais tem proporcionado várias pesquisas com tarefas próprias de NLP, como classificação de texto, análise semântica, extração de informação e outros. Essas tarefas de NLP são próprias das atividades da Ciência da Informação (CI), pois entre as várias atividades da CI, pode-se citar a organização e recuperação da informação. Apesar da Ciência da Informação não possuir uma definição única (HJØRLAND, 2018), Zins (2007) buscou agrupar as principais visões até então como "[...] o estudo de fenômenos de sistemas, seus subsistemas e processos de informação e suas inter-relações

em diferentes contextos ambientais." (ZINS, 2007, p. 337). Essa definição traz de forma subjacente a ideia da interdisciplinaridade, apontada por vários autores como uma característica da Ciência da Informação (SARACEVIC, 1996; 1999), e muito presente na Ciência da Computação.

Nesse sentido, a Ciência da Computação, por meio de subáreas como *Machine Learning* e Inteligência Artificial (*Artificial Intelligence* em inglês) aplicadas em NLP, fornece um ferramental inovador e avançado para a Ciência da Informação executar tarefas como criação, indexação, armazenamento, recuperação e disseminação de informações. Logo, pode-se presumir que há um aumento do recurso computacional, principalmente em NLP, que pode ser apropriado pela Ciência da Informação para automatização e ampliação de suas funções. Entretanto, há múltiplas tarefas de NLP com objetivos distintos. Diante disso surge a seguinte pergunta: para quais objetivos as pesquisas envolvendo NLP podem ser apropriadas pela Ciência da Informação?

O foco dessa pesquisa é compreender como esses recursos computacionais de redes neurais, aplicados em NLP, vêm sendo desenvolvidos e em quais áreas eles têm sido mais aplicados. Assim será possível apontar o quanto a Ciência da Informação tem se apropriado desses recursos computacionais para suas atividades. Para isso foi conduzida uma pesquisa de Revisão Sistemática da Literatura, em artigos que aplicaram redes neurais em tarefas de NLP.

Para responder à pergunta dessa pesquisa, esse artigo está organizado da seguinte forma: introdução (seção 1); seção 2, aborda a Revisão da Literatura a respeito do Processamento de Linguagem Natural (NLP), Ciência da Informação e Redes Neurais; seção 3, na qual se discute a metodologia da pesquisa; seção 4, exposição e discussão dos resultados; seção 5, conclusão; e seção 6, bibliografia.

2 Revisão da literatura

2.1 A inter-relação da Ciência da Informação com a Ciência de Dados

A Ciência da Informação (CI) é uma disciplina jovem, pois remonta aos anos 1960, mas enraizada nas profundezas do tempo, datando do surgimento da linguagem escrita, da impressão de livros e das bibliotecas (RODIONOV; TSVETKOVA, 2015). Por ser nova, os limites e definições acerca da Ciência da Informação não são claros, gerando paralelos conceituais com outras áreas do conhecimento.

Diante dessa diversidade conceitual, Hjørland (2018, p. 232), descreveu "[...] a história da biblioteca e da ciência da informação (LIS - *Library and Information Science*), desde suas raízes na ciência da biblioteca, ciência da informação e documentação.". Para isso, ele separou o termo LIS em Biblioteconomia (LS - *Library Science*), Ciência da Informação (IS - *Information Science*) e Documentação, e buscou os principais conceitos que fundamentam cada termo, e consequentemente LIS.

A partir de uma visão mais pragmática, Chang (2018), investigou as contribuições de pesquisadores externos à área de conhecimento de Bibliotecas e Ciências da Informação (LIS), que não eram afiliados a instituições relacionadas à LIS, mas publicaram seus resultados de pesquisa em revistas da área. Dentre os resultados encontrados foi observado uma tendência crescente nos graus de interdisciplinaridade da LS e IS. Também observou que pesquisadores com formação em Ciência da Computação foram os contribuintes "não-LIS" mais prevalentes no campo de Ciência da Informação e preferiram publicar individualmente.

Logo, ambos os autores, Hjørland (2018) e Chang (2018) concluem que a Ciência da Informação é multidisciplinar, sendo esse um elo com a Ciência da Computação. Outro elo é a subdivisão chamada Recuperação da Informação (HJØRLAND, 2018).

Recuperação da Informação está relacionada ao design e avaliação de sistemas que ajudam as pessoas a acessar grandes coleções e a encontrar itens de interesse nessas coleções (FURNER, 2015). Porém, a recuperação de informações pode ser afetada pela representação e organização das informações

(WEISSENBERGER, 2015), gerando ilhas isoladas de informação. Para evitar a criação dessas ilhas informacionais, a classificação da informação precisa ser objetiva, baseada em ontologia, e subjetiva, baseada em epistemologia (LI; WANG, 2019). O aperfeiçoamento da classificação e recuperação da informação tem ocorrido, dentre outros fatores, principalmente devido a universalização da Inteligência Artificial (IA), e às aplicações derivadas dela como, por exemplo, as técnicas analíticas no campo do *Machine Learning* (ML).

Esses avanços no campo da IA e seus subcampos, como *Machine Learning* e o *Deep Learning*, ocorrem dentro do conceito de Ciência de Dados, como uma "[...] nova série de tecnologias, processos e sistemas para extrair valor e fazer descobertas." (WANG, 2018). Dentre as várias definições de Ciência de Dados, o entendimento proposto por Dhar (2013) merece destaque quando define essa abordagem como o estudo sistemático da organização, propriedades e análise de dados e seu papel na inferência. Segundo o Data Science Association (2020, documento online), Ciência de Dados é o "[...] estudo científico da criação, validação e transformação de dados para criar significado.". A Ciência de Dados é um campo emergente e interdisciplinar porque envolve estatística, mineração de dados, aprendizado de máquina e análise de dados, com o objetivo de produzir insights analíticos e estabelecer modelos de previsão (WANG, 2018).

Nesse ponto, a interdisciplinaridade tem um elo entre a Ciência de Dados e a Ciência da Informação e há um conjunto coerente de premissas meta-teóricas que se aplicam a estas ciências (WANG, 2018). Ao pensar na missão de ambas, a missão da Ciência de Dados é transformar dados brutos e confusos em conhecimento acionável (STANTON, 2012), enquanto a missão da Ciência da Informação é tornar mais acessível uma reserva desconcertante de conhecimento (BUSH, 1945).

Em termos de atividades, os cientistas de dados coletam, transformam, analisam, visualizam e fazem a curadoria de dados. Os cientistas da informação conduzem principalmente pesquisas sobre geração, coleta, organização, interpretação, armazenamento, recuperação, disseminação, transformação e uso de informações. Ambos se concentram na extração de valor e insight dos dados (WANG, 2018).

2.2 Processamento de Linguagem Natural (Natural Language Processing -NLP)

O Processamento de Linguagem Natural (NLP), enquanto disciplina, pode ser entendido como um elo entre Ciência da Computação e Ciência da Informação. Há 40 anos, aproximadamente, os computadores foram dotados com as capacidades básicas de entendimento inicial da linguagem natural. Com o crescimento da capacidade dos sistemas de computadores de processar texto em linguagem humana, também houve o crescimento do volume de texto legível por computador. Esses esforços foram originalmente chamados de "entendimento da linguagem natural", que agora é mais frequentemente chamado de processamento de linguagem natural (NLP) (CROWSTON; ALLEN; HECKMAN, 2012; MARTINEZ, 2010).

O NLP, também conhecido como linguística computacional, pertence ao campo da Ciência da Computação e da Linguística como um subcampo da Inteligência Artificial (IA) (MARTINEZ, 2010; ZEROUAL; LAKHOUAJA, 2018). Se preocupa com a adoção de máquinas para analisar a linguagem natural para um determinado objetivo (FALESSI; CANTONE; CANFORA, 2013), e visa aprender, entender, reconhecer e produzir conteúdo da linguagem humana (ZEROUAL; LAKHOUAJA, 2018). As tecnologias de NLP permitem extrair automaticamente informações e conhecimentos legíveis pela máquina a partir de dados de linguagem natural não estruturados (NESI; PANTALEO; SANESI, 2015).

O NLP fornece uma abordagem automatizada para medir a semelhança entre os pares de requisitos. A similaridade é então adotada como critério para classificar/filtrar os pares de requisitos de acordo com a probabilidade de serem equivalentes (FALESSI; CANTONE; CANFORA, 2013). O NLP possibilita várias abordagens, pois existe uma infinidade de técnicas de NLP, com várias destas apresentando desempenho diferente de acordo com o contexto. Por exemplo, alguns de seus efeitos foram observados em tradução automática, desambiguação de sentido de palavra, resumo, anotação sintática, reconhecimento de entidade nomeada, compreensão e classificação supervisionada de documentos de texto, extração de conteúdo,

design de ferramentas de recomendação e sistemas de suporte a decisões, expansão de consultas, estruturas de perguntas e respostas e análise de sentimentos (NESI; PANTALEO; SANESI, 2015; ZEROUAL; LAKHOUAJA, 2018).

Dado o aumento do fluxo de dados textuais e da diversidade de técnicas, houve uma mudança de paradigma na arquitetura da computação e no processamento de dados em larga escala. Esses desafios exigem processamento intensivo e em larga escala (AGERRI et al., 2015). Uma das respostas dada pela Ciência da Computação foi a Rede Neural.

2.3 Rede Neural

O que é uma Rede Neural? É um método de inspiração biológica para criar programas de computador capazes de aprender e encontrar conexões de dados independentemente (SKALSKI, 2018). É uma técnica para criar um programa de computador, também chamada de arquitetura, que aprende com os dados, baseado na estrutura e funções de redes neurais biológicas. Primeiro, uma coleção de "neurônios" de *software* é criada e conectada, permitindo que eles enviem mensagens uns para os outros. Em seguida, solicita-se à rede que resolva um problema. Ela tenta resolver repetidamente, sempre fortalecendo as conexões que levam ao sucesso e diminuindo as que levam ao fracasso (KOCEJKO, 2018; MITRA, 2018).

Redes Neurais, ou Redes Neurais Artificiais, representam uma tecnologia baseada em muitas disciplinas, como: neurociências, matemática, estatística, física, ciência da computação e engenharia. As Redes Neurais encontram aplicações em campos diversos como modelagem, análise de séries temporais, reconhecimento de padrões, processamento de sinais e controle. Essas aplicações são possíveis em virtude de uma propriedade importante: a capacidade de aprender com dados de entrada com ou sem um instrutor (HAYKIN, 1999).

Esse aprender automaticamente com os dados era conhecido desde 2006, mas apenas para abordagens mais tradicionais. Isso mudou em 2006 com o *Deep Neural Networks* (Redes Neurais Profundas). Essa técnica também é conhecida como *Deep Learning*. Tanto Redes Neurais Profundas quanto *Deep Learning* alcançaram um desempenho excepcional em muitos problemas importantes em computação, como reconhecimento de fala e processamento de linguagem natural. Logo esses algoritmos se tornaram poderosos e populares, e muito do sucesso deles reside no design cuidadoso da arquitetura da Rede Neural (CULURCIELLO, 2017; GOODFELLOW; BENGIO; COURVILLE, 2016; NIELSEN, 2015).

Em Rede Neural há vários tipos de arquiteturas. Dentre as mais modernas destacam-se as Redes Neurais Convolucionais (CNNs, *Convolutional Neural Networks* -). As CNNs são um tipo especializado de Rede Neural para processamento de dados que possui uma topologia conhecida como grade. Usam convolução no lugar da multiplicação geral da matriz em pelo menos uma de suas camadas. As arquiteturas de rede baseadas nas CNNs dominam o campo da visão computacional a tal ponto que hoje em dia quase ninguém desenvolveria um aplicativo comercial ou participaria de uma competição relacionada ao reconhecimento de imagens, detecção de objetos ou segmentação semântica, sem basear sua abordagem em CNNs (GOODFELLOW; BENGIO; COURVILLE, 2016; ZHANG et al., 2020).

Redes Neurais Recorrentes (RNNs, *Recurrent Neural Networks*) são uma família de Redes Neurais para o processamento de dados sequenciais. Enquanto as Redes Neurais Convolucionais podem processar informações espaciais com eficiência, as Redes Neurais Recorrentes são projetadas para lidar melhor com as informações sequenciais. Enquanto as Redes Convolucionais podem ser facilmente dimensionadas para imagens com grande largura e altura, ou então processar imagens de tamanho variável, as Redes Recorrentes podem ser dimensionadas para sequências muito mais longas do que seria esperado para redes sem especialização, baseadas em sequências. A maioria das Redes Recorrentes também podem processar sequências de comprimento variável. Essas redes introduzem variáveis para armazenar informações passadas e, em seguida, determinam as saídas atuais, juntamente às entradas atuais (GOODFELLOW; BENGIO;

COURVILLE, 2016; ZHANG et al., 2020). Entre as RNNs modernas estão *Gated Recurrent Units* (GRU), *Long Short Term Memory* (LSTM), e *Encoder-Decoder Architecture*.

As arquiteturas CNN e RNN possuem modelos de treinamento para a representação de texto para diferentes tarefas de NLP, como *Word Embedding* (word2vec), *The Continuous Bag of Words* (CBOW), *Word Embedding with Global Vectors* (GloVe), *Subword Embedding – fastText, Bidirectional Encoder Representations from Transformers* (BERT), *Context Vectors* (CoVe), *Embeddings from Language Models* (ELMo) (ZHANG *et al.*, 2020).

3 Metodologia

Segundo Lakatos e Marconi (2008, p. 83):

> o método é o conjunto das atividades sistemáticas e racionais que, com maior segurança e economia, permite alcançar o objetivo – conhecimentos válidos e verdadeiros – traçando o caminho a ser seguido, detectando erros e auxiliando as decisões do cientista.

Com base nessa definição, a metodologia que suportou essa pesquisa, quanto à abordagem, pode ser caracterizada como uma pesquisa qualitativa (CRESWELL, 2014) e, quanto ao objetivo como pesquisa exploratória.

Para Gil (2008, p. 27), "[...] as pesquisas exploratórias têm como principal finalidade desenvolver, esclarecer e modificar conceitos e ideias, tendo em vista a formulação de problemas mais precisos ou hipóteses pesquisáveis para estudos posteriores.". De acordo com Wazlawick (2014, p. 22) "[...] na pesquisa exploratória, o autor vai examinar um conjunto de fenômenos, buscando anomalias que não sejam ainda conhecidas e que possam ser, então, a base para uma pesquisa mais elaborada.".

Quanto ao procedimento técnico, a pesquisa pode ser classificada em bibliográfica. A pesquisa bibliográfica é desenvolvida a partir de material já elaborado (GIL, 2008) e implica no estudo de artigos, teses, livros e outras publicações usualmente disponibilizadas por editoras e indexadas (WAZLAWICK, 2014). Segundo Gil (2008, p. 50), "[...] a principal vantagem da pesquisa bibliográfica reside no fato de permitir ao investigador a cobertura de uma gama de fenômenos muito mais ampla do que aquela que poderia pesquisar diretamente.".

A metodologia escolhida como a mais apropriada para responder à pergunta desse estudo foi a Revisão Sistemática da Literatura, ou Revisão Sistemática. Segundo Sampaio e Mancini (2007, p. 84), "[...] uma revisão sistemática, assim como outros tipos de estudo de revisão, é uma forma de pesquisa que utiliza como fonte de dados a literatura sobre determinado tema.". De acordo com Faria (2019, p. 20):

> caracteriza-se por empregar uma metodologia de pesquisa com rigor científico e de grande transparência, cujo objetivo é minimizar o enviesamento da literatura, na medida que é feita uma pesquisa exaustiva dos textos publicados sobre o tema em questão.

Em termos de características, "[...] as revisões sistemáticas são consideradas estudos secundários, que têm nos estudos primários sua fonte de dados, sendo estudos primários os artigos científicos, que relatam os resultados de pesquisa em primeira mão." (GALVÃO; PEREIRA, 2014, p. 183). De acordo com Faria (2019, p. 21), deve-se "[...] estruturar todos os procedimentos de forma a garantir a qualidade das fontes, logo pela definição de uma equação de pesquisa, de critérios de inclusão e exclusão e de todos os critérios que se julguem convenientes para o caso.". Segundo Sampaio e Mancini (2007, p. 84):

> [...] esse tipo de investigação disponibiliza um resumo das evidências relacionadas a uma estratégia de intervenção específica, mediante a aplicação de métodos explícitos e sistematizados de busca, apreciação crítica e síntese da informação selecionada.

A Revisão Sistemática da Literatura possui várias formas, por isso foi adotada a Revisão Sistemática e, dentro desta, a meta-análise (BOTELHO; CUNHA; MACEDO, 2011). Segundo Sampaio e Mancini (2007, p. 84), meta-análise:

> [...] é a análise da análise, ou seja, é um estudo de revisão da literatura em que os resultados de vários estudos independentes são combinados e sintetizados por meio de procedimentos estatísticos, de modo a produzir uma única estimativa ou índice que caracterize o efeito de (uma) determinada intervenção.

Para Botelho, Cunha e Macedo (2011, p. 127), "[...] na meta-análise, cada estudo é sintetizado, codificado e inserido num banco de dados quantitativo.".

As etapas para a condução da revisão sistemática seguiram os passos apontados por Galvão e Pereira (2014). As etapas 1 (elaboração da pergunta de pesquisa) e 2 (busca na literatura) já constam no estudo. A etapa 3 (seleção dos artigos) foi realizada no Portal de Periódicos CAPES, no dia 16/06/2020, com a seguinte query de consulta, representada na Fórmula 1:

"nlp" AND "transformers" AND ("LSTM" OR "GRU" OR "word embedding" OR "Word2vec" OR "Doc2vec" OR "CBOW" OR "Char2vec" OR "Glove" OR "fastText" OR "BERT" OR "CoVe" OR "Elmo")

FÓRMULA 1
Query de Consulta
Fonte: Elaborado pelos autores (2020).

Os termos foram retirados da Revisão da Literatura, dos textos que subsidiaram a Seção 2. O Portal de Periódicos CAPES foi utilizado como referência pois permite acesso a uma gama de periódicos, de diversas áreas do conhecimento, que podem conter de alguma forma o tema de interesse. O escopo de busca foi delimitado, permitindo apenas a busca por artigos científicos. No quesito delimitação temporal, foram selecionados os artigos do ano de 2010 até 2020, tendo em vista a contemporaneidade do tema.

Com esse procedimento especificado, a quarta etapa (extração dos dados) foi realizada. Foram encontrados 169 artigos, nos quais se buscou respostas para as seguintes questões: a) quais *transformers* foram usados?; b) quais aplicações NLP foram usadas?; c) qual o objetivo do artigo?; d) em qual(is) área(s) do conhecimento foi aplicada o artigo? Também foram analisados os resumos dos objetivos dos artigos. Para cada pergunta há a possibilidade de haver mais de uma resposta, exceto para "objetivo do artigo" e "resumo do objetivo". O objetivo do artigo foi inserido para fornecer a base para o resumo. Foi criado um banco de dados em Excel com os seguintes campos: código do artigo; título do artigo; autores; jornal; ano; *transformers*; *task* NLP; objetivo; resumo; e áreas. Foi realizada a leitura e análise de cada um dos artigos selecionados e os seus dados foram utilizados para alimentar a base de dados.

Na quinta etapa, avaliação da qualidade metodológica, 54 artigos foram retirados por não estarem relacionados à temática NLP e Rede Neural ou por estarem duplicados. A sexta etapa, síntese dos dados/meta-análise, ocorreu por meio de estatística descritiva, na qual os termos que aparecem nas variáveis *Transformers*, *Task* NLP, Resumo e Áreas foram contados e agrupados. Os termos nessas variáveis foram padronizados para a realização da estatística descritiva. As demais etapas, 7 (avaliação da qualidade das evidências) e 8 (redação e publicação dos resultados), constam no capítulo a seguir: análise dos resultados.

4 Análise dos resultados

Os 115 artigos selecionados datam do ano de 2018 a 2020, sendo que 98% são dos anos de 2019 e 2020. Isso evidência o quanto essa temática é atual. Essa concentração nessas datas é justificada pelo avanço das abordagens em NLP e, principalmente, pelo Transformer BERT, lançado em 2018 e rapidamente absorvido

pela comunidade científica. Segundo Devlin et al. (2018, p. 1), "[...] BERT foi projetado para pré-treinar representações bidirecionais profundas de texto não rotulado, condicionando conjuntamente os contextos esquerdo e direito em todas as camadas.". Por isso BERT possui uma capacidade maior de processamento de tarefas de NLP.

Foram citados 45 transformers distintos de Rede Neural empregados em NLP. Sendo que um mesmo artigo pode ter usado mais de um transformer em sua análise. Isso gerou 227 citações de Transfomers. Variações dos transformers, como Bi-LSTM ou BioBERT foram padronizadas para a "raiz"do transformer. Com isso, todas as citações de transformers foram padronizadas para a sua "raiz".

O transformer BERT foi o mais citado, mesmo sendo o transformer mais novo, justamente por ser muito versátil em NLP. BERT foi construído originalmente com bases de testes em inglês, mas tem funcionalidades em outros idiomas. Muitos artigos utilizaram BERT em tradução para idiomas específicos e obtiveram um resultado melhor do que utilizando as bases internas de treinamento do BERT. Os autores dos artigos analisados apontaram uma preferência por usarem BERT, LSTM, GloVe e ELMo. Esses transfomers correspondem por 60% das citações, e possuem funcionalidades próprias para NLP. Além desses, foram citados os seguintes transformers e arquiteturas: CNN, GPT, GRU, RNN, BART, ConvNet, ERNIE, Flair, BLEU, Byte Pair Embeddings, CapsNet, Connectionist Text Proposal Network (CTPN), Context2vec, Cui2Vec, Dict2vec, Embeddings of Semantic Predications (ESP), Graph2Graph, GTrXL, KT-NET, MINILM, ngram2vec, OpenAI transformer, Reasoner, ROUGE-L, SAO2Vec, Sent2Vec, Seq2Seq, SeqVec, SGNS, SOFSAT, SRU, Tree Transformer, TrXL, T-XL, USE e XLM. Esses resultados estão apresentados na Tabela 1:

TABELA 1
Quantidade de Citação por Transformer

Transformers	Número de Citações	Part. %	Part. % Acumulada
BERT	72	31,72	31,72
LSTM	43	18,94	50,66
GloVe	11	4,85	55,51
ELMo	10	4,41	59,91
word embedding	9	3,96	63,88
Word2Vec	9	3,96	67,84
FastText	8	3,52	71,37
XLNet	8	3,52	74,89
Bag-of-words (BoW)	7	3,08	77,97
Others	50	22,03	100,00
Total	**227**	**100,00**	

Elaborado pelos autores (2020).

Nessa pesquisa foram identificadas cerca de 60 tarefas distintas de NLP, com 188 citações. Cerca de 10 tarefas correspondem por 60% das citações de Tarefas de NLP empregadas nos artigos analisados. Em vários artigos mais de uma Tarefa de NLP foi utilizada. A tarefa mais citada é *Named Entity Recognition* (NER), com 26 citações, correspondente a 14% do total de citações. NER é uma tarefa de extração de informações que localiza e classifica rótulos em texto, por categorias predefinidas, como nomes de pessoas, organizações, localizações, códigos médicos, expressões de tempo, quantidades, valores monetários, porcentagens, etc. Essa tarefa possui variações, como *Medical Entity Recognition* (MER) e *Clinical Named Entity Recognition* (CNER). Artigos citam uma melhora considerável quando se utiliza transformer em idiomas específicos ou em domínios específicos, como da área médica. Esse achado justifica o desenvolvimento de tarefas específicas como transformers específicos tanto para idioma quanto para tarefas.

Dentre as Tarefas de NLP, é possível identificar aquelas com forte correlação com a Ciência da Informação, como *Text Classification, Information Extraction* (IE), *Semantic Textual Similarity* (STS), *Relation Extraction* (RE), *Document Classification (DC), Reinforcement Learning* (RL), *Text Representation, Transfer Learning* (TL), *Extractive Summaries, Relation Classification* (RC), *Semantic Role Labeling* (SRL) e *Subject–Action–Object* (SAO). Essas tarefas totalizam 28% das citações. Se NER for considerado um tipo de classificação, esse total sobe para 42%. Esse tipo de resultado aponta o quanto os conceitos e ferramentas, próprias da Ciência da Informação são utilizadas em outras áreas do conhecimento para otimização de funções.

O grupo de "Demais" agrupa uma série de tarefas específicas, como por exemplo, *Reinforcement Learning* (RL), *Represent Protein Sequences, Response Generation, Search Query Understanding, Synonym Discovery, Relation Extraction* (RE), *Chunking, Document Classification* (DC), *Machine Reading, Comprehension* (MRC), *Medical Entity Recognition* (MER), *Natural Language Understanding* (NLU), *Text Representation, Transfer Learning* (TL) e outras 41 tarefas distintas. Esses resultados estão apresentados na Tabela 2:

TABELA 2
Quantidade de Citações por Tarefas de NLP

***Task* NLP**	**Número de Citações**	**Part. %**	**Part. % Acumulada**
Named Entity Recognition (NER)	26	13,83	13,83
Text Classification	17	9,04	22,87
Question Answering (QA)	14	7,45	30,32
Sentiment Analysis (SA)	14	7,45	37,77
Information Extraction (IE)	10	5,32	43,09
Part Of Speech (POS)	7	3,72	46,81
Sentiment Classification (SC)	7	3,72	50,53
General Language Understanding Evaluation (GLUE)	6	3,19	53,72
Semantic Textual Similarity (STS)	6	3,19	56,91
Machine Translation (MT)	5	2,66	59,57
Others	76	40,43	100,00
Total	**188**	**100,00**	

Fonte: Elaborado pelos autores (2020).

Foram identificadas 21 áreas distintas do conhecimento em que os *transformers* foram aplicados, com 132 citações ao todo. Como alguns artigos citavam áreas correlatas, como *clinical* e *medical*, ambas as áreas foram utilizadas, justamente para não perder a abrangência do assunto do qual o artigo trata. A área *Computation and Language* teve 67 citações, correspondendo por 50,76% do total de citações. Como principal usuária de NLP, pode-se citar a área médica, nesse trabalho representada por medical, clinical e biomedical, com 29,55% das citações.

Entre as aplicações na área médica estão tarefas como registros médicos eletrônicos, para atividade clínica, detecção de atributos de conceitos médicos no texto clínico, encontrar respostas para perguntas de textos biomédicos, construção de bases de conhecimento para associações de microbiomas e doenças humanas, extrair informações clínicas abrangentes para câncer de mama, descoberta de sinônimos médicos e outros. Esse resultado aponta que a área médica tem buscado utilizar mais os recursos de NLP para otimizar suas atividades operacionais e de pesquisa.

Foram identificados dois artigos focados em Ciência da Informação. Dentre *Others* foi identificado um artigo na área de risco de crédito, no qual os autores utilizarem NLP para ponderar empréstimos, e outro na área de *Maintenance*, voltado para manutenção de geradores. Os dados sobre a quantidade de citações por áreas do conhecimento estão descritos na Tabela 3.

TABELA 3
Quantidade de Citações por Áreas do Conhecimento

Áreas do Conhecimento	Número de Citações	Part. %	Part. % Acumulada
Computation and Language	67	50,76	50,76
Medical	18	13,64	64,39
Biomedical	13	9,85	74,24
Clinical	8	6,06	80,30
Social Media	5	3,79	84,09
Radiology	3	2,27	86,36
Biology	2	1,52	87,88
Educational	2	1,52	89,39
Information Science	2	1,52	90,91
Others	12	9,09	100,00
Total	**132**	**100,00**	

Elaborado pelos autores (2020).

No que tange aos objetivos, foram identificados 35 objetivos distintos, totalizando 115 citações de objetivos. O principal objetivo foi "*propose by new/improvement architecture*" (proposta de uma nova arquitetura ou melhoria), com 74 citações, correspondente a 64% do total. Esse objetivo está relacionado à proposição de uma nova arquitetura ou se propõe a melhorar um *transformer*. Exemplos desse caso: exBERT, BioBERT, M-BERT, Q8BERT, X-BERT, TRANS-BLSTM, entre outros. A maioria tem como raiz BERT, por ainda estar em fase de exploração. Esse resultado aponta a busca por *transformers* mais específicos em tarefas e idiomas, eficientes e que possibilitem uma resposta de predição melhor.

Ao cruzar "áreas do conhecimento" com "tarefas" foram identificados 15 artigos que utilizaram NER em "medicina, clínica e biomédica". A tarefa *text classification* foi a segunda tarefa mais citada (9%). Ao todo, 16 artigos utilizaram essa tarefa, sendo nove na área de conhecimento Linguagem e Computação. Os outros sete artigos se distribuíram em: *social media* (três) e *medical* (os outros quatro).

Esse tipo de resultado corrobora o fato dos conceitos, tarefas e nomenclaturas próprias da Ciência da Informação estarem sendo plenamente utilizados, principalmente em Ciência da Computação, quando aplicados em NLP. Entretanto, não se observou o uso de recursos de NLP com a mesma intensidade pela Ciência da Informação, mesmo com suas diversas tarefas.

5 Conclusões

O objetivo do artigo foi identificar o quanto as áreas do conhecimento estão se apropriando dos recursos de NLP, uma vez que, o custo de captação, processamento e armazenagem de dados tornou uma série de atividades mais acessíveis. Com tarefas de NLP é possível analisar inúmeros blocos de textos desestruturados e assim obter *insights*. Esse tem sido, por exemplo, um dos objetivos da área médica quando utiliza ferramentas de NLP em registros médicos.

Ao considerar as possibilidades, qualquer área do conhecimento deveria utilizar esse tipo de insumo para analisar textos, que na maioria das vezes são gravados para fins de auditoria e nunca mais visitados. Entretanto, o que foi constatado é um uso muito intenso das ferramentas de NLP pela área Computação e Linguagem, principalmente para melhorar a performance dos *transformers*, apresentar uma melhoria em um *transformer* ou até mesmo lançar um *transformer*. Um outro fato recorrente nos artigos é o teste entre *transformers*, para identificar qual proporciona um resultado de predição melhor por tarefa de NLP.

Os *transformers* mais citados foram BERT, mesmo tendo sido lançado em 2018, porém com vários recursos, e com a possibilidade de ser adaptado para obter um desempenho melhor por tarefa ou domínio, ou

tarefa por domínio. O LSTM foi o segundo *transformer* mais citado, também sendo muito versátil em tarefas de NLP, e possui recursos de otimização, citado em diversos artigos.

Ao considerar especificamente a Ciência da Informação, foram identificados apenas dois artigos nessa área, um para a construção de uma base de conhecimento e outro de extração de informação. Enquanto a Ciência da Computação, utiliza os termos da Ciência da Informação para desenvolver e melhorar os recursos computacionais, não se observou o uso dessas melhorias computacionais na Ciência da Informação.

Ao fazer uma revisão da literatura a partir de uma base de artigos abrangente como a CAPES, que indexa as principais revistas de diversas áreas, não foi possível identificar a aplicação de tarefas de NLP em Ciência da Informação. Isso indica que as publicações de Ciência da Informação têm utilizado pouco as tarefas de NLP em suas pesquisas. Conforme aponta Chang (2018), autores da Ciência da Computação tem publicado com frequência em periódicos de Ciência da Informação.

Logo, pode-se deduzir que há uma gama de possibilidades de pesquisa nas quais se utilize os conceitos de Ciência da Informação e tarefas de NLP. Dentre as alternativas está a classificação de textos, a classificação de documentos, a extração de partes da informação (NER), análise semântica, partes de discurso e outros, que podem otimizar a recuperação da informação.

Pelos resultados da pesquisa, é possível concluir que um dos focos da Ciência da Computação é melhorar a performance dos *transformers* e testá-los, para identificar os ajustes que possibilitem uma predição cada vez melhor. Diante disso, surge a possibilidade da Ciência da Informação aplicar as melhorias dos *transformers* de NLP em suas atividades, como representação do conhecimento, armazenamento e recuperação da informação, representação e organização das informações e classificação da informação. A aplicação dos recursos de NLP nessas atividades de CI as aperfeiçoaria, evitando a geração de ilhas isoladas de informação. O resultado seria a obtenção de *insights* em bases de conhecimento antes intocáveis, como a melhoria da gestão do conhecimento.

Referências

AGERRI, Rodrigo *et al*. Big data for Natural Language Processing: a streaming approach. **Knowledge-Based Systems**, Amsterdam, v. 79, p. 36-42, 2015. DOI: 10.1016/j.knosys.2014.11.007.

BOTELHO, Louise Lira Roedel; CUNHA, Cristiano Castro de Almeida; MACEDO, Marcelo. O método da revisão integrativa nos estudos organizacionais. **Gestão e Sociedade**, Belo Horizonte, v. 5, n. 11, p. 121-136, 2011. DOI: 10.21171/ges.v5i11.1220.

BUSH, Vannevar. As we may think. **Atlantic Monthly**, Boston, v. 176, n. 1, p. 101-108, 1945.

CHANG, Yu-Wei. Examining interdisciplinarity of library and information science (LIS) based on LIS articles contributed by non-LIS authors. **Scientometrics**, Dordrecht, v. 116, n. 3, p. 1589-1613, 2018. DOI: 10.1007/s11192-018-2822-7.

CRESWELL, John W. **Research design:** qualitative, quantitative, and mixed methods approaches. 4. ed. Los Angeles: Sage Publications, 2014.

CROWSTON, Kevin; ALLEN, Eileen E.; HECKMAN, Robert. Using natural language processing technology for qualitative data analysis. **International Journal of Social Research Methodology**, London, v. 15, n. 6, p. 523-543, 2012.

CULURCIELLO, Eugenio. Neural Network Architectures. **Towards data science**, *[S.l.]*, 23 Mar., 2017.

DATA SCIENCE ASSOCIATION. **About data science**. [S.l.]: Data Science Association, 2020.

DEVLIN, Jacob *et al.* BERT: pre-training of deep bidirectional transformers for language Understanding. **Computation and Language,** New York, 2018.

DHAR, Vasant. Data science and prediction. **Communications of the ACM**, New York, v. 56, n. 12, p. 64-73, 2013.

FALESSI, Davide; CANTONE, Giovanni; CANFORA, Gerardo. Empirical principles and an industrial case study in retrieving equivalent requirements via natural language processing techniques. **IEEE Transactions on Software Engineering**, Washington, D.C., v. 39, n. 1, p. 18-44, 2013.

FARIA, Paulo M. **Revisão sistemática da literatura:** contributo para um novo paradigma investigativo. Champaign: CG Publisher, 2019.

FOSSO WAMBA, Samuel; AKTER, Shahriar; EDWARDS, Andrew; CHOPIN, Geoffrey; GNANZOU, Denis. How "big data" can make big impact: findings from a systematic review and a longitudinal case study. **International Journal of Production Economics**, Amsterdam, v. 165, p. 234-246, 2015.

FURNER, Jonathan. Information Science is neither. **Library Trends**, Baltimore, v. 63, n. 3, p. 362-377, 2015.

GALVÃO, Taís Freire; PEREIRA, Mauricio Gomes. Revisões sistemáticas da literatura: passos para sua elaboração. **Epidemiologia e Serviços de Saúde**, Brasília, DF, v. 23, n. 1, p. 183–184, 2014.

GIL, Antônio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. Cambridge: MIT Press, 2016.

HAYKIN, Simon O. **Neural networks:** a comprehensive foundation. 2. ed. Ontario.

HJØRLAND, Birger. Library and Information Science (LIS), Part 1. **Knowledge Organization**, Kent, v. 45, n. 3, p. 232-254, 2018.

ISOTANI, Seiji; BITTENCOURT, Ig Ibert. **Dados abertos**. São Paulo: Novatec, 2015.

JIN, Xiaolong *et al*. Significance and Challenges of Big Data Research. **Big Data Research**, Amsterdam, v. 2, n. 2, p. 59-64, 2015.

KOCEJKO, Szymon. **Neural networks:** what are neural networks and how do they work? 2018.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Fundamentos de metodologia científica**. 5. ed. São Paulo: Atlas, 2008.

LAUSCH, Angela; SCHMIDT, Andreas; TISCHENDORF, Lutz. Data mining and linked open data: new perspectives for data analysis in environmental research. **Ecological Modelling**, Amsterdam, v. 295, p. 5-17, 2015.

LI, Fan; WANG, Zhisen. The study on the out-structure information classification based on observers. **Journal of Physics**: Conference Series, *[S. l.]*, v. 1168, n. 2, 2019.

MARTINEZ, Angel R. Natural language processing. **Wiley Interdisciplinary Reviews**: Computational Statistics, Hoboken, v. 2, n. 3, p. 352-357, 2010.

MILLER, Jerry P. **O milênio da inteligência competitiva**. Porto Alegre: Bookman, 2002.

MITRA, Manu. Neural processor in artificial intelligence advancement. **Journal of Autonomous Intelligence**, *[S. l.]*, v. 1, n. 1, p. 2, 2018.

NESI, Paolo; PANTALEO, Gianni; SANESI, Gianmarco. A hadoop based platform for natural language processing of web pages and documents. **Journal of Visual Languages and Computing**, Pittsburgh, v. 31, n. 2015, p. 130-138, 2015.

NIELSEN, Michael A. **Neural networks and deep learning**. *[S.l.]*: Determination Press, 2015.

RODIONOV, I. I.; TSVETKOVA, V. A. Information management in information science. **Scientific and Technical Information Processing**, *[S. l.]*, v. 42, n. 2, p. 73–77, 2015.

SAMPAIO, RF; MANCINI, MC. Estudos de revisão sistemática: um guia para síntese criteriosa da evidência científica. **Revista Brasileira de Fisioterapia**, São Carlos, v. 11, n. 1, p. 83–89, 2007.

SARACEVIC, Tefko. Ciência da informação: origem, evolução e relações. **Perspectiva em Ciência da Informação**, Belo Horizonte, v. 1, n. 1, p. 41–62, 1996.

SARACEVIC, Tefko. Information science. **Journal of the American Society for Information Science**, New York, v. 50, n. 12, p. 1051–1063, 1999.

SKALSKI, Piotr. Deep dive into math behind deep networks. **Towards data science**, 17 Aug. 2018.

STANTON, Jeffrey M. Data science: what's in it for the newlibrarian? **Syracuse University**, Syracuse, 16 Jul. 2012.

WANG, Lin. Twinning data science with information science in schools of library and information science. **Journal of Documentation**, West Yorkshire, v. 74, n. 6, p. 1243-1257, 2018.

WAZLAWICK, Raul Sidnei. **Metodologia de pesquisa para ciência da computação**. 2. ed. Rio de Janeiro: Grupo Gen, 2014.

WEISSENBERGER, Lynnsey. Toward a universal, meta-theoretical framework for music information classification and retrieval. **Journal of Documentation**, West Yorkshire, v. 71, n. 5, p. 917–937, 2015.

ZEROUAL, Imad; LAKHOUAJA, Abdelhak. Data science in light of natural language processing: an overview. **Procedia Computer Science**, *[S. l.]*, v. 127, p. 82–91, 2018.

ZHANG, Aston *et al.* **Dive into deep learning**. *[S. l]:* arXiv preprint, 2021

ZINS, Chaim. Conceptual approaches for defining data, information, and knowledge. **Journal of the American Society for Information Science and Technology**, *[S. l.]*, v. 58, n. 4, p. 479-493, 2007.

Información adicional

Como citar: FALCÃO, Luander; LOPES, Brenner; SOUZA, Renato Rocha. Absorção das tarefas de processamento de Linguagem Natural (NLP) pela Ciência da Informação (CI): uma revisão da literatura para tangibilização do uso de NLP pela CI. Em Questão, Porto Alegre, v. 28; n. 1, p. 13-34, 2022. DOI: http://dx.doi.org/10.19132/1808-5245281.13-34