EXTRAÇÃO DE CONHECIMENTO EM DISCURSOS POLÍTICOS: REVISÃO SISTEMÁTICA

Márcio Welter[1];
Lyvia Mendes Corrêa[2];
Alexandre Leopoldo Gonçalves[3]

Abstract: *Detecting underlying hidden dynamics in political attitudes and expressions, as well as similarities between them, is crucial for our society. This article contribute to the synthesis of the main scientific efforts in extracting knowledge from political discourses, through a systematic review. It was possible to contribute introducing the main techniques, strategies and impacts of the application of clustering and social network analysis from political speeches, as well as the study scenarios adopted by parliaments around the world.*
Keywords: Discovery Knowledge in Text; Clustering;Social Network Analysis; Policy.

Resumo: Detectar dinâmicas ocultas subjacentes em atitudes e expressões políticas, bem como semelhanças entre elas, é crucial para nossa sociedade. O propósito deste artigo é contribuir com a síntese dos principais esforços científicos na extração de conhecimento de discursos políticos, por meio de uma revisão sistemática. Foi possível contribuir apresentando quais são as principais técnicas, estratégias e os impactos da aplicação de agrupamentos e análise de redes a partir de discursos políticos, bem como os cenários de estudo adotados pelos parlamentos ao redor do mundo.
Palavras-chave: Descoberta de Conhecimento em Texto; Agrupamentos; Análise de Redes Sociais; Política.

1 INTRODUÇÃO

Uma quantidade crescente de dados de transcrições de discursos políticos está disponível e, em alguns parlamentos remontam há centenas de anos, criando possibilidades de diversas análises do conteúdo (Diermeier et al., 2012; Abercrombie & Batista-Navarro, 2020; Moreira, Vaz-de-Melo & Pappa, 2020).

[1]Mestre do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento – Universidade Federal de Santa Catarina (UFSC) Florianópolis – Brasil. ORCID: https://orcid.org/0000-0002-5442-1041.e-mail: contato@marciowelter.com.br
[2]Mestranda do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento – Universidade Federal de Santa Catarina (UFSC) Florianópolis – Brasil. ORCID: https://orcid.org/0000-0002-2260-2462.e-mail: lyviacorrea@gmail.com
[3]Prof. Dr. do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento – Universidade Federal de Santa Catarina (UFSC) Florianópolis – Brasil. ORCID: http://orcid.org/0000-0002-6583-2807.e-mail: alexandre.l.goncalves@gmail.com

Detectar dinâmicas ocultas subjacentes em atitudes e expressões políticas, bem como semelhanças entre elas, é crucial para nossa sociedade (Silva, 2020). Todas as percepções devem considerar os contornos pessoais, diferentes filtros, perspectivas de análise destacadas pela sensibilidade humana (Lorenzini, 2020). Contudo, tal tarefa desafiadora é difícil de alcançar sem métodos estatísticos avançados e ferramentas de mineração de dados.

O uso de dados dos discursos parlamentares, classificados por um modelo de tópico, pode servir para construir redes compostas por um conjunto de parlamentares que estão vinculados a um conjunto de tópicos, ou redes bimodais que constituem uma rede onde uma ligação entre parlamentares indica a existência de um interesse mútuo, fornecendo informações importantes sobre as atividades políticas. Por exemplo, como descobrir se a gama típica de interesses de um deputado está mudando ou não, bem como os padrões desse comportamento ao longo do tempo (Moreira, Vaz-de-Melo & Pappa, 2020).

Portanto, com base nos registros existentes de discursos políticos dos parlamentares e no processo de descoberta de conhecimento em texto, a presente pesquisa sustenta-se em avanços dos estudos científicos sobre as análises de temáticas e suas relações presentes em discursos políticos nos parlamentos.

2 REFERENCIAL TEÓRICO

O contexto principal acerca do tema deste trabalho passa pelas definições de Descoberta de Conhecimento em Texto e Análise de Redes Sociais.

2.1 DESCOBERTA DE CONHECIMENTO EM TEXTO

A Descoberta de Conhecimento em Texto (do inglês *Knowledge Discovery in Text* - KDT) é entendida como “um processo não trivial de identificação de padrões implícitos, a partir de dados textuais, válidos, potencialmente úteis e compreensíveis” (Gonçalves et al., 2018, p. 3). O termo “processo” indica que o KDT é constituído de várias etapas insterconectadas, permitindo múltiplas iterações. Chistol (2020, p. 209) complementa afirmando que se trata de um processo complexo, iterativo e interativo com um forte caráter interdisciplinar.

Entre as etapas do KDT encontra-se a Mineração de Texto (do inglês *Text Mining* - TM), que aplica métodos e técnicas, com o intuito de revelar padrões e ativos do conhecimento a partir de dados não estruturados na forma de textos. A TM utiliza uma combinação de técnicas de *data mining*, *machine learning*, processamento de linguagem natural (PLN), recuperação da informação (RI) e gerenciamento de conhecimento de grandes

bases textuais (Delen & Crossland, 2008; Yang et al., 2018; Gonçalves et al., 2018), visando promover subsídios para diferentes tarefas intensivas em conhecimento, tais como a categorização e o agrupamento de texto.

A tarefa de categorização de textos (*text categorization*) com aplicação direta ao domínio da gestão de documentos, sendo uma das aplicações mais comuns de PLN (Sarkar, Bali & Sharma, 2017). Por exemplo, dado um conjunto de categorias (assuntos, tópicos) e uma coleção de documentos de texto, esse processo objetiva encontrar o tópico correto (ou tópicos) para cada documento (Feldman & Sanger, 2007).

No contexto de TM, o agrupamento (*clustering*) pode ser utilizado em aplicações com foco na organização de documentos, quando estes não possuem um assunto associado. Trata-se da tarefa de localizar grupos de documentos semelhantes em uma coleção de documentos. A similaridade é calculada usando uma função de similaridade. O agrupamento de texto pode estar em diferentes níveis de granularidades, onde os agrupamentos podem ser documentos, parágrafos, sentenças ou termos. *Clustering* é uma das principais técnicas usadas para organizar documentos para melhorar a recuperação de informação e dar suporte à pesquisa (Allahyari et al., 2017).

2.2 ANÁLISE DE REDES SOCIAIS

A Análise de Redes Sociais (ARS) também pode contribuir para realizar análises na área da Política (Cassi et al., 2017; Bhattacharya, 2020; Wei, Jiamin & Jiming, 2020). Diversas perspectivas e relacionamentos entre agentes políticos e públicos, grupos e tópicos podem ser estudados com suas técnicas.

Trata-se de uma abordagem à pesquisa social que exibe quatro características: uma intuição estrutural, dados relacionais sistemáticos, imagens gráficas e modelos matemáticos ou computacionais (Freeman, 2004).

O modo sistemático que a Análise de Redes auxilia na revelação de estruturas ocultas que seriam difíceis de serem observadas por outros meios (Newman, 2003). A rede permite a visualização da relação entre elementos ou atores a partir da representação por meio de nós, também denominados vértices, e suas conexões ou arestas. Essa visualização da informação é crucial para apresentar resultados de fácil compreensão (Ernst, 2003; Mattos et al., 2020).

3 REVISÃO DOS TRABALHOS RELACIONADOS

A revisão da literatura científica adotou o protocolo de Kitchenham (2004) e Kitchenham e Charters (2007).

3.1 PLANEJAMENTO DA REVISÃO

O planejamento da presente pesquisa inicia com:

a) Identificação de necessidade da revisão sistemática: Analisar na literatura científica os trabalhos sobre agrupamento, modelagem de tópicos e análise de redes empregadas na mineração de texto em discursos políticos.

b) Especificação das perguntas de pesquisa: A pergunta que guia o desenvolvimento dessa pesquisa é: "Qual o estado da arte em pesquisas sobre agrupamento, modelagem de tópicos e análise de redes na mineração de textos em discursos políticos identificadas nos artigos selecionados?"

c) Desenvolvimento de um protocolo de revisão: O protocolo de pesquisa foi composto das bases de dados científicas: i – Scopus; ii – WoS - *Web of Science*; iii - IEEE Xplorer; iv – Springer Link; v – ProQuest - ASSIA; vi – Science Direct; e vii – ACM Digital.

A *string* base com os termos: *(cluster* or "topic model*" or "network analysis") and (parliament or legislative or congress* or politic*) and (speech* or discours*)*.

Os termos da *string* deveriam estar constantes no Título, Resumo ou Palavras-chave. A ausência dos termos de visão sistêmica na *string* teve como objetivo ampliar os resultados para a avaliação. Foi ainda necessário especializar a *string* base para cada base de dados devido aos seus respectivos motores de pesquisa dos seus *sites*.

Houve delimitação para retorno de apenas artigos publicados nos últimos cinco anos, já que a utilização de processamento de linguagem natural no contexto ao qual abrange esta pesquisa é recente, inclusive na literatura científica.

Os critérios de exclusão aplicados foram: i – falta de acesso ao artigo mesmo por intermédio da rede da Comunidade Acadêmica Federada (CAFe); ii – língua diferente da inglesa ou portuguesa; iii – trabalhos duplicados, e iv - excluídos *conference papers, reviews, books chapters, notes*.

Os critérios de inclusão e seleção foram: i – tratar de descoberta de conhecimento ou mineração em texto; ii – o foco da aplicação é sobre discursos políticos; iii – trabalho ser um artigo ("*article*") de periódico; e iv – artigo revisado por pares.

Foram extraídas as informações: i – base de dados; ii – ano da publicação; iii – autores; iv – título; v – resumo; e vi – palavras-chave.

d) Avaliação do protocolo de revisão: O protocolo foi revisado por especialistas. Dois professores pesquisadores doutores oriundos do Programa de Pós-Graduação de Engenharia e Gestão do Conhecimento da UFSC não integrantes da respectiva análise.

3.2 CONDUÇÃO DA REVISÃO

A pesquisa foi conduzida da seguinte maneira:

a) Identificação da pesquisa: A busca inicial nas bases foi realizada entre os dias 02 e 03 de dezembro de 2020 e as seguintes quantidades de trabalhos na Tabela 1:

Tabela 1 – Artigos retornados em cada base de dados

Base de artigos científicos	Quantidade inicial de artigos com base na *string* de busca	Eliminados por título, palavras-chave e resumo	Selecionados para avaliação
Springer	66	64	2
WoS - Web of Science	29	10	19
Scopus	39	19	20
Science Direct	117	107	10
ProQuest	27	18	9
IEEE	12	8	4
ACM Digital	1	-	1
TOTAL	291	225	65

Fonte: Elaborada pelos autores (2021).

b) Seleção dos estudos primários: Foram obtidos como resultado 65 artigos, sendo eliminados 19 por duplicidade, 1 por não estar escrito nas línguas previstas, resultando em 46 para leitura e avaliação da qualidade.

c) Avaliação da qualidade dos estudos: Os artigos selecionados foram analisados quanto à aderência aos critérios e ao tema em estudo. Eliminou-se 24 após a leitura integral e que não atenderam aos critérios. Por fim, 22 artigos selecionados, conforme Tabela 2.

Tabela 2 – Artigos selecionados

Base de artigos científicos	Selecionados
Springer	-
WoS - Web of Science	3
Scopus	12
Science Direct	2
ProQuest	2
IEEE	2
ACM Digital	1
TOTAL	22

Fonte: Elaborada pelos autores (2021).

d) Extração de dados: Os dados foram extraídos em uma planilha eletrônica para catálogo, bem como apresentados no Quadro 1.

e) Síntese dos dados: Os artigos selecionados foram sintetizados e constam na seção de trabalhos encontrados sendo analisados na sequência.

3.3 TRABALHOS ENCONTRADOS

Os estudos selecionados foram elencados conforme Quadro 1.

Quadro 1 – Artigos selecionados obtidos na pesquisa e analisados

ID	Estudo		Ano
3	Analyzing the topic distribution and evolution of foreign relations from parliamentary debates: A framework and case study	Lu Wei; Wang Jiamin; Hu Jiming	2020
5	Automatic Content Analysis of Legislative Documents by Text Mining Techniques	F. Lin; S. Chou; D. Liao; D. Hao	2015
6	Automatic content analysis of media framing by text mining techniques	Lin F.-R., Hao D., Liao D.	2016
10	Cross-national measurement of polarization in political discourse: Analyzing floor debate in the U.S. the Japanese legislatures	T. Sakamoto; H. Takikawa	2017
11	Documents as data: A content analysis and topic modeling approach for analyzing responses to ecological disturbances	Altaweel M., Bone C., Abrams J.	2019
13	Exploring the political agenda of the european parliament using a dynamic topic modeling approach	Greene D., Cross J.P.	2017
17	Gatekeeping the plenary floor: Discourse network analysis as a novel approach to party control	Bhattacharya C.	2020
21	How electoral reform alters legislative speech: Evidence from the parliament of Victoria, Australia 1992–2017	Ishima H.	2020
22	Immigration-related Speechmaking in a Party-constrained Parliament: Evidence from the 'Refugee Crisis' of the 18th German Bundestag (2013–2017)	Geese L.	2020
23	Improving fitness: Mapping research priorities against societal needs on obesity	Cassi L. et al.	2017
25	Investigating political herd mentality: A community sentiment based approach	Bhavan A. et al.	2019
26	Land, Wood, Water, and Space: Senator Robert S. Kerr, Congress, and Selling the Space Race to the American Public	Hayden, J. et al.	2017
29	Look who's talking: Two-mode networks as representations of a topic model of New Zealand parliamentary speeches	Curran B. et al.	2018
30	Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech	Gentzkow, M; Shapiro, JM; Taddy, M	2019
35	Political rhetoric through the lens of non-parametric statistics: are our legislators that different?	Iliev I.R., Huang X., Gel Y.R.	2019
40	The politics of climate finance: Consensus and partisanship in designing green state investment banks in the United Kingdom and Australia	Anna GeddesNicolas SchmidTobias S. SchmidtBjarne Steffen	2020
41	The polycentricity of climate policy blockage	Fisher, DR; Leifeld, P	2019
42	Topic identification techniques applied to dynamic language model adaptation for automatic speech recognition	Echeverry-Correa, JD et al.	2015
43	Towards topic modeling swedish housing policies: Using linguistically informed topic modeling to explore public discourse	Lindahl A., Borjeson L.	2018

44	Understanding Congressional Coalitions: A Discourse Network Analysis of Congressional Hearings for the Every Student Succeeds Act	Wang, Yinying	2020
45	Unveiling the Political Agenda of the European Parliament Plenary: A Topical Analysis	Derek Greene; James P. Cross	2015
46	Verba volant, scripta manent? Intra-party politics, party conferences, and issue salience in France	Ceron A., Greene Z.	2019

Fonte: Elaborado pelos autores (2021)

4 ANÁLISE DOS RESULTADOS

A seguir estão analisados todos os estudos selecionados encontrados e que possuem relação com o tema desta pesquisa. Também foi adicionado, após a citação do respectivo trabalho, o número do estudo entre colchetes visando a facilitar a identificação e a vinculação ao respectivo referencial extraído e presente no Quadro 1.

Para Wei, Jiamin e Jiming (2020) [#3], os textos parlamentares são como registros de discussões de assuntos nacionais e internacionais, que refletem atitudes nacionais e tendências de desenvolvimento nas relações exteriores Reino Unido-China. Primeiro, as palavras do tópico são extraídas de textos parlamentares e, em seguida, uma rede de *co-words* é construída para representar a estrutura de correlação das palavras do tópico.

As estatísticas básicas, cálculo de indicadores de rede, detecção de comunidade e visualização de mapas de rede e localização de evolução, bem como a representação de um diagrama estratégico, elucidam características e conotações profundas das relações externas.

Lin et al. (2015) [#5] utilizaram de uma estrutura categórica legislativa e, em seguida, um agrupamento de dois estágios foi aplicado para realizar a seleção de recursos para documentos legislativos. O método SVM foi utilizado para construir um modelo para classificar o documento na categoria apropriada. Já para manter as categorias de classificação atualizadas, também os autores avaliaram a diferença entre os conteúdos de rotulagem por especialistas do domínio e o público em geral.

Para minimizar a lacuna entre os documentos legislativos e o público em geral, Lin, Hao e Liao (2016) [#6] agruparam documentos legislativos e de notícias para identificar enquadramentos de mídia (*media frames*) e, em seguida, representar a proporção de cada enquadramento correspondente às fontes de informação.

O sistema de *clustering* automático determinou o enquadramento de mídia (*media framing*) com o mínimo de interferência humana. Os resultados do estudo fornecem aos especialistas do domínio político promovem evidências concretas do enquadramento da mídia e auxiliam o público em geral a descobrir o fenômeno do enquadramento da mídia.

Sakamoto e Takikawa (2017) [#10] propuseram uma nova medida de polarização, que, por intermédio da modelagem de tópicos, quantifica diferenças na articulação coletiva de agendas públicas entre atores políticos relevantes.

Os autores identificaram que os atores políticos japoneses são muito mais polarizados em sua articulação de questões do que suas contrapartes nos EUA. Por segundo, que no Japão, fatores estruturais como os papéis do partido no poder e a oposição geralmente dominam essa dinâmica, enquanto nos EUA ocorrem diferenças ideológicas persistentes sobre questões específicas entre os principais partidos políticos.

Altaweel, Bone e Abrams (2019) [#11] aplicaram análise de conteúdo e modelagem de tópicos em documentos governamentais contendo informações ecológicas. A abordagem utilizou *Latent Dirichlet Allocation* (LDA), Hierarchical Dirichlet Process (HDP) e *Term Frequency–Inverse Document Frequency* (TF-IDF) para a compreensão do discurso e do conteúdo em políticas relacionadas a distúrbios ecológicos.

Bhattacharya (2020) [#17] demonstra que no parlamento alemão, o tempo de uso da palavra é uma fonte escassa e é alocada aos parlamentares por líderes de seus respectivos grupos de partidos parlamentares. Pesquisas anteriores indicam que debates em plenário tendem a ser dominados por líderes partidários e outros defensores leais. Os discursos plenários podem, portanto, oferecer apenas percepções limitadas sobre a unidade partidária.

A fim de avaliar o efeito do controle partidário na unidade partidária observada e na contestação parlamentar, a análise da rede do discurso foi empregada neste estudo para comparar o discurso legislativo em debates sobre a crise grega entre 2010 e 2015.

O trabalho de Ishima (2020) [#21] observou uma reforma eleitoral no Parlamento de Victoria, Australia, com a mudança do Sistema de Distrito Único para Sistema de Representação Proporcional e como isso afetou o comportamento dos políticos.

Analisou-se um conjunto de dados recém-coletados de discursos de posse dos legisladores de 1992 a 2017 usando um modelo de tópico. Os resultados mostraram que a reforma eleitoral aumentou a atenção dos políticos para novas questões econômicas, mas não diminuiu a atenção aos interesses locais, como a promoção de indústrias primárias.

Geese (2020) [#22] estudou o 18º mandato do *Bundestag*, na Alemanha, cuja agenda política foi dominada pela questão dos refugiados e de asilo. Em um estudo de caso, o artigo questionou quais foram os fatores que levam os legisladores a falar sobre a imigração no plenário do parlamento. Um *corpus* de mais de 10.000 discursos foi analisado, concluindo-se que a fala dos legisladores sobre a questão dos refugiados e asilo foi moldada principalmente por fatores específicos dos grupos partidários parlamentares, e não por motivos individuais.

Cassi et al. (2017) [#23] apresentaram uma abordagem que usa modelagem de tópicos utilizando LDA para explorar: como as publicações científicas podem ser usadas para descrever as prioridades existentes na produção científica sobre a obesidade; como os registros de políticas (questões colocadas no parlamento europeu) podem ser usados como uma instância de mapeamento do discurso das necessidades sociais relativas ao assunto obesidade.

O estudo concluiu que a maioria das pesquisas relacionadas à obesidade se concentra na biologia e na medicina e uma pequena parte do portfólio de pesquisas sobre obesidade está relacionada a agendas de políticas, principalmente por meio da saúde pública e social.

O experimento de Bhavan et al. (2019) [#25] analisou as transcrições de *Hansard* disponíveis publicamente dos debates conduzidos no Parlamento do Reino Unido. A abordagem proposta pelos autores utilizou informações gráficas baseadas na comunidade para aumentar recursos feitos à mão com base na modelagem de tópicos e detecção de emoções em debates, atualmente superando os resultados de *benchmark* no mesmo conjunto de dados.

Hayden et al. (2017) [#26] examinaram a contribuição histórica do senador Robert S. Kerr (D-OK) na promoção da NASA e da exploração espacial, enquanto atuava como presidente do Comitê do Senado em Ciências Aeronáuticas e Espaciais. Os autores estimaram um modelo de tópico usando artigos de jornal, bem como discursos e comunicados de imprensa dos documentos de arquivo para descobrir as dimensões do debate sobre o espaço.

Concluíram que embora seja fácil dar crédito ao Executivo por atingir as metas nacionais, é quase impossível obter os resultados desejados sem a adesão do Legislativo, significando que os membros do Congresso precisam vender a política para os cidadãos.

Curran et al. (2018) [#29] afirmam que faltam métodos quantitativos para descrever a participação no debate dos deputados e dos partidos a que pertencem. Os autores propuseram uma nova abordagem que combina modelagem de tópicos com técnicas de redes complexas e a usaram para caracterizar o discurso político no Parlamento da Nova Zelândia.

Implementaram um modelo LDA para descobrir a estrutura temática de discursos parlamentares e construir, a partir dela, redes de dois modos que ligam os membros do Parlamento aos tópicos que discutem. Os resultados mostraram como a popularidade do tema muda ao longo do tempo e permitem relacionar as tendências seguidas pelos partidos políticos em seus discursos com eventos sociais, econômicos e legislativos específicos.

Gentzkow, Shapiro e Taddy (2019) [#30] estudaram o problema de medir as diferenças de grupo nas escolhas quando a dimensionalidade do conjunto de escolhas é

grande. Mostraram que as abordagens padrão sofreram de um viés severo de amostra finita e, para isto, propuseram um estimador que aplica avanços recentes no aprendizado de máquina.

Iliev, Huang e Gel (2019) [#35] apresentaram uma nova análise estatística da retórica legislativa no Senado dos Estados Unidos que lança uma luz sobre os padrões ocultos no comportamento dos senadores em função de seu tempo no cargo. Criaram um novo conjunto de dados abrangente com base nos discursos de todos os senadores que serviram no Comitê de Energia e Recursos Naturais dos Estados Unidos entre 2001 e 2011.

A abordagem é baseada em uma metodologia de aprendizagem híbrida supervisionada e não supervisionada de dois estágios e, a partir dela, descobriram que os legisladores se tornam muito mais parecidos após os primeiros anos de seu mandato, independentemente de seu partidarismo e promessas de campanha.

Geddes (2020) [#40] analisou o discurso parlamentar por trás do estabelecimento e projeto do Banco de Investimento Verde do Reino Unido e da Corporação Financeira de Energia Limpa da Austrália. Afirmou que o debate sobre o estabelecimento do *green investment banks* (GIBs) se concentrou em argumentos relacionados a objetivos políticos de alto nível e ao papel do estado.

O artigo de Fisher e Leifeld (2019) [#41] partiu de pesquisas recentes sobre governança policêntrica e Ecologia dos Jogos para compreender a política climática nos EUA. Complementando o trabalho anterior de 2005 a 2009, mapearam as redes ideológicas de atores políticos engajados na rede de políticas climáticas usando dados do Congresso dos Estados Unidos como uma arena de interação simbólica.

Para os autores, o conceito de policentricidade tende a ser normativamente associado à inovação de políticas, ao invés de estagnação. Na análise longitudinal, demonstram, usando a Análise de Rede do Discurso, como o aumento da participação em vários níveis pode ser associado ao bloqueio de políticas climáticas progressivas, em vez de permitir a inovação política.

Echeverry-Correa et al. (2015) [#42] apresentaram uma abordagem de reconhecimento de fala eficiente para fala multitópica, combinando técnicas de recuperação de informação e modelagem de linguagem (LM) baseada em tópicos. Técnicas baseadas na recuperação da informação, como a identificação do tópico por meio da *Latent Semantic Analysis*, foram usadas para identificar o tópico em uma transcrição reconhecida de um segmento de áudio.

Para a avaliação do sistema proposto, utilizaram a partição espanhola da base de dados das Sessões Plenárias do Parlamento Europeu (EPPS), selecionaram um subconjunto do banco de dados com 67 tópicos rotulados para a avaliação. Para a tarefa de identificação de

tópico, nos experimentos mostraram uma redução relativa no erro de identificação de tópico de 44,94% quando comparado ao método *baseline*, o *Generalized Vector Model* com um esquema de ponderação TF-IDF clássico.

Lindahl e Börjeson (2018) [#43] investigaram o efeito do pré-processamento com informação linguística na modelagem de tópicos. Por meio da avaliação humana, a filtragem dos dados com base na classe gramatical tem o maior efeito na qualidade do tópico.

A área de políticas habitacionais suecas foi escolhida, representada por documentos do parlamento sueco e textos novos da Suécia. Tópicos não lematizados foram considerados mais avaliados do que tópicos lematizados. Os tópicos de filtros baseados em relações de dependência apresentam classificações baixas.

O objetivo do estudo de Wang (2020) [#44] é investigar as coalizões de políticas do *Every Student Succeeds Act* (ESSA) em audiências no Congresso dos EUA. Realizou uma análise de rede de discurso para examinar 30 depoimentos nas audiências do Congresso sobre a ESSA desde sua aprovação em 2015.

A análise da rede de discursos sugere quatro coalizões com base nas reivindicações de política dos atores sobre (1) equidade, (2) avaliação e responsabilidade, (3) os estados mudaram/aprovaram legislação para alinhar os sistemas de responsabilidade do Estado com os objetivos ESSA, e (4) a aprovação do plano estadual do Departamento de Educação era inconsistente com as disposições legais da ESSA. Os resultados forneceram percepções sobre o processo contínuo de implementação da ESSA nos níveis federal, estadual e local.

Greene e Cross (2015) [#45] analisaram as interações políticas no Parlamento Europeu (PE), considerando como a agenda política das sessões plenárias tem evoluído ao longo do tempo e a forma como os deputados do Parlamento Europeu (MEPs) reagiram a estímulos externos e internos ao fazerem discursos parlamentares. Isso é feito considerando o contexto em que os discursos foram feitos e o conteúdo desses discursos.

Para detectar temas latentes em discursos legislativos ao longo do tempo, o conteúdo do discurso foi analisado usando modelagem de tópicos dinâmicos baseado em duas camadas de fatoração de matriz. O método foi aplicado a um *corpus* composto por discursos legislativos de língua inglesa proferidos no plenário do PE no período de 1999-2014.

As conclusões sugerem que a agenda política do PE evoluiu significativamente ao longo do tempo, afetada pela estrutura de comissões do Parlamento, reagindo a eventos exógenos que influenciam nos assuntos debatidos no Parlamento.

Ceron e Greene (2019) [#46] examinaram reuniões do Partido Socialista Francês. Os autores utilizaram Modelos de Tópicos Estruturais para analisar 74 moções, 1.439 discursos e

9 manifestos de congressos realizados entre 1969 e 2015. Avaliaram se moções *faccionais* ou discursos individuais refletiam adequadamente o conteúdo dos manifestos, e avaliaram o processo de definição da agenda interna.

Concluíram que grupos intrapartidários influenciam as prioridades políticas dos partidos e que os congressos do partido servem como locais para a tomada de decisões, permitindo discursos e moções para apoiar prioridades diferentes. Considerando o processo interno das partes, eles propuseram que as deliberações e moções alternativas afetam de forma independente as declarações de política resultantes.

4 CONCLUSÕES

A partir da perspectiva da engenharia e gestão do conhecimento este artigo contempla uma revisão sistemática trazendo o esrtado da arte acerca do tema proposto.

Foi possível observar que grande parte dos estudos utilizaram métodos e técnicas já dominadas nas Ciências da Computação como LDA, SVM, TF-IDF, HDP, ou então, os pesquisadores utilizaram *software* de PLN para gerar o conteúdo de agrupamentos ou redes.

Não foi observado a adoção de técnicas recentes como análise por gráficos de conhecimento (*Knowledge Graph Analysis*), e ainda são iniciais aqueles com uso aprofundado de *embeddings* e *transformers*, e, por exemplo, adoção de redes neurais convulucionais.

Ressalta-se que a inovação potencializada na política com abordagens de descoberta de conhecimento em texto em um ambiente parlamentar devem considerar diversos fatores circunstanciais (Welter, Corrêa & Alves, 2020), dada sua complexidade e dinâmica.

Assim, observa-se um avanço das pesquisas sobre as temáticas de descoberta de conhecimento em texto (KDT), agrupamentos e análise de redes sociais e sua aplicação na política, oferecendo importantes oportunidades de pesquisa.

REFERÊNCIAS

Abercrombie, G., & Batista-Navarro, R. (2020). Sentiment and position-taking analysis of parliamentary debates: A systematic literature review. Journal of Computational Social Science, 1-26.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

Altaweel, M., Bone, C., & Abrams, J. (2019). Documents as data: A content analysis and topic modeling approach for analyzing responses to ecological disturbances. *Ecological Informatics*, *51*, 82-95.

Bhattacharya, C. (2020). Gatekeeping the plenary floor: Discourse network analysis as a novel approach to party control. *Politics and Governance*, *8*(2), 229-242.

Bhavan, A., Mishra, R., Sinha, P. P., Sawhney, R., & Shah, R. (2019, July). Investigating political herd mentality: A community sentiment based approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop* (pp. 281-287).

Cassi, L., Lahatte, A., Rafols, I., Sautier, P., & De Turckheim, E. (2017). Improving fitness: Mapping research priorities against societal needs on obesity. *Journal of Informetrics*, *11*(4), 1095-1113.

Ceron, A., & Greene, Z. (2019). Verba volant, scripta manent? Intra-party politics, party conferences, and issue salience in France. *Party Politics*, *25*(5), 701-711.

Chistol, M. (2020, May). A Comparative Study of Parametric Versus Non-Parametric Text Classification Algorithms. In 2020 International Conference on Development and Application Systems (DAS) (pp. 208-213). IEEE.

Curran, B., Higham, K., Ortiz, E., & Vasques Filho, D. (2018). Look who's talking: Two-mode networks as representations of a topic model of New Zealand parliamentary speeches. *PloS one*, *13*(6), e0199072.

Delen, D., & Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*, *34*(3), 1707-1720.

Diermeier, D., Godbout, J. F., Yu, B., & Kaufmann, S. (2012). Language and ideology in Congress. *British Journal of Political Science*, *42*(1), 31-55.

Echeverry-Correa, J. D., Ferreiros-López, J., Coucheiro-Limeres, A., Córdoba, R., & Montero, J. M. (2015). Topic identification techniques applied to dynamic language model adaptation for automatic speech recognition. *Expert Systems with Applications*, *42*(1), 101-112.

Ernst, H. (2003). Patent information for strategic technology management. *World patent information*, *25*(3), 233-242.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Fisher, D. R., & Leifeld, P. (2019). The polycentricity of climate policy blockage. *Climatic Change*, *155*(4), 469-487.

Freeman, L. (2004). The development of social network analysis. *A Study in the Sociology of Science*, *1*(687), 159-167.

Geddes, A., Schmid, N., Schmidt, T. S., & Steffen, B. (2020). The politics of climate finance: consensus and partisanship in designing green state investment banks in the United Kingdom and Australia. *Energy Research & Social Science*, *69*, 101583.

Geese, L. (2020). Immigration-related speechmaking in a party-constrained parliament: Evidence from the 'refugee crisis' of the 18th German Bundestag (2013–2017). German Politics, 29(2), 201-222.

Gentzkow, M., Shapiro, J. M., & Taddy, M. (2019). Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, *87*(4), 1307-1340.

Greene, D., & Cross, J. P. (2015, June). Unveiling the political agenda of the european parliament plenary: A topical analysis. In *Proceedings of the ACM web science conference* (pp. 1-10).

Greene, D., & Cross, J. P. (2017). Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis*, *25*(1), 77-94.

Gonçalvez, A. L., Faraco, F. M., de Souza, J. A., Todesco, J. L., & Nunes, R. C. T. (2018). Análise de agrupamentos sobre textos: um estudo dos resumos do banco de teses e dissertações da capes: um estudo dos resumos do banco de teses e dissertações da CAPES. *Anais do Congresso Internacional de Conhecimento e Inovação – Ciki*, *1*(1). Recuperado de https://proceeding.ciki.ufsc.br/index.php/ciki/article/view/589

Hayden, J. M., Geras, M. J., Gerth, N. M., & Crespin, M. H. (2017). Land, Wood, Water, and Space: Senator Robert S. Kerr, Congress, and Selling the Space Race to the American Public. *Social Science Quarterly*, *98*(4), 1189-1203.

Iliev, I. R., Huang, X., & Gel, Y. R. (2019). Political rhetoric through the lens of non-parametric statistics: are our legislators that different?. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(2), 583-604.

Ishima, H. (2020). How electoral reform alters legislative speech: Evidence from the parliament of Victoria, Australia 1992–2017. *Electoral Studies*, *67*, 102192.

Kitchenham, B. *Procedures for performing systematic reviews*. Keele: Keele University, 33 (TR/SE-0401), 28. 2004.

Kitchenham, B.; Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. *Evidence-Based Software Engineering (EBSE).*Keele: Keele University; Durham: University of Durham.

Lin, F. R., Chou, S. Y., Liao, D., & Hao, D. (2015, January). Automatic content analysis of legislative documents by text mining techniques. In 2015 48th Hawaii International Conference on System Sciences (pp. 2199-2208). IEEE.

Lin, F. R., Hao, D., & Liao, D. (2016, January). Automatic content analysis of media framing by text mining techniques. In *2016 49th Hawaii International Conference on System Sciences (HICSS)* (pp. 2770-2779). IEEE.

Lindahl, A., & Börjeson, L. (2018). Towards Topic Modeling Swedish Housing Policies: Using Linguistically Informed Topic Modeling to Explore Public Discourse.

In *DHN* (pp. 427-444).

Lorenzini, D. (2020). Biopolítica em tempos de coronavírus. Instituto Humanitas, 14.

Mattos, Rafael S.; Artese, Leticia S.; Wolski, Luciano Z. Gonçalves, Alexandre Leopoldo. (2021). Um método voltado à Análise de Redes de termos em Bases de Patentes. Proceedings of International Conference on Information Systems and Technology Management (CONTECSI). São Paulo: Usp, 2020. p. 1-17. Recupado em: www.tecsi.org/contecsi/index.php/contecsi/17thCONTECSI/paper/view/6551

Moreira, R. C., Vaz-de-Melo, P. O., & Pappa, G. L. (2020). Elite versus mass polarization on the Brazilian impeachment proceedings of 2016. *Social Network Analysis and Mining*, *10*(1), 1-23.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, *45*(2), 167-256.

Sakamoto, T., & Takikawa, H. (2017, December). Cross-national measurement of polarization in political discourse: Analyzing floor debate in the US the Japanese legislatures. In *2017 IEEE international conference on big data (Big Data)* (pp. 3104-3110). IEEE.

Sarkar, D., Bali, R., & Sharma, T. (2018). Practical machine learning with Python. *A Problem-Solvers Guide To Building Real-World Intelligent Systems. Berkely: Apress*.

Silva, K. F. D. (2020). A reorganização da direita brasileira e o papel do Movimento Brasil Livre (MBL): da fundação ao impeachment de Dilma Rousseff (2013-2016).

Wang, Y. (2020). Understanding Congressional Coalitions: A Discourse Network Analysis of Congressional Hearingsfor the Every Student Succeeds Act. *Archivos Analíticos de Políticas Educativas= Education Policy Analysis Archives*, *28*(1), 183.

Wei, L., Jiamin, W., & Jiming, H. (2020). Analyzing the topic distribution and evolution of foreign relations from parliamentary debates: A framework and case study. *Information Processing & Management*, *57*(3), 102191.

Welter, M., Corrêa, L. M., & Alves, J. B. da M. (2020). Inovação Aberta para uma Assembleia Legislativa. Anais do Congresso Internacional de Conhecimento e Inovação – Ciki, 1(1). https://doi.org/10.48090/ciki.v1i1.957

Welter, M., Corrêa, L., & Gonçalves, A. (2020). Análise de agrupamentos em discursos políticos no parlamento. International Conference on Information Systems and Technology Management - *ISSN 2448-1041*. Acesso https://www.tecsi.org/contecsi/index.php/contecsi/17thCONTECSI/paper/view/6502

Yang, D., Kleissl, J., Gueymard, C. A., Pedro, H. T., & Coimbra, C. F. (2018). History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining. Solar Energy, 168, 60-101.