CLUSTERING ANALYSIS IN POLITICAL SPEECHES ON PARLIAMENT
ANÁLISE DE AGRUPAMENTOS EM DISCURSOS POLÍTICOS NO PARLAMENTO

ABSTRACT

Parliamentarians have the floor to defend ideologies and discuss topics of social interest. This work seeks to identify themes debated by parliamentarians in a Legislative Assembly in Brazil, by analyzing the speeches recorded in text format in minutes of ordinary and extraordinary sessions at different times of two legislatures. For that, we use Text Knowledge Discovery Methodologies, especially Text Mining and clustering techniques. As a contribution, the work highlights the potential of KDT to obtain non-obvious clusters in large text bases, when the use of statistical techniques is not adequate. For the context of social sciences, it enables the exploration and discovery of knowledge from the elements present in political speeches in a quick, visual and dynamic way.

RESUMO

Parlamentares a palavra para defender ideologias e discutir temas de interesse social. Este trabalho busca identificar temáticas debatidas pelos parlamentares na Assembleia Legislativa do Estado de Santa Catarina, a partir dos discursos registrados no formato de texto em atas de sessões ordinárias e extraordinárias em momentos distintos de duas Legislaturas. Para tanto, utilizamos Metodologias de Descoberta de Conhecimento em Texto, especialmente Mineração de Texto e técnicas de agrupamentos. Como contribuição, o trabalho evidencia o potencial do KDT para obtenção de agrupamentos não óbvios em grandes bases de textos, quando a utilização de técnicas estatísticas não é adequada. Para o contexto das ciências sociais, habilita a exploração e descoberta de conhecimento a partir dos elementos presentes nos discursos políticos de maneira rápida, visual e dinâmica.

Palavras-chave: Knowledge Discovery in Text; Documents Clustering; Political Speech; Parliament; Descoberta de Conhecimento em Texto; Agrupamento de Documentos; Discurso Político; Parlamento

1. INTRODUÇÃO

O discurso é elemento da política e, portanto, também do processo legislativo (Rocha, 2010). Na formação discursiva as "palavras, expressões, proposições, etc., mudam de sentido segundo as posições sustentadas por aqueles que as empregam, o que quer dizer que elas adquirem seu sentido em referência a essas posições", ou seja, são representações ideológicas (Pêcheux, 1990, p. 202).

A atuação política é intrinsecamente ligada à capacidade de argumentação, retórica, e expressão dos seus agentes. Políticos investidos em cargos públicos de representação parlamentar ― vereadores, deputados, senadores, ― ocupam espaços de fala para defender seus projetos políticos e interagir com pares, com a sociedade e com outros agentes, em sessões ou reuniões no curso de seu mandato parlamentar em instituições legislativas (Miguel & Feitosa, 2009).

Como resultado, muito do conhecimento obtido pelo cidadão sobre como os legisladores descrevem seu trabalho aos eleitores é derivado de pesquisas realizadas sobre alguns representantes parlamentares e de um pequeno subconjunto de declarações realizadas por eles aos eleitores (Grimmer, 2010).

Lazer, Tsur e Eliassi-Rad (2016) afirmam que a mineração de dados online pode revolucionar a compreensão dos sistemas sócio-políticos. No entanto, existe uma lacuna entre estudiosos em ciência da computação e ciências sócio-políticas. Essa lacuna se manifesta no fracasso dos cientistas políticos e da computação em preverem eventos políticos locais e globais.

Os cientistas da computação muitas vezes desconhecem as teorias sociais e políticas relevantes, o que pode levar a perguntas de pesquisa errôneas e a tirar conclusões equivocadas. Os cientistas sócio-políticos têm profundo conhecimento das teorias e *insights* sobre sistemas complexos, mas geralmente não possuem a modelagem e o conhecimento algorítmico para aprender e minerar a partir de grandes quantidades de dados. Essa disparidade pode levar ao uso de métodos insuficientes, a criação incorreta de experimentos, a má interpretação dos resultados empíricos e a previsão incorreta da existência ou magnitude dos eventos principais (Lazer, Tsur, & Eliassi-Rad, 2016).

Tal lacuna tem potencial de pesquisa dada a sua possibilidade de colaboração interdisciplinar. O tema cobre as teorias sócio-políticas relevantes e os algoritmos de última geração no processamento de linguagem natural, mineração, oriundos das técnicas e ferramentas aplicados na descoberta de conhecimento. Portanto, com base nos registros existentes de discursos políticos dos parlamentares, emerge a seguinte pergunta: c*omo identificar e representar os principais temas debatidos em instituições legislativas?*

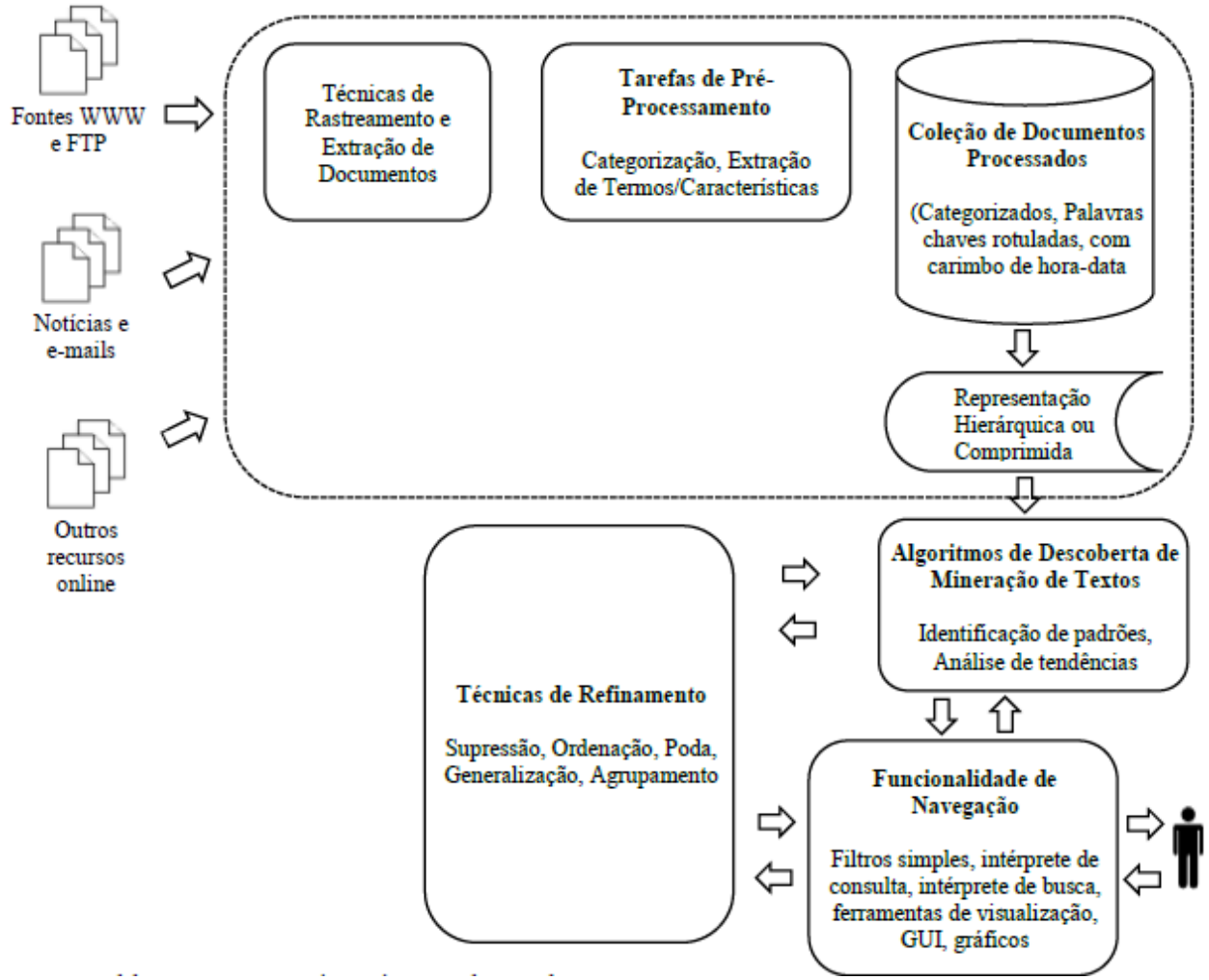
2. METODOLOGIA

Gonçalves et al. (2018) apontam que a Descoberta de Conhecimento em Textos (do inglês *Knowledge Discovery in Text* - KDT) e Mineração de Textos (do inglês *Text Mining* - TM) surgem da necessidade do uso de técnicas adequadas para propiciar uma forma de descoberta de informações úteis e relevantes de massas de dados, especialmente quando complexas, tal qual como em textos de transcrições ou resumos de discursos políticos.

Dentro do processo de KDT, a TM é um processo semiautomático de extração de conhecimento de uma grande quantidade de dados não estruturados. Onde técnicas como mineração de dados, aprendizado de máquina, processamento de linguagem natural,

recuperação da informação e *big data* são utilizados para a descoberta de conhecimento (Delen & Crossland, 2008; Gonçalves et al., 2018; Yang et al., 2018). Um sistema de TM pode contar com uma arquitetura conforme a Figura 1.

Figura 1. Arquitetura de um sistema genérico de mineração de texto



Fonte: Feldman & Sanger (2006) – Traduzido por Gonçalves et al. (2018)

Entre as tarefas de mineração de dados a análise de agrupamento caracteriza-se como uma das mais utilizadas para exploração e análise de conjuntos de dados quando não se considera nenhuma classificação prévia (Gonçalves et al., 2018), pois conforme Jain (2010), organizar os dados em grupos semelhantes permite a compreensão e a aprendizagem.

Já as descrições legíveis e inequívocas dos grupos temáticos são fatores importantes da qualidade geral do agrupamento. Eles fornecem aos usuários uma visão geral dos temas abordados nos resultados do agrupamento e os ajudam a identificar o grupo específico que estavam procurando (Osiński, Stefanowski, & Weiss, 2004).

O processo de agrupamento de dados de texto (*Document Clustering*) não é trivial e requer uma série de etapas adicionais. Igualmente aos demais processos de mineração de dados, os dados, nesse caso os textos, precisam ser tratados, limpos, trabalhados, transformados, indexados e trazidos a uma base comum. Podendo utilizar a normalização *tf-idf* (*term frequency – inverse document frequency*) para então aplicar

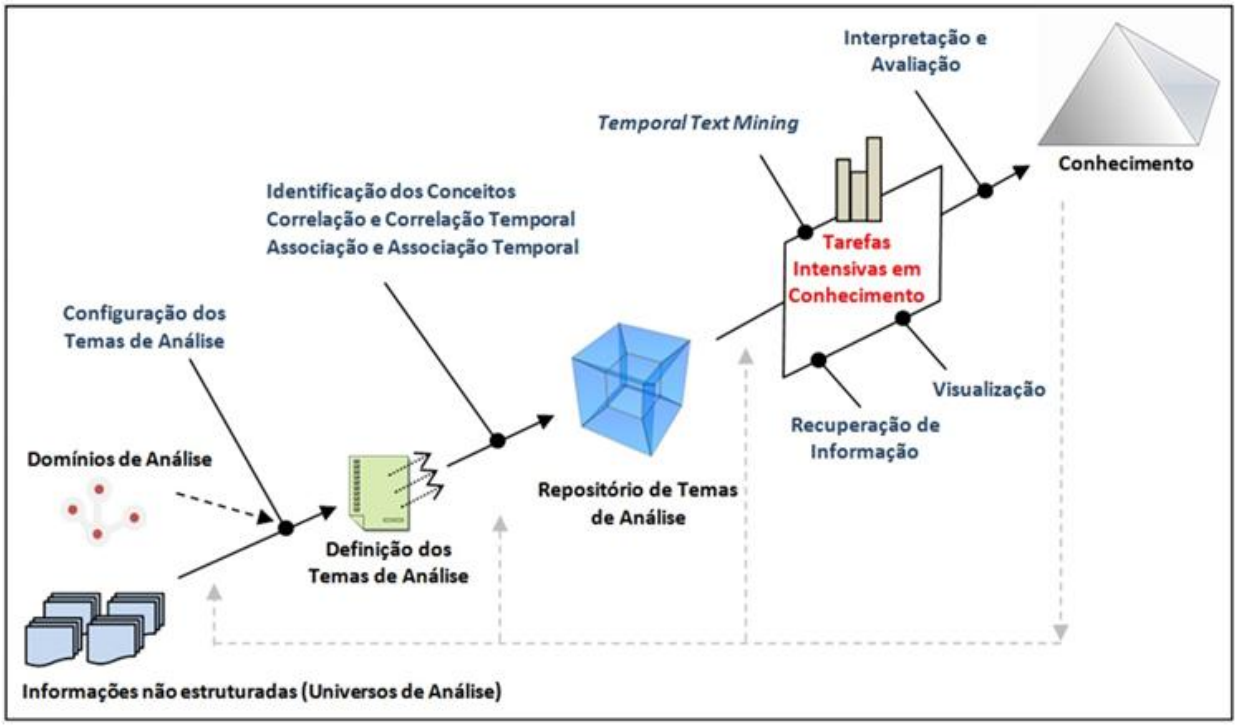
determinado algoritmo de agrupamento sobre os dados transformados (Gonçalves et al., 2018).

Conforme Schreiber et al. (2010) a Engenharia do Conhecimento permite identificar as oportunidades e os gargalos na forma como as organizações desenvolvem, distribuem e aplicam seus recursos de conhecimento, fornece os métodos para obter um entendimento completo das estruturas e processos utilizados pelos profissionais do conhecimento e ajuda, como resultado, a construir melhores sistemas de conhecimento.

Apesar de adotar Pêcheux (1990) ao introduzir o conceito e a importância do discurso, especialmente o político, esse trabalho não explora o método da Análise de Discursos, nem na visão da Linguística, Comunicação, nem avança na Ciência Política. Entretanto, possam os resultados ou potencial de replicação interessar tais campos e áreas.

O trabalho seguiu o modelo proposto por Bovo (2011), ilustrado na Figura 2, compondo 4 etapas: i) revisão na literatura; ii) seleção dos temas de análise para aplicação do algoritmo de agrupamento; iii) mineração de textos utilizando técnica de agrupamentos; e iv) interpretação e avaliação dos resultados.

Figura 2. Modelo de KDT Temporal



Fonte:Adaptado de Bovo (2011)

A aplicação ocorreu sobre discursos políticos resumidos e registrados em formato de atas taquigráficas de sessões denominadas ordinárias e extraordinárias, e o espaço temporal analisado foi a 4ª sessão legislativa da 18ª legislatura e o primeiro semestre da 1ª sessão legislativa da 19ª legislatura na Assembleia Legislativa do Estado de Santa Catarina (ALESC), ou seja, o ano de 2018 e 1º semestre de 2019, marcando o último ano dos deputados estaduais eleitos para o mandato 2015-2018 e o primeiro ano do mandato dos deputados estaduais eleitos para 2019-2022.

As atas ficam disponíveis em arquivos eletrônicos no formato *Adobe PDF*® do portal da Taquigrafia do Plenário (ALESC, 2019), os quais foram coletados e seus conteúdos extraídos para uma base de dados *MySQL*® por intermédio do uso da técnica

de programação em linguagem *python*®. O banco de dados foi instalado e disponível para consulta online em uma instância *EC2* na *Amazon Web Services*® (AWS)

As atas referentes ao tema de análise precisaram ser extraídas do corpo do texto dos arquivos por interpretação das marcações padronizadas, exemplificado na Figura 3. Foram usados os dados do cabeçalho das atas e quando encontrado início de novas subseções ou término do arquivo como fim da ata. Utilizou-se programação em *scripts* diretamente na base de dados em procedimentos e funções do próprio banco de dados *MySQL*®.

Figura 3. Trecho de uma Ata de Sessão Ordinária

ATA DA 107ª SESSÃO ORDINÁRIA DA
1ª SESSÃO LEGISLATIVA DA 19ª LEGISLATURA
REALIZADA EM 14 DE NOVEMBRO DE 2019
PRESIDÊNCIA DO SENHOR DEPUTADO JULIO GARCIA

Às 9h, achavam-se presentes os seguintes srs. deputados: Ana Campagnolo – Bruno Souza – Dr. Vicente Caropreso – Fernando Krelling – Ismael dos Santos – Ivan Naatz – Jair Miotto – Jessé Lopes – João Amin – Laércio Schuster – Nazareno Martins – Neodi Saretta – Nilso Berlanda – Sergio Motta.

PRESIDÊNCIA – Deputado Nilso Berlanda

DEPUTADO NILSO BERLANDA (Presidente) – Abre os trabalhos da sessão ordinária. Solicita a leitura da ata da sessão anterior para aprovação e a distribuição do expediente aos senhores deputados.

Neste momento, a Presidência suspende a sessão para que a Coordenadora de Cultura da Acic – Associação Catarinense para Integração do Cego, senhora Marcilene Alberton, e a Assistente Social, senhora Lidiane Barbosa, façam a divulgação do espetáculo "Os Olhos da Arte".
Na sequência, manifestou-se o escritor, senhor Afonso Rocha, para discorrer sobre temas do livro de sua autoria "*Pecador Me Confesso*!".

****
Breves Comunicações

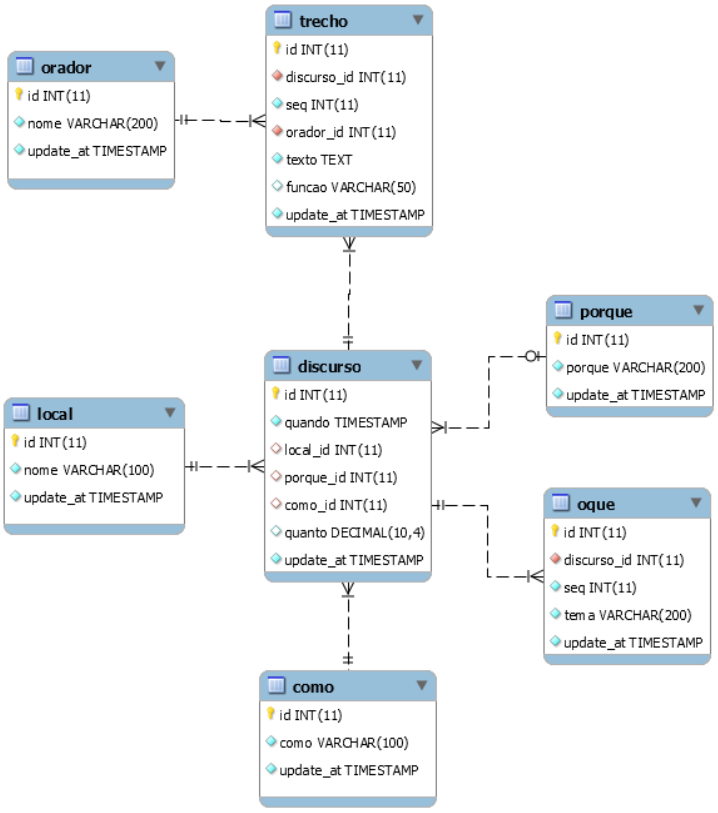
DEPUTADO NILSO BERLANDA (Presidente) – Reabre a sessão, passando às Breves Comunicações.

DEPUTADO JOÃO AMIN (Orador) – Inicia direcionando seu pronunciamento à Prefeitura Municipal de Florianópolis e à Casan – Companhia Catarinense de Águas e Saneamento.
Acusa a Casan de irresponsabilidade nos serviços de saneamento prestados à cidade de Florianópolis, que contribui com um terço da arrecadação total da empresa pública, bem como

Fonte: ALESC (2019)

A partir do texto da ata foi extraído cada trecho pertinente ao espaço de fala dos deputados estaduais conforme o Regimento Interno do parlamento (ALESC, 2019b), delimitado às Breves Comunicações (Art. 107), horário dos Partidos Políticos (Art. 108) e Explicação Pessoal (Art. 113). Por fim, identificado a fala de quem era o parlamentar, podendo ser ele na função de Orador ou de Aparteante dentro do discurso de outro orador. As informações do trecho do discurso dos parlamentares foram armazenadas no próprio banco de dados, conforme a estrutura da Figura 4.

Figura 4. Estrutura de armazenamento dos discursos na base de dados



Fonte: Autores

Visando facilitar a utilização, pesquisa e a futura expansão em forma de modelo, a estrutura de armazenamento dos discursos está inspirada na ferramenta 5W2H, composta dos elementos *Why*, *What*, *Where*, *Who*, *When*, *Howmany* e *Howmuch*, utilizada em diversos contextos para identificar fatores de influência, onde as respostas factuais a essas elementos são consideradas elaboradas o suficiente para que as pessoas entendam o discurso todo (Carmagnola, 2008; Wang, Zhao, & Wang, 2010; Koons, Schenke-Layland, & Mikos, 2019)

Selecionados e estruturados os conjuntos de dados contendo os discursos e seus elementos de compreensão, foi realizada a importação dos dados na plataforma de busca *Apache Solr*®, também em uma instância *EC2* no serviço *AWS*®, por meio da sua extensão *Data Import Handler* diretamente ao banco de dados *MySQL*®, procedimento que potencializou o tempo de resposta nas buscas.

Flexibilizado o mecanismo de pesquisa de conteúdos dos discursos, tanto pelo seu teor, como pelos demais atributos (*fields*) foram vinculados a cada elemento da resposta da ferramenta 5W2H (porque, onde, quando, quem, como, quanto, como) e posteriormente reunidos os temas dos discursos no *software Carrot*²®, o qual após agrupar pela temática do discurso, gerará o conteúdo referente ao elemento *What*, ou o "o quê".

O processamento dos agrupamentos foi realizado em computador pessoal utilizando o algoritmo *Lingo*®, disponível no *software Carrot*²®. A diferença dele em relação a outros algoritmos é quanto à descrição do agrupamento, pois objetiva descobrir o melhor termo para descrever determinado agrupamento (Osiński, Stefanowski, & Weiss, 2004), sendo que esse *software* ainda oferece gratuitamente os algoritmos STC

(*Sufix Tree Clustering*) e *K-Means* e outros comercialmente, inclusive evoluções do próprio *Lingo*®.

Cabe ressaltar que, o objetivo desse trabalho, o algoritmo escolhido não possui como foco o desempenho ou a precisão da distribuição dos discursos em torno de um termo central (centróide) em um agrupamento, e sim a geração de rótulos com carga semântica possível de auxiliar no entendimento do teor dos discursos e que atinja um ponto de equilíbrio entre a viabilidade dos agrupamentos conforme a similaridade matemática e sua significância entre rótulo do grupo e o teor do discurso.

O algoritmo *Lingo*® apresenta vantagens sobre os outros algoritmos de agrupamento (*cluster*), como suporte a *cluster* dinâmico conforme a consulta do usuário, ou seja, não é estático. Além disso, identifica primeiro o rótulo do *cluster* e depois atribui o documento a esse *cluster*. *C*om base no modelo de espaço vetorial, pode funcionar para vários idiomas e suporta várias pesquisas baseadas em palavras-chave. Ele supera a desvantagem do algoritmo *K-Means* em ter que definir previamente um número de *k clusters*, e ao SHOC criando primeiro um rótulo de *cluster* compreensível pelo usuário e, em seguida, adiciona documentos a esse *cluster*. O *Lingo*® primeiro descobre as frases com maior frequência nos documentos, em seguida, usa o método SVD (*Singular Value Decomposition*) para reduzir a matriz de documentos, para então descobrir os títulos dos *clusters* e, com base no valor de similaridade, atribuir documentos aos títulos desse *cluster* (Osiński, Stefanowski, & Weiss, 2004; Vispute & Potey, 2013).

Foram realizados processamentos no *software Carrot²* com os algoritmos STC (*Sufix Tree Clustering*) e *K-Means*, inclusive reproduzindo as parametrizações propostas por Gonçalves et al. (2018), com diversas interações e que não obtiveram resultados satisfatórios sobre o *dataset* adotado nesse contexto de aplicação.

Como o objetivo do trabalho é identificar as temáticas dos discursos parlamentares, portanto, responder o quê está sendo discutido e agrupar discursos por expressões relevantes, promovendo uma melhor visualização do domínio das temáticas, dessa forma o *Lingo*® apresenta-se como uma boa alternativa e está sendo adotado também por outros trabalhos similares (Vispute & Potey, 2013; Sergio, De Souza, & Goncalves, 2017; Goncalves et al., 2018).

Não foi aplicado explicitamente o processo de *stemming* que reduz para o radical do termo em análise, diminuindo a dimensionalidade do texto para identificar palavras similares, o *software Carrot²*® aplica tal processo de forma não assistida e automática, inclusive ao optar pelo algoritmo *Lingo*.

O algoritmo de agrupamento *Lingo*®, além de ignorar palavras consideradas *stopwords*, oferece controle mais preciso sobre os rótulos de *cluster* por meio de expressões regulares denominadas *stoplabel*. Se o rótulo de um *cluster* corresponder a um dos rótulos de parada, o rótulo não aparecerá na lista de *clusters* produzidos. Cada linha do arquivo de rótulos *stoplabel* corresponde a um rótulo de parada utilizando a notação de expressão regular do *Java*®. Para ser removido, um rótulo como um todo deve corresponder a pelo menos uma das expressões de rótulo de parada.

A ferramenta de visualização *FoamTree*® do *Carrot²*® foi utilizada para visualizar os agrupamentos gerados, onde cada elemento guarda o rótulo do grupo e a sua dimensão representa proporcionalmente o tamanho do agrupamento, ou seja, quantidade de documentos/discursos agrupados. Os dados quantitativos e os termos (*labels*) identificados pelo algoritmo de agrupamento também foram analisados, e diversas interações com reprocessamento dos resultados, atualizando a lista de *stoplabels*. Os resultados são apresentados e discutidos na seção seguinte.

3. RESULTADOS

A consulta para delimitar o período de análise foi realizada utilizando o *software Carrot² Workbench*® que permitiu o processamento e a visualização das informações para análise.

O *Lingo*® possui diversas parametrizações, todas acessíveis via o *Workbench*, sendo que foram testadas interativamente para alcançar os resultados apresentados. Todas as configurações e parametrizações não referenciadas neste trabalho foram mantidas conforme o valor padrão do *software*. Todos os processamentos do estudo utilizaram elementos padronizados para o Português, influenciando no uso das *stopwords* e das *stoplabels* no *Lingo*®. O arquivo de *stoplabels.pt* (para Português) foi elaborado de maneira incremental para melhorar os resultados e está disponível no portal do estudo (Welter, 2020).

Por padrão o algoritmo de fatoração é o *non-negative matrix factorization*. Essa configuração registrou agrupamentos interessantes após análise dos documentos presentes em cada *cluster* (análise por amostragem). Entretanto, o número de agrupamento chegou a quantidades elevadas, que para o presente contexto é até viável pela segmentação das temáticas geradas na rotulagem dos grupos. Todavia, para a visualização e para estudos mais avançados pode ser inviável tal distribuição esparsa.

Com a utilização do algoritmo de fatoração *partial singular value decomposition* (*SVD*) menos *clusters* foram produzidos, melhorando a visualização. Como o objetivo é extrair a descrição da temática do discurso, os resultados agrupados em menos *clusters* obtiveram termos com menor distinção semântica, ou seja, sintetizando em níveis onde o significado dos temas discursados não são evidenciados.

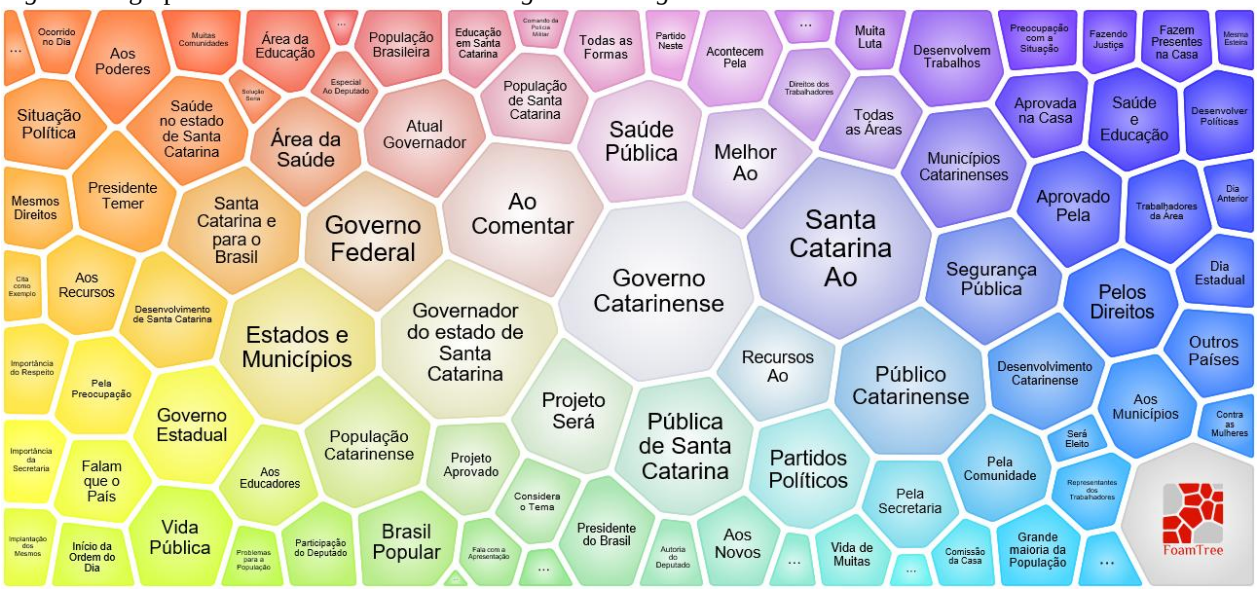
O mesmo efeito sintetizado foi observado ao testar com algoritmo *K-Means* conforme Figura 5, especialmente que o *Carrot²* esse não utiliza das expressões *stoplabel* nesse algoritmo e as palavras ou frases com pouco significado não são retiradas da rotulagem dos agrupamentos.

Figura 5. *Clusters* usando *K-Means* e SVD com 1 e 3 itens para rotulagem do grupo - *labelcounts.*

Fonte: Autores

Para atingir o objetivo da exploração das temáticas, foi observado de maneira empírica que espaços temporais muito grandes como o de 1 ano, onde são processados mais de 1.000 discursos, não se tornou eficaz na criação automática da nomenclatura dos agrupamentos, mesmo investindo na adequação dos *stoplabels*, conforme Figura 6.

Figura 6. Agrupamentos de discursos 2018 - *Lingo*® *-Non negative Matriz Factorization ED*



Fonte: Autores

Assim sendo, o período de análise considerado foi o trimestral. Entretanto, a quantidade de discursos agrupados em cada *cluster* foi considerada baixa em alguns casos, porém, ao especificar uma quantidade de *clusters* como referência ou um tamanho mínimo de documentos/discursos por *cluster* (*minimum cluster size - MCS*) o resultado

da rotulagem se tornou mais expressivo. Para a apresentação do trabalho o parâmetro *MCS* foi fixado em 2.

As figuras 7, 8, 9 e 10 apresentam os agrupamentos dos discursos em 2018.

Figura 7. Discursos de janeiro a março de 2018 - *Lingo*® – SVD

Fonte: Autores

Figura 8. Discursos de abril a junho de 2018 - *Lingo*® – SVD

Fonte: Autores

Figura 9. Discursos de julho a setembro de 2018 - *Lingo*® – SVD

Fonte: Autores

Figura 10. Discursos de outubro a dezembro de 2018 - *Lingo*® - SVD

Fonte: Autores

Nas figuras 7 e 8 correspondentes ao primeiro semestre de 2018 podem ser observados uma variedade maior de temas nos discursos. Alguns grupos poderiam ser unidos, entretanto, a diversidade abrange saúde, educação, trabalho e trabalhadores, segurança e polícia, governo estadual e nacional e os municípios.

Já na figura 9 que abrange o período pré-eleitoral e ao longo do processo eleitoral de 2018, nos quais os parlamentares reduzem a atividade na instituição legislativa para se dedicarem ao processo eleitoral, é observada uma redução na quantidade tanto de discursos como de temas debatidos.

A figura 10 repercute o momento após processo eleitoral de 2018 onde surgem agrupamentos com temas de discursos condizentes, tais como Presidente e Governo eleito, a metáfora da “luta” para expressar a atuação pretérita e futura no mandato do parlamentar. Outra validação dos agrupamentos é percebida ao se observar temas sobre

“pedágio”, devido ao fato desse assunto repercutir no ambiente legislativo ao anúncio pela Agência Nacional de Transportes Terrestres (ANTT) nos últimos meses do ano de 2018 de novas praças de pedágio no Estado de Santa Catarina.

No início do ano de 2019 ocorre a posse da nova legislatura e ocorre uma renovação dos membros do parlamento catarinense. Observa-se um aumento do uso do espaço de fala e a quantidade de discursos é superior aos mesmos períodos do ano anterior. Assim também reflete na diversidade e generalização dos agrupamentos para visualização. As figuras 11 e 12 representam o primeiro semestre de 2019.

Figura 11. Discursos de janeiro a março de 2019 - *Lingo*® - SVD

Fonte: Autores

Figura 12. Discursos de abril a junho de 2019 - *Lingo*® - SVD

Fonte: Autores

As figuras 11 e 12 demonstram a retomada da diversidade de temas em torno da saúde, educação, segurança, municípios e aos governos do estado e federal. Com o aumento da quantidade de documentos/discursos em análise pelo algoritmo, a perda semântica é novamente observada, mas mesmo assim há explicitação dos temas com

relevância e frequência maiores e que, nas representações gráficas, são demonstrados como figuras maiores e centralizadas.

Figura 13. Discursos de fevereiro de 2019 no início da 19a.legislatura - Lingo - SVD

Fonte: Autores

Como todo início da legislatura as menções ao governo anterior são notáveis e aparecem nos agrupamentos da figura 13. Outro agrupamento que comprova a validação das temáticas refere-se à situação das obras de restauração e manutenção da histórica Ponte Hercílio Luz na capital Catarinense frente ao pedido e instauração de Comissão Parlamentar de Inquérito (CPI) protocolizado no referido mês, bem como menções e agrupamentos dedicados ao governo são predominantes.

4. CONCLUSÕES

Conforme Osiński; Stefanowski e Weiss (2004) a aplicação do algoritmo *Lingo*® carece adicionar mais elementos de reconhecimento linguístico de frases sem sentido. A fase de separação de tópicos requer transformações algébricas com elevada carga computacional e abordagens incrementais com baixa utilização de memória seriam de grande importância para a escalabilidade do algoritmo. Entretanto, o conjunto de ferramentas e soluções tecnológicas em ambientes de estudo exploratório como este demonstraram ser adequadas, com processamento em computadores com característica aos de uso pessoal, tornaram viável para realizar essa pesquisa.

Para trabalhos futuros para identificação de similaridade e classificação de textos serão adotadas *word embedding models* do processamento de linguagem natural e da inteligência artificial como *Word2vec* (Mikolov et al., 2013), *Glove* (Pennington, Socher, & Manning, 2014), *FastText* (Bojanowski et al., 2017) e *Wang2vec* (Ling et al., 2015), ainda considerando suas especificidades e desempenhos para textos em português apontados por Hartmann et al. (2017). As opções de visualização das informações, bem como representações das variações das temáticas no espaço temporal são outros pontos de aperfeiçoamento.

Uma técnica de avaliação mais elaborada será necessária para estabelecer pontos fracos no método proposto, tanto sobre a solução tecnológica adotada, quanto da engenharia do conhecimento, quanto dos especialistas do domínio. Um ponto importante

de evolução é a avaliação pelas técnicas da linguística e do processamento de linguagem natural das *stoplabels* em português, visto que foi comprovado que ao aumentar ou diminuir o agrupamento, certas expressões ou palavras possuem mais ou menos valor semântico, podendo ser excluídas ou mantidas.

Como já foi ressaltado, é possível melhorar a qualidade dos agrupamentos, todavia, a contribuição deste trabalho consiste em evidenciar o potencial do método, como forma de obtenção de agrupamentos não óbvios em grandes bases de textos, especialmente discursos políticos, onde a utilização de técnicas estatísticas não se mostra adequada. Visto que o procedimento é viável e reprodutível, pode ser aplicado em diversos outros conjuntos de dados, para diversas aplicações. Para o contexto da ciência sócio-política habilita a exploração e descoberta de conhecimento a partir dos elementos presentes nos discursos políticos de maneira rápida, visual e dinâmica.

Esse trabalho faz parte dos estudos dos autores no desenvolvimento de uma plataforma de *political mining* inspirados em discussões que demonstram um ambiente promissor (Lazer et al., 2009; Lazer, Tsur, & Eliassi-Rad, 2016) na união das ciências das engenharias e da sócio-política.

5. REFERÊNCIAS

ALESC. (2019, 21 de novembro). Assembleia Legislativa do Estado de Santa Catarina: Taquigrafia do Plenário. Recuperado de http://www.alesc.sc.gov.br/taquigrafia-plenario.

ALESC. (2019, 21 de novembro). Assembleia Legislativa do Estado de Santa Catarina: Legislação. Recuperado de http://www.alesc.sc.gov.br/legislacao.

Bojanowski, P. et al. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135-146.

Bovo, A. B. (2011). Um modelo de descoberta de conhecimento inerente à evolução temporal dos relacionamentos entre elementos textuais. (Tese de Doutorado). Curso de Engenharia e Gestão do Conhecimento, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, Brasil.

Carmagnola, F. (2008, janeiro). The five Ws for user model interoperability. In 5th International Workshop on Ubiquitous User Modeling (pp. 30-36), Gran Canária, Las Palmas.

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.

Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge, England: Cambridge University Press.

Gonçalves, A. L. et al. (2018). Análise de Agrupamentos sobre Textos: Um Estudo dos Resumos do Banco de Teses e Dissertações da CAPES. In: 8th International Congress of Knowledge and Innovation-Ciki. Guadalajara. Recuperado de http://proceeding.ciki.ufsc.br/index.php/ciki/article/view/589/246

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, 18 (1), 1-35.

Hartmann, N. et al. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. Recuperado de https://arxiv.org/pdf/1708.06025.pdf. arXiv preprint arXiv:1708.06025

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, 31 (8), 651-666.

Koons, G. L., Schenke-Layland, K., Mikos, A. G. (2019). Why, When, Who, What, How, and Where for Trainees Writing Literature Review Articles. *Annals of biomedical engineering*, 47 (11), 2334-2340.

Lazer, D. et al. (2009). Computational social science. *Science*, 323 (5915), 721-723.

Lazer, D., Tsur, O., Eliassi-Rad, T. (2016). Understanding offline political systems by mining online political data. In: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining - WSDM '16. San Francisco, California.

Ling, W. et al. (2015). Two/too simple adaptations of word2vec for syntax problems. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado.

Miguel, L. F., & Feitosa, F. (2009). O gênero do discurso parlamentar: mulheres e homens na tribuna da Câmara dos Deputados. *Dados - Revista de Ciências Sociais*, 52 (1), pp. 201-221.

Mikolov, T. et al. (2013). Efficient estimation of word representations in vector space. In: Proceedings of the International Conference on Learning Representations (ICLR 2013). Scottsdale, Arizona. arXiv preprint arXiv:1301.3781.

Morgan, G. (1980). Paradigms, metaphors, and puzzle solving in organization theory. *Administrative Science Quarterly*, 25 (4), pp. 605-622.

Osiński, S., Stefanowski, J., Weiss, D. (2004). Lingo: Search results clustering algorithm based on singular value decomposition. In: Intelligent information processing and web mining. Springer, Berlin, Heidelberg. pp. 359-368.

Pêcheux, M. (1990). *O Discurso: estrutura ou acontecimento*. Campinas, SP: Pontes.

Pennington, J., Socher, R., Manning, C. (2014). Glove: Global vectors for word representation. In: Proceedings of the Conference on empirical methods in natural language processing (EMNLP). Doha, Qatar.

Ribeiro, A. C. (2018). Modelo de reconhecimento de padrões em ideias usando técnicas de descoberta de conhecimento em textos. (Dissertação de Mestrado). Curso de Engenharia e Gestão do Conhecimento, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, Brasil.

Rocha, M. M. (2008). A critique of the discursive conception of democracy. *Teoria Social*, Belo Horizonte , 4 (Selected Edition), pp. 94-117.

Schreiber, G. et al. (2000). Knowledge engineering and management: the CommonKADS methodology. Massachusetts: MIT Press.

Sergio, M. C., De Souza, J. A., & Gonçalves, A. L. (2017). Idea identification model to support decision making. *IEEE Latin America Transactions*, 15 (5), pp. 968-973.

Vispute, S. R., & Potey, M. A. (2013). Automatic text categorization of Marathi documents using clustering technique. In: 2013 15th International Conference on Advanced Computing Technologies (ICACT), New Boyanapalli, Rajampet, India.

Wang, W., Zhao, D., Wang, D. (2010). Chinese news event 5w1h elements extraction using semantic role labeling. In: 2010 Third International Symposium on Information Processing. Qingdao, China.

Welter, M. (2020, 10 de julho). Policital Mining. Recuperado de http://www.marciowelter.com.br/policitalmining

Woszezenk, C. R. (2016). Modelo para Descoberta de Conhecimento baseado em Associação Semântica e Temporal entre Elementos Textuais. (Tese de Doutorado), Curso de Engenharia e Gestão do Conhecimento, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, Brasil.

Yang, D. et al. (2018). History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining. *Solar Energy*, 168, pp. 60-101.