Revista de Sistemas de Informação da FSMA
n. 27 (2021) pp. 41-53

SISTEMAS DE INFORMAÇÃO
http://www.fsma.edu.br/si/sistemas.html

TRILHA PRINCIPAL

Mineração de Textos e Semântica: desafios, abordagens e aplicações

Roberta A. Sinoara, *Instituto Federal de São Paulo, Campus Boituva*,
Ricardo M. Marcacini, *Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo*,
Solange O. Rezende, *Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo*

***Resumo*—No cenário atual, com avanços e disseminação constantes da tecnologia da informação, tem-se uma crescente geração, armazenamento e disponibilização de textos, tanto internamente nas organizações quanto na *web* e nas redes sociais. Nesse contexto, o processo de Mineração de Textos torna-se essencial no apoio ao processamento destes dados e à descoberta de conhecimento para ser aplicado nos mais variados domínios. No entanto, a natureza dos dados textuais traz desafios adicionais ao processo de mineração, sendo que aspectos semânticos devem ser considerados. Neste artigo, a Mineração de Textos é abordada sob a perspectiva da semântica, discutindo-se os desafios inerentes ao processo de extração de conhecimento de textos. Também são apresentadas abordagens para a incorporação de aspectos semânticos nesse processo, bem como aplicações recentes que se beneficiam do enriquecimento semântico.**

***Palavras-chave*—Mineração de Textos, Semântica, Pré-processamento de Dados, Aplicações.**

Text Mining and Semantics: challenges, approaches, and applications

***Abstract*—Currently, due to the constant advance and dissemination of information technology, we face a growth in the generation, storing, and availability of texts, both internally in organizations, and widely in the web and social networks. In this context, the Text Mining process becomes essential for supporting textual data processing and knowledge discovery, which is applied in various domains. On the other hand, the nature of textual data brings additional challenges to the mining process, as semantic aspects must be considered. In this paper, we address Text Mining from a semantics perspective, discussing the challenges of extracting knowledge from textual data. We also present approaches for incorporating semantic aspects in the Text Mining process and recent applications that benefit from semantic enhancement.**

***Index Terms*—Text Mining, Semantics, Data Pre-Processing, Applications.**

I. Introdução

FAZER com que computadores consigam analisar e compreender textos escritos em língua natural tem sido foco de pesquisas há vários anos. Por exemplo, nos anos 1990, quando as primeiras pesquisas em Mineração de Dados (*Knowledge Discovery in Databases*) estavam sendo desenvolvidas, Feldman and Dagan [1] chamaram a atenção para os dados não estruturados, sendo proposto um método para estruturar os textos, permitindo assim o uso de técnicas de Mineração de Dados para a descoberta de conhecimento nesses dados não estruturados.

Voltando ainda um pouco mais no tempo, em 1950, Turing [2] discutia a pergunta "*Can machines think?*" (em português: As máquinas são capazes de pensar?) e propunha um jogo de imitação em que um computador passa-se por um humano e tenta enganar uma pessoa que não sabe se está falando com um humano ou com uma máquina. Em tal jogo de imitação, que mais tarde ficou conhecido como Teste de Turing, a comunicação entre humanos e máquinas se dá por meio de troca de mensagens textuais. Portanto, para passar no Teste de Turing, o computador deve interpretar perguntas escritas em linguagem natural e respondê-las, também em linguagem natural.

Atualmente, no cenário de *Big Data*, no qual um grande volume de dados é gerado diariamente, a aplicação de técnicas de extração automática de conhecimento torna-se essencial. Visando a identificação de padrões em textos escritos em língua natural, como os documentos textuais gerados internamente nas empresas, os artigos científicos, as revisões e comentários sobre produtos e serviços em páginas da web, e as postagens em redes sociais, pode-se aplicar o processo de Mineração de Textos (MT). Esse processo pode ser visto como uma sequência de etapas genéricas, que devem ser instanciadas de acordo com os dados disponíveis e o conhecimento que se espera obter [3], [4]. Por exemplo, a partir de um conjunto de documentos rotulados pode-se empregar técnicas de classificação de textos com o objetivo de obter classificadores que relacionem novos documentos ao conjunto de classes previamente estabelecido [5]. Já em aplicações que têm como objetivo uma organização da informação textual, porém sem conhecimento prévio sobre as classes existentes, pode-se utilizar técnicas de agrupamento de dados [6].

As possíveis aplicações desse processo são diversas e, portanto, este deve ser instanciado de acordo com a necessidade de cada aplicação. Além da organização de coleções de documentos por meio da classificação ou agrupamento de textos, a grande diversidade de textos disponíveis também tem possibilitado o uso de técnicas de MT em diversas outras aplicações, como para análise de sentimentos e opiniões, detecção de notícias falsas, construção de mapas conceituais para descoberta baseada em literatura,

Autor correspondente: Solange O. Rezende, *solange@icmc.usp.br*

recomendação de produtos e serviços, identificação de tendências sociais, políticas ou comerciais, monitoramento de desastres naturais, e predição de epidemias [7]–[14].

A Mineração de Textos pode ser vista como uma especialização da Mineração de Dados, sendo que a diferença entre elas está na natureza dos dados. Enquanto a Mineração de Dados lida com dados estruturados, a Mineração de Textos lida com dados não estruturados (textos escritos em língua natural). Assim, a principal diferença entre os processos de Mineração de Dados e de Textos está nas atividades realizadas no pré-processamento dos dados, que envolvem métodos para obter uma representação estruturada dos textos. Independentemente da aplicação ou técnica utilizada, a natureza não estruturada dos textos traz grandes desafios para a descoberta automática de conhecimento.

Para que os algoritmos de Aprendizado de Máquina (AM) possam ser utilizados na MT, os dados não estruturados devem passar por uma série de transformações para se obter uma representação estruturada dos mesmos. Nesse processo, os textos são normalmente representados no modelo espaço-vetorial (*vector space model*), formando uma matriz atributo-valor, também chamada de matriz documento-termo. Nessa matriz cada linha representa um documento e cada coluna corresponde a um atributo (ou termo) presente na coleção de documentos. As palavras são comumente utilizadas como atributos, dando origem à representação chamada de *bag-of-words*. Esse formato é simples e possibilita o uso direto de métodos de AM, porém assume que os termos são independentes e as relações entre esses termos não são consideradas.

Uma alternativa ao modelo espaço-vetorial é a representação em redes, sendo composta por objetos, que representam as entidades de um problema, e as relações entre esses objetos. Com isso, pode-se, por exemplo, representar relações entre documentos textuais ou entre as entidades que compõem os documentos, como autoria e citações. No entanto, apesar de representar informações sobre relacionamentos, assim como as tradicionais representações no modelo espaço-vetorial, as redes ainda podem ser limitadas em relação à semântica dos textos.

Um dos desafios da Mineração de Textos refere-se à esparsidade e à dimensionalidade da representação dos dados textuais. Uma coleção de documentos pode conter milhares de palavras enquanto que um de seus documentos pode ser formado por apenas um conjunto pequeno dessas palavras. Esse fato pode, por exemplo, tornar o processo de análise muito custoso computacionalmente, ou até mesmo inviável, além de afetar negativamente o resultado de alguns algoritmos utilizados para extração de conhecimento.

Os problemas de alta dimensionalidade e esparsidade têm sido tratados no pré-processamento dos dados com atividades como remoção de *stopwords*, normalização de termos (como radicalização e lematização), uso de termos compostos (n-gramas) em substituição a termos simples, e aplicação de técnicas para extração ou seleção dos atributos mais relevantes [4], [15]–[17]. Além de diminuir a dimensionalidade e/ou esparsidade da representação dos textos, essas atividades também podem levar ao aumento da representatividade dos atributos. No entanto, ainda existe outro grande desafio relacionado com a semântica dos dados textuais.

Uma representação dos textos como um conjunto não ordenado de palavras, como a representação *bag-of-words*, ignora propriedades relacionadas à sintaxe e às relações semânticas existentes em textos em língua natural, como voz ativa e passiva, termos sinônimos e hiperônimos. Isso tem como consequência a perda de parte das informações contidas nos textos. O seguinte exemplo, apresentado por Sinoara [18], ilustra esse fato. Considere uma coleção formada por quatro documentos, $D1$, $D2$, $D3$ e $D4$, apresentados a seguir.

- $D1$ = *A Empresa Alfa adquiriu a Empresa Beta.*
- $D2$ = *A Empresa Beta adquiriu a Empresa Alfa.*
- $D3$ = *A Empresa Beta foi adquirida pela Empresa Alfa.*
- $D4$ = *A Empresa Alfa comprou a Empresa Beta.*

Uma matriz que representa essa coleção de documentos, considerando o formato *bag-of-words* é apresentada na Figura 1.

Fig. 1. Exemplo ilustrativo de uma *bag-of-words* dos documentos $D1$, $D2$, $D3$ e $D4$ [18]

	empr	alf	adquirir	bet	foi	compr
$D1$	2	1	1	1	0	0
$D2$	2	1	1	1	0	0
$D3$	2	1	1	1	1	0
$D4$	2	1	0	1	0	1

Nesse exemplo é possível notar duas limitações da representação utilizada. Considerando a *bag-of-words*, os documentos $D1$ e $D2$ possuem a mesma representação e são considerados iguais, apesar de apresentarem sentidos opostos. Analisando as sentenças apenas por suas palavras, não é possível diferenciá-las. Porém, ao se considerar a sintaxe das sentenças, pode-se perceber que elas são opostas. As sentenças de $D1$ e $D2$ possuem o mesmo verbo e o sujeito de uma é o objeto da outra. Analisando um pouco mais a fundo, considerando a semântica das sentenças por meio dos papéis semânticos, pode-se identificar quais são os agentes de cada sentença. Assim, considerando os papéis semânticos, é possível perceber que $D1$ é igual a $D3$, apesar de terem sujeitos e objetos opostos. Já considerando as relações semânticas entre as palavras (como sinonímia, por exemplo), pode-se perceber que $D4$ expressa o mesmo evento de $D1$ e $D3$. Com a *bag-of-words* essas relações semânticas não são representadas.

Dependendo da aplicação, o tratamento adequado de informações semânticas dos textos pode levar a resultados mais adequados [3]. Jain [19] afirma que a representação dos dados é um dos fatores que mais impacta na qualidade do resultado obtido. Apesar deste autor tratar o caso específico da tarefa de agrupamento de dados, o mesmo pode ser generalizado para as demais tarefas. Em se tratando de dados textuais, relações semânticas têm impacto sobre o significado do conteúdo dos documentos e podem servir

para, por exemplo, diferenciar documentos que utilizam a mesmo vocabulário mas que apresentam ideias diferentes sobre um mesmo assunto [18].

Neste artigo, a Mineração de Textos é abordada sob a perspectiva da semântica. Em especial, são discutidos os desafios inerentes à extração de conhecimento de textos e são apresentadas abordagens que têm sido desenvolvidas visando a incorporação de aspectos semânticos no processo de Mineração de Textos, bem como exemplos de aplicações recentes que se beneficiam desse enriquecimento semântico.

Este artigo está organizado da seguinte maneira. Na Seção II é apresentada uma visão geral do processo de Mineração de Textos. Conceitos relacionados à análise semântica de textos, tais como relações semânticas e tarefas de Processamento de Língua Natural, são apresentados na Seção III. O desafio semântico inerente ao processo de Mineração de Textos é discutido na Seção IV e, na Seção V, são apresentadas abordagens que podem ser adotadas para incorporar aspectos semânticos nesse processo. Aplicações da Mineração de Textos que se beneficiam de um tratamento voltado à semântica são apresentadas na Seção VI. Por fim, na Seção VII são apresentadas as considerações finais.

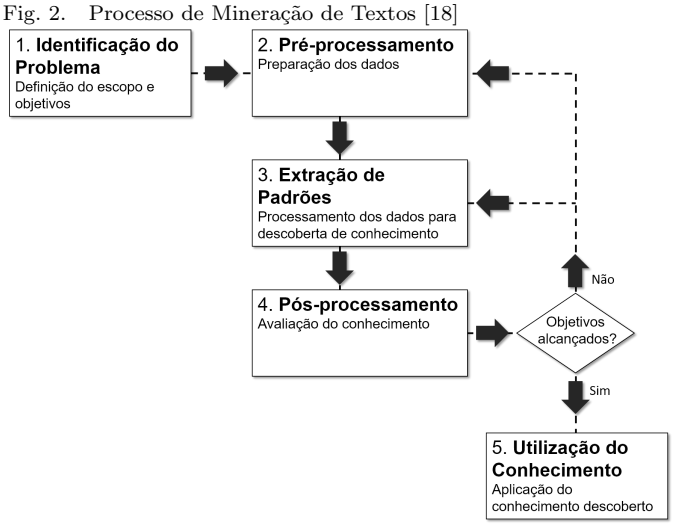
II. Mineração de Textos

APESAR de não haver um consenso entre diferentes comunidades de pesquisa [20], a Mineração de Textos pode ser vista como a aplicação de um conjunto de técnicas usadas para analisar dados não estruturados e descobrir padrões que não eram conhecidos previamente [3]. Com o crescente aumento e variedade de documentos textuais, tanto internamente em organizações quanto em redes sociais e web, as técnicas de MT têm se tornado essenciais no apoio à descoberta de conhecimento. Com isso, as fontes de textos, bem como as aplicações da MT, são variadas.

De forma geral, o processo de MT pode ser visto como um processo formado por cinco etapas, conforme ilustrado na Figura 2. Esse processo se inicia com a especificação de seus objetivos na etapa de Identificação do Problema. Nesta etapa, o analista, especialista em MT, deve delimitar o escopo da mineração em conjunto com um especialista do domínio de aplicação. Devem ser definidos as coleções de textos que serão mineradas e como os resultados serão utilizados. As especificações definidas nessa etapa irão guiar as próximas etapas do processo de MT, as quais podem ser executadas em ciclos de preparação dos dados (etapa de Pré-processamento), descoberta de conhecimento (etapa de Extração de Padrões) e avaliação do conhecimento (etapa de Pós-processamento).

A etapa de Pré-processamento consiste na preparação dos dados para a extração de padrões. Definidos o escopo e os objetivos do processo, na etapa de Pré-processamento busca-se colocar os dados em um formato adequado para a extração de conhecimento, normalmente sendo realizadas atividades de tratamento, limpeza e redução do volume de dados disponível na base. É nessa etapa que os documentos

Fig. 2. Processo de Mineração de Textos [18]



são representados de maneira a torná-los processáveis pelos algoritmos usados para extração de padrões. As atividades realizadas na etapa de Pré-processamento são cruciais para o sucesso de todo o processo de MT. Os dados pré-processados devem preservar os padrões ocultos nos documentos para que os padrões de interesse possam ser descobertos na próxima etapa do processo. Desse fato vem a importância do modelo de representação de textos adotado.

Com a coleção de documentos devidamente formatada e tratada, pode-se iniciar a etapa de Extração de Padrões. As tarefas a serem realizadas são definidas de acordo com o objetivo final do processo. Na etapa de Extração de Padrões, o analista aplica um algoritmo de aprendizado adequado para extrair os padrões dos dados pré-processados. A escolha do algoritmo é feita com base nos dados disponíveis e no tipo de conhecimento que se deseja descobrir.

Uma vez obtidos os padrões acerca dos dados, esses devem ser avaliados e interpretados na etapa de Pós-processamento. Assim como as demais, essa etapa também deve ser guiada pelos objetivos definidos no início do processo. Pode-se avaliar diversos aspectos do conhecimento extraído, como representatividade, novidade, validade e aplicabilidade. Essa avaliação pode ser realizada junto a um especialista do domínio e/ou por meio da aplicação de medidas objetivas de avaliação.

Após a etapa de Pós-processamento, caso o conhecimento extraído cumpra os objetivos estabelecidos para o processo de MT, o mesmo pode ser disponibilizado aos usuários, dando início à etapa final do processo, a Utilização do Conhecimento. Caso contrário, outro ciclo deve ser executado, realizando mudanças nas atividades de preparação dos dados e/ou em parâmetros da extração de padrões. Se forem necessárias mudanças nos objetivos estabelecidos ou nas coleções de textos utilizadas, o processo de MT deve ser reiniciado na etapa de Identificação do Problema.

III. Análise Semântica

NA Linguística, Semântica é o ramo que compreende o estudo do significado das palavras [21]. A Linguística apresenta diversas frentes de estudo da semântica, cada uma das quais dá um enfoque diferente no estudo do significado. O estudo da semântica vai desde a relação entre as palavras e seres/coisas do mundo real até as mudanças de sentido que as palavras sofrem com o tempo. Assim, quando pensa-se em semântica pensa-se em sentido, em significado de algo. Riemer [22] apresenta uma introdução sobre Semântica, mostrando as diferentes abordagens para o estudo do significado. Tal tema é bastante abrangente e não há consenso entre os pesquisadores da Linguística sobre os limites da Semântica.

Neste artigo, o termo semântica é usado em um sentido geral, considerando o sentido ou significado de itens linguísticos, sejam eles palavras, expressões ou documentos completos. Neste contexto, nesta seção são apresentadas algumas relações semânticas que têm impacto no significado dos textos, bem como algumas tarefas de Processamento de Língua Natural que lidam com a semântica.

As palavras apresentam diversas relações entre si. Conforme apresentado por Pietroforte [23], entre essas relações estão:

- **Sinonímia** - Termos sinônimos podem se substituir em determinado contexto. Por exemplo: as palavras "novo" e "jovem" são sinônimas quando se trata da característica de um ser vivo, como em "homem novo"/"homem jovem", porém "jovem" não pode substituir "novo" quando se trata da característica de um objeto, como em "livro novo" [23].
- **Antonímia** - Termos antônimos possuem significados contrários. Assim como os sinônimos, os antônimos também dependem do contexto. Palavras diferentes podem ter o mesmo antônimo desde que possuam ao menos um sentido em comum. Por exemplo: "velho" pode ser antônimo tanto de "fresco" quanto de "novo" [23].
- **Hiperonímia** - Um termo é hiperônimo de outro quando existe uma relação de englobamento entre eles em uma hierarquia de classificação. Hiperônimo é o termo englobante. Por exemplo: "esporte" é hiperônimo de "futebol".
- **Hiponímia** - Um termo é hipônimo de outro termo que é seu hiperônimo. Exemplo: "futebol" é hipônimo de "esporte".
- **Homonímia** - Termos homônimos são termos com origens distintas e significados distintos, mas que apresentam a mesma forma gráfica (termos homógrafos), fonética (termos homófonos) ou ambas (homônimos perfeitos). Exemplo: o termo "cobra" pode ser tanto o substantivo, nome de um animal, quanto o verbo cobrar conjugado no presente para a terceira pessoa do singular.
- **Holonímia** - Um termo é holônimo de outro termo quando existe uma relação parte - todo entre eles. Holônimo é o termo que corresponde ao todo na relação parte-todo. Exemplo: "carro" (todo) é holônimo de "freio" (parte).
- **Meronímia** - Um termo é merônimo de outro termo que é o seu holônimo. Por exemplo: "freio" é parte de "carro", portanto "freio" é merônimo de "carro".
- **Polissemia** - Um termo é polissêmico quando possui mais de um significado. Exemplo: o termo "manga" pode tanto se referir à uma fruta quanto à uma parte de uma camisa.

Além do significado das palavras, o significado de uma frase (ou sentença) também depende da sua estrutura gramatical. Por exemplo, as frases "João matou o bandido" e "O bandido matou João" possuem as mesmas palavras, porém, dada a estrutura sintática diferente (alternância entre sujeito e objeto), elas apresentam significados distintos [24]. Além disso, assim como as palavras, as sentenças também possuem relações entre si:

- **Paráfrase** - Corresponde à noção de sinonímia estendia para sentenças.
- **Acarretamento** - Corresponde à noção de hiponímia estendida para sentenças.
- **Contradição** - Duas sentenças são contraditórias quando elas não podem ser simultaneamente verdadeiras.
- **Ambiguidade** - Uma sentença é ambígua quando ela pode ter mais de um sentido. A ambiguidade de uma sentença pode ser causada por uma palavra ambígua, por diferentes estruturas sintáticas possíveis, ou por uma ambiguidade semântica (causada por relações anafóricas, dêiticas ou de escopo, relações descritas na sequência).
- **Relação Anafórica** - Ocorre quando um pronome presente na sentença se refere a um nome citado anteriormente na mesma sentença.
- **Relação Dêitica** - Ocorre quando um pronome presente na sentença se refere a um ente que existe no contexto.
- **Relação de Escopo** - Ocorre quando a interpretação de uma expressão da sentença depende da interpretação de outra. Exemplo: "Cada aluno leu dois livros" pode significar que cada aluno leu quaisquer dois livros ou que dois determinados livros foram lidos pelos alunos.

Essas relações semânticas entre palavras e sentenças influenciam como as pessoas interpretam os textos e podem ser importantes para a Mineração de Textos. O entendimento de textos escritos em língua natural é um processo complexo, que se dá por meio do conhecimento das palavras e de seus significados, das relações existentes entre as palavras, bem como do conhecimento de mundo e do contexto no qual o texto foi escrito.

Buscando uma representação mais rica de documentos escritos em língua natural, é possível utilizar recursos de algumas tarefas da área de Processamento de Língua Natural. Algumas dessas tarefas são apresentadas brevemente a seguir.

- **Reconhecimento de Entidades Nomeadas** - Tarefa de extração de informação que envolve processar

um texto e identificar as ocorrências de palavras ou expressões pertencentes a categorias de entidades nomeadas [25], [26]. São exemplos de categorias de entidades nomeadas: Pessoa, Organização e Local. Além das entidades identificadas por nome próprio, também é comum o reconhecimento de expressões temporais e numéricas.

- **Anotação de Papéis Semânticos** - Tarefa que busca identificar o predicado de uma oração e atribuir papéis semânticos a seus argumentos [27], [28]. Com os papéis semânticos obtém-se informações do tipo "quem fez o que para quem", além de "como fez" e "quando fez". Como exemplos de papéis semânticos tem-se Agente (aquele que inicia a ação), Paciente (aquele afetado pela ação), Instrumento (algo ou meio utilizado para efetuar a ação) e Local (lugar de um objeto ou ação).
- **Desambiguação Lexical de Sentidos** - Tarefa que busca determinar qual sentido uma palavra apresenta quando é utilizada em determinado contexto [29]–[31]. Normalmente essa tarefa é realizada com o apoio de recursos léxicos, como a WordNet [32], que agrupa as palavras em conjuntos de sinônimos e apresenta relacionamentos entre esses conjuntos e seus membros.
- **Tratamento de sinônimos** - Tarefa relacionada ao tratamento das relações de sinonímia que as palavras podem apresentar. Diferentes palavras podem se substituir em uma sentença sem que o significado expresso seja alterado. A WordNet é um recurso bastante utilizado nessa tarefa, visto que apresenta uma lista de sinônimos para cada sentido de uma palavra.
- **Resolução de correferências** - Tarefa que busca identificar todas as expressões que se referem a uma mesma entidade no texto. Uma expressão anafórica, que se refere a uma entidade que foi apresentada anteriormente no texto, pode ser pronomial (como ele, ela, meu) ou definida (a aluna, o presidente) [33].
- **Similaridade Semântica** - Tarefa que busca determinar o grau de equivalência semântica entre um par de itens linguísticos, que podem ser palavras, conceitos ou sentenças, por exemplo. Diversas medidas e abordagens têm sido propostas para medir similaridade semântica e corpus têm sido construídos para possibilitar a avaliação das propostas [34]–[38].

As relações semânticas apresentadas e o tratamento delas por meio das tarefas da área de Processamento de Língua Natural podem contribuir para a obtenção de melhores resultados na Mineração de Textos, auxiliando no tratamento do desafio semântico discutido na próxima seção.

IV. Desafio semântico

CONFORME apresentado na Seção III, o entendimento da língua natural é um processo complexo. A fim de se entender o significado de textos escritos em língua natural é necessário ter conhecimento sobre: (i) vocabulário utilizado, ou seja, conhecer o significado das palavras; (ii) gramática do idioma, ou seja, conhecer as regras que definem como as palavras são utilizadas e combinadas; (iii) relações semânticas entre os itens linguísticos, tais como sinonímia e hiperonímia; e (iv) conhecimento de mundo e do contexto no qual os textos foram escritos. Textos são uma fonte rica de conhecimento, porém seu formato não estruturado e passível de ambiguidade, sarcasmo, ironia e outros fenômenos que podem alterar o significado composicional do que é dito [22], trazem grandes desafios ao processo de descoberta automática de conhecimento.

É possível perceber algumas características importantes e que podem ter impacto no processo de MT analisando alguns exemplos. Por exemplo, para o entendimento das quatro sentenças apresentadas na Seção I (documentos $D1$, $D2$, $D3$ e $D4$) e a relação entre elas, além do vocabulário, estão envolvidos:

- **Relação sujeito x objeto** - As sentenças $D1$ e $D2$ são diferenciadas em uma análise sintática, pela ordem com que as mesmas palavras aparecem. Ambas possuem as mesmas palavras, porém o sujeito de $D1$ é o objeto de $D2$ e vice-versa.
- **Voz ativa x voz passiva** - As sentenças $D1$ e $D3$ reportam o mesmo fato, apesar de terem sujeitos e objeto/agente da passiva opostos.
- **Sinonímia** - As sentenças $D1$ e $D3$ reportam o mesmo fato, apesar de usarem verbos diferentes.

A fim de ir um pouco mais adiante nessa análise e ilustrar como esses e outros fatores podem afetar o problema de classificação automática de documentos, apresenta-se um outro exemplo. Considere a existência de uma coleção de notícias sobre diferentes esportes e que precisa ser classificada por esporte. Neste cenário, a questão que envolve a organização dessa coleção seria "Qual é o assunto do documento?". Suponha que nessa coleção exista os seguintes documentos.

- $D5$ = *Guga é o campeão do Tennis Masters Cup. Ele venceu Agassi por três sets a zero no jogo final.*
- $D6$ = *Hamilton larga na pole position e vence o Grande Prêmio do Canadá. Após colisão, Massa abandona a prova.*

O documento $D5$ possui os termos "Guga", "Tennis Masters Cup", "sets", "Agassi" e "jogo". E o documento $D6$ possui os termos "Hamilton", "pole position", "Grande Prêmio", "Massa" e "prova". Esses termos são bem característicos de seus esportes. Assim, considerando esses termos, pode-se dizer que os documentos são de dois grupos (ou classes) distintos: $D5$ é sobre Tênis e $D6$ é sobre Fórmula 1. Nesse exemplo, cada esporte (cada grupo esperado ou classe conhecida) possui seus termos (ou palavras-chave) específicos. Os documentos de um mesmo esporte terão palavras similares. Portanto, a classe (ou grupo esperado) pode ser determinada em grande parte pelo vocabulário utilizado.

No entanto, usuários diferentes ou situações diferentes podem requerer outras naturezas de classificação ou organização dos mesmos documentos. Considere agora que é desejado organizar a mesma coleção de notícias de

esportes por uma outra perspectiva, a do desempenho de atletas brasileiros em competições. Portanto, nesse novo cenário, a questão que envolve a organização da coleção de documentos seria "Esse documento refere-se a vitória de um atleta brasileiro?". Considerando novamente os documentos $D5$ e $D6$, para este caso, as informações importantes são "Guga é o campeão" e "Massa abandona a prova". E para organizar corretamente esses documentos é necessário saber que Guga e Massa são atletas brasileiros e que "campeão" e "abandona" indicam, respectivamente, vitória e derrota. Como isso, pode-se dizer que $D5$ refere-se a uma vitória de brasileiro e $D6$ refere-se a uma derrota.

Nesse contexto, pode-se separar os problemas de classificação de documentos em dois níveis de complexidade semântica [18], [39]. O primeiro nível, chamado de organização por tópico, consiste em problemas de classificação que dependem basicamente do vocabulário. Nesse problema, cada classe possui termos bastante característicos, e, portanto, o léxico (vocabulário) possui grande relevância para representar o conteúdo dos documentos. Pode-se dizer que os documentos podem ser diferenciados em grande parte pelas palavras utilizadas.

O segundo nível de complexidade semântica engloba os demais problemas de classificação de documentos. Esse segundo nível é chamado de organização semântica, no sentido de que se necessita mais do que apenas o léxico para resolvê-lo. Tais problemas requerem uma análise mais profunda, além apenas das palavras, visto que os documentos de classes distintas podem usar o mesmo vocabulário.

Os problemas de organização de documentos tratados na área de MT são tradicionalmente problemas do primeiro nível de complexidade semântica, a organização por tópico. Seja por meio da classificação ou do agrupamento, normalmente espera-se organizar os documentos com base no assunto dos mesmos. Esse fato fica claro logo na introdução de [5]. Em sua revisão, o autor apresenta a tarefa de classificação automática de textos como sendo uma tarefa de detecção de tópicos, com a rotulação de textos escritos em língua natural por meio da atribuição de uma categoria temática presente em uma lista de categorias pré-definidas. Nas palavras do autor: "*Text categorization (TC — a.k.a. text classification, or topic spotting), the activity of labeling natural language texts with thematic categories from a predefined set*" [5, p. 1]. Pode-se perceber que essa definição corresponde aos problemas do primeiro nível de complexidade semântica (organização por tópico).

A predominância dos problemas de organização por tópico nas pesquisas de MT também pode ser verificada ao se analisar coleções de textos de *benchmarking*. Sinoara [18] apresenta uma análise de 45 coleções de *benchmarking* apresentadas por Rossi et al. [40], sendo que sete delas estão entre as coleções de textos mais citadas em estudos de Mineração de Textos que consideram aspectos semânticos, identificados por Sinoara et al. [41]. Apenas três das 45 coleções são classificadas como organização semântica, sendo coleções de análise de sentimentos, que visam a classificação do sentimento (polaridade) no nível de documento.

A classificação do sentimento é um caso particular de problemas do segundo nível de complexidade semântica. Nesse tipo de classificação, palavras e expressões de sentimento são indicadores importantes do sentimento manifestado no documento. Palavras de sentimento são palavras que normalmente são utilizadas para expressar sentimentos positivos ou negativos, tais como "bom", "péssimo", "incrível". Dada a importância de tais palavras na análise de sentimentos, várias pesquisas têm o foco na criação de listas de palavras de sentimento (*sentiment lexicon*) [42]. Um exemplo de recurso léxico bastante utilizado em aplicações de análise de sentimentos para textos em inglês é a SentiWordNet [43]. Para textos em português, Balage Filho et al., [44] comparam três recursos léxicos de sentimento: LIWC, OpinionLexicon e SentiLex.

Enquanto em problemas de classificação por tópico os termos de domínio são importantes para distinguir as classes dos documentos, na classificação de sentimento as palavras de sentimento, que normalmente são adjetivos ou advérbios, são de grande importância. No entanto, apesar de importantes, as palavras de sentimento não são suficientes para resolver o problema de análise de sentimentos [42], encaixando esse problema no segundo nível de complexidade semântica. Entre os desafios da análise de sentimentos estão: (i) palavras de sentimento podem ter orientações (positiva ou negativa) diferentes em diferentes contextos ou domínios de aplicação; (ii) a presença de *sentiment shifters*, como as palavras de negação, alteram a orientação de palavras de sentimento; (iii) uma sentença contendo palavras de sentimento pode não expressar um sentimento; (iv) o autor do texto pode estar sendo sarcástico; e (v) sentenças sem palavras de sentimento podem conter opiniões implícitas.

Na próxima seção são apresentadas abordagens que podem ser utilizadas para incorporar aspectos semânticos na Mineração de Textos. A incorporação de aspectos semânticos no processo é importante especialmente para os casos do segundo nível de complexidade semântica, para os quais apenas o léxico não é suficiente para resolver o problema.

V. Abordagens para incorporação de aspectos semânticos

UMA revisão da literatura realizada sobre estudos de Mineração de Textos que consideram aspectos semânticos revela que o uso de representações mais ricas é o foco de muitos estudos [41]. Grande parte dos trabalhos concentra-se na proposta e/ou uso de atributos mais elaborados para representar os documentos no modelo espaço-vetorial.

O uso de conceitos com base em fontes externas de conhecimento, como a WordNet ou Wikipedia, e a aplicação de métodos de Processamento de Língua Natural são abordagens exploradas para enriquecer a representação de textos [45]–[50]. Técnicas de Processamento de Língua Natural podem auxiliar a Mineração de Textos em relação ao tratamento da semântica, visto que essa área estuda tarefas de análise semântica como as apresentadas na

Seção III, que contribuem para uma melhor definição do conteúdo dos textos.

Outra abordagem comum em trabalhos realizados na direção do uso da semântica na Mineração de Textos é a aplicação de técnicas de modelagem de tópicos (*topic modeling)*, como *Probabilistic Latent Semantic Indexing* (PLSI) e *Latent Dirichlet Allocation* (LDA), para obter atributos semanticamente mais ricos [3], [51]–[53]. Tais atributos formam um espaço de semântica latente (*latent semantic space*), que é um espaço de vetores de dimensão fixa e normalmente baixa, no qual formas alternativas de se expressar determinado conceito são projetadas para uma representação comum. Assim, tais representações lidam com a semântica dos textos de forma latente e reduzem ruídos causados por sinonímia e polissemia.

Ainda considerando a semântica latente, outra abordagem que apresenta resultados bastante promissores é o uso de modelos preditivos de semântica distribucional. A semântica distribucional é uma área de pesquisa que estuda e desenvolve teorias e métodos para o cálculo de similaridade semântica entre itens linguísticos, como palavras e expressões. Os modelos de semântica distribucional (*distributional semantic models*) baseiam-se na hipótese distribucional (*distributional hypothesis*), que afirma que palavras que ocorrem em contextos similares tendem a ter significados similares [54]. Aplicando essa hipótese a representações de palavras no modelo espaço-vetorial, tem-se que palavras podem ser representadas por vetores cujas dimensões são contextos. Assim, palavras cujos vetores são similares tendem a ter sentidos similares.

Analogamente à representação de documentos por meio de uma *bag-of-words*, os modelos de semântica distribucional tradicionais são baseados na contagem de contextos. Um contexto pode ser definido de diversas maneiras, como janelas de palavras ou dependências sintáticas [54], [55]. Modelos construídos dessa maneira tradicional, também podem ser chamados de modelos de contagem (*count models*) [56].

Nos últimos anos surgiu uma nova abordagem para construção de modelos de semântica distribucional. Tal abordagem, inicialmente desenvolvida para modelagem de língua, faz uso de redes neurais artificias para predizer a próxima palavra dado um contexto. Como palavras similares aparecem em contextos similares, a rede aprende a atribuir vetores similares a palavras similares. Assim, como resultado desse treinamento, obtém-se um conjuntos de vetores que modelam os contextos em que as palavras do corpus são observadas. Esses modelos de semântica distribucional são chamados de modelos preditivos (*predictive models*), *neural language models* ou, simplesmente, *embeddings*.

Avanços como esses na representação de itens linguísticos (como palavras, expressões e sentidos de palavras) têm sido importantes em tarefas de semântica lexical, tais como desambiguação lexical de sentidos, anotação de papéis semânticos, identificação de similaridade semântica e analogia [54], [56], [57]. A representação semântica de itens linguísticos é fundamental para o entendimento mais profundo de textos escritos em línguas naturais. Em especial, tem-se a representação dos sentidos, que é mais precisa do que a representação de palavras, visto que essa última é afetada pela polissemia e homonímia. Como uma mesma palavra pode assumir diferentes significados, a representação de sentidos resulta em representações mais precisas [57].

Vários trabalhos relacionados à geração de representações vetoriais para palavras utilizando-se modelos de redes neurais têm sido desenvolvidos na área de semântica lexical nos últimos anos. Um destaque nessa área foi a proposta dos modelos de aprendizado de vetores chamados de *Continuous Bag-of-Words* e *Skip-gram* [58]. Após essa proposta, diversos trabalhos foram desenvolvidos com extensões ou aplicações desses modelos. Para verificar a eficiência das representações vetoriais geradas por meio de modelos de redes neurais, Baroni et al. [56] realizaram um estudo comparando os vetores gerados por meio da predição (*context-predicting vectors* ou *word embeddings*) com os modelos tradicionais baseados em contagem de co-ocorrências (*context-counting vectors*) em diferentes tarefas de semântica lexical. Os autores concluíram que as *word embeddings* são modelos de semântica distribucional superiores, obtendo resultados excelentes mesmo com apenas poucas variações dos parâmetros de treinamento.

Nesse contexto, Levy and Goldberg [55] propõem o uso de um novo método para definir o contexto na geração de *word embeddings*. Enquanto o método proposto por Mikolov et al. [58] usa janelas de palavras para definir o contexto de palavras, Levy and Goldberg [55] propõem o uso de informações de dependência sintática entre as palavras. Os autores mostram que os contextos obtidos pelo método original dão enfase a similaridades topicais, enquanto que os contextos gerados com base em dependências sintáticas são mais indicados para identificar similaridades funcionais entre as palavras. Outros trabalhos fazem uso de outros tipos de anotações no texto de entrada para gerar as *word embeddings*, como *supersenses* [59] e sentidos das palavras [60].

Nessa linha, Camacho-Collados et al. [57] apresentam a abordagem NASARI (*Novel Approach to a Semantically-Aware Representation of Items*), que inclui a variação NASARI *embedded* para construção de *embeddings* de sentidos de palavras em um mesmo espaço vetorial de *word embeddings* pré-treinadas. Em avaliações experimentais realizadas pelos autores, a representação NASARI *embedded* apresentou bons resultados, tanto para identificação de similaridade entre palavras quanto para o agrupamento de sentidos. Camacho-Collados et al. [57] disponibilizaram a representação vetorial para milhões de conceitos e entidades nomeadas pertencentes à base BabelNet [61]. Por serem representações ligadas aos *synsets* da BabelNet, que estão disponíveis em vários idiomas, apesar de terem sido construídos a partir de textos no idioma inglês, os vetores NASARI *embedded* são independentes de idioma.

Considerando a representação de parágrafos ou documentos completos, Le and Mikolov [62] propuseram uma abordagem chamada *Paragraph Vector*, que posterior-

mente ficou conhecida como Doc2Vec. Essa abordagem foi inspirada nos trabalhos de construção de representações vetoriais de palavras por meio de redes neurais. Na abordagem Doc2Vec, um rede neural de uma camada oculta é treinada para prever o conteúdo do documento. Após o treinamento, os pesos da rede são extraídos e usados como *embeddings* para representar os documentos. Os resultados obtidos por Le and Mikolov [62] são competitivos com outros métodos estado da arte.

Motivado pela cobertura e propriedade multilíngue dos vetores NASARI *embedded*, Sinoara et al. [63] propõem modelos de representação de documentos baseados em *embeddings* de palavras e sentidos pré-treinadas. Em contraste com a abordagem Doc2Vec, a construção dos modelos propostos não requer uma grande quantidade de dados para o aprendizado do modelo de representação. Essa é uma vantagem da proposta deste trabalho, pois a quantidade de documentos disponível para o treinamento de *embeddings* pode ser um fator crítico, especialmente para pequenas coleções de documentos. Por outro lado, atualmente existem *embeddings* de palavras e sentidos pré-treinadas e de boa qualidade.

Mais recentemente, estudos em aprendizado de representação para dados textuais têm investigado *word embeddings* contextuais ou *word embeddings* dinâmicas [64]. Diferente das estratégias discutidas anteriormente, que são *word embeddings* estáticas, esses estudos visam computar a representação de *embedding* de palavras de acordo com sua ocorrência em uma sentença [65]. Um exemplo é o modelo BERT (*Bidirectional Encoder Representations from Transformers*) [66], desenvolvido por pesquisadores do Google, que é inicialmente pré-treinado em grandes coleções textuais de forma não supervisionada. Esse treinamento obtém um modelo neural da linguagem, representando textos por meio de suas características semânticas e sintáticas. Uma característica importante desse modelo é a possibilidade de refinamento do modelo na aplicação final a partir de um conjunto de dados rotulados, também chamado de *fine-tuning* [66], [67]. Do ponto de vista de enriquecimento semântico, essa é uma estratégia interessante pois permite aproveitar o pré-treinamento em coleções textuais de propósito geral (transferência de conhecimento) e refiná-lo considerando as características específicas do problema. Essa estratégia tem recebido destaque pois vários modelos pré-treinados estão sendo disponibilizados para a comunidade científica, considerando diferentes domínios de aplicação textual[1].

Na próxima seção são apresentadas aplicações do processo de Mineração de Textos que se beneficiam com o tratamento de aspectos semânticos.

VI. Aplicações da Mineração de Textos

APLICAÇÕES provenientes de um processo de Mineração de Textos se tornaram parte da rotina de uma sociedade digital. Aplicações como organização automática dos e-mails, recomendação de conteúdo, recuperação de informação, e análise de opiniões de consumidores são exploradas diariamente por empresas, consumidores, estudantes e pesquisadores. Essas aplicações também podem ser analisadas de acordo com os diferentes níveis de complexidade semântica do problema que visam resolver. Nesta seção, serão apresentadas três aplicações que cada vez mais exigem enriquecimento semântico para atender as expectativas de usuários: sistemas de recomendação, análise de sentimentos e mineração de eventos.

Aplicações baseadas em **sistemas de recomendação** são investigadas em cenários em que a quantidade de dados e informações disponíveis na *web* é abundante, por exemplo, com grande variedade de produtos, serviços, notícias e filmes disponíveis. Essa variedade aumenta a dificuldade de consumidores em identificar e escolher os itens que melhor atendam às suas necessidades. De forma geral, o objetivo do sistema de recomendação é filtrar a informação de forma inteligente e fornecer sugestões de itens que atendam a expectativa dos usuários [68].

A recomendação pode ser realizada por meio de diferentes estratégias. Tradicionalmente, a tarefa de recomendação consiste em identificar os relacionamentos existentes entre os usuários e os itens do sistema, considerando o histórico das interações passadas. Entretanto, além dessas informações, existem inúmeros outros fatores que também podem influenciar a preferência de um usuário, como o contexto no qual ele está inserido [69]–[71]. Pesquisadores começaram a perceber que a qualidade das recomendações aumenta quando essas informações adicionais, como tempo, local e outros, são utilizadas [71]–[73]. Dessa forma, a integração dessa informação contextual nos sistemas de recomendação tornou-se um tópico de crescente importância em pesquisas.

Nesse cenário, um sistema de recomendação sensível ao contexto é uma tecnologia de filtragem de informação que, além do comportamento e do interesse do usuário, utiliza também informação contextual para recomendar itens que lhe são de interesse. Pesquisas recentes indicaram que o uso do contexto em sistemas de recomendação pode aumentar a confiança do usuário e fornecer recomendações mais precisas, superando as abordagens tradicionais [7], [70], [74], [75]. A tendência nessa área é a utilização de novos tipos de informações com o objetivo de gerar recomendações mais personalizadas, precisas e relevantes aos usuários.

É nesse sentido que o enriquecimento semântico se torna relevante para sistemas de recomendação. Parte relevante das informações utilizadas para determinar o contexto da recomendação pode ser extraída de dados textuais. Exemplos típicos são *posts* em redes sociais e *reviews* sobre produtos e serviços. Tal conhecimento é útil para determinar o comportamento do usuário para inferir, por exemplo, se há maior probabilidade em consumir um determinado tipo de conteúdo naquele momento, ou até mesmo usar a representação semanticamente enriquecida para explicar motivos de uma determinada recomendação. Esse último exemplo tem recebido cada vez mais atenção atualmente, dada a exigência crescente de apresentar transparência e

[1] https://huggingface.co/models

interpretabilidade às ferramentas usadas diariamente pela sociedade [76].

Já a **análise de sentimentos**, ou mineração de opiniões, envolve aplicações para processar as opiniões das pessoas, sentimentos, avaliações, atitudes, e emoções relacionadas a entidades como produtos, serviços, organizações, indivíduos, assuntos, eventos, tópicos, e seus atributos [77]. Aplicações relacionadas à análise de sentimentos visam inferir a atitude do autor da opinião em relação a um determinado objeto, sendo que essa opinião é expressada de forma textual, em determinado tempo e contexto [42], [78].

Diferentes níveis de complexidade semântica podem ser explorados na análise de sentimento. Em sua forma mais simples, a análise de sentimentos é realizada em nível do documento, no qual a tarefa é classificar se todo o documento de opinião expressa um sentimento. De forma similar, o sentimento pode ser analisado em nível da sentença. Por fim, o sentimento pode ser analisado no nível da entidade e do aspecto, também conhecido como nível da característica, no qual é analisado o sentimento associado a cada aspecto da entidade. Esse último é atualmente mais investigado nas aplicações, e envolve diferentes etapas e algoritmos, como extração e categorização das entidades da uma opinião, extração e categorização dos aspectos, identificação do titular da opinião, extração e padronização do tempo e a classificação dos sentimentos dos aspectos.

Um outro ponto que afeta a complexidade semântica da análise de sentimentos está relacionado às fontes de informação. Existem diversas fontes de informação das quais é possível extrair as opiniões dos usuários, como redes sociais, *sites* especializados em analisar produtos de um segmento e os próprios *sites* de comércio eletrônico. Cada uma das fontes de informação possui suas peculiaridades, incluindo a estrutura da linguagem utilizada (formal e informal) e o conteúdo das opiniões. Por exemplo, pode-se ter, em um mesmo *site*, avaliações nas quais o produto foi analisado de forma completa e outras avaliações em que somente alguns de seus aspectos foram analisados.

A análise de sentimentos utilizando fontes de dados heterogêneas é um desafio de pesquisa em aberto, sendo preciso definir uma estrutura para representar esses dados. Nesse quesito, o enriquecimento semântico de textos de opinião tem sido objeto de estudo [10], [79]. Em particular, representações baseadas em redes para relacionar diferentes tipos de informação de maneira unificada, combinando informação de várias fontes, têm obtido resultados promissores [10]. Também vale destacar os modelos de *word embedding* para análise de sentimentos, que são particularmente úteis para descobrir similaridade entre aspectos de produtos e serviços, bem como relacionar as diferentes formas negativas e positivas de se expressar [79].

Por fim, um terceiro exemplo de aplicação é a **mineração de eventos** [80], [81]. Essa aplicação surgiu inicialmente da necessidade de sensoriamento de eventos extraídos de notícias e redes sociais para enriquecer modelos preditivos, por exemplo, em aplicações financeiras [82], [83], na análise de impactos da propagação de epidemias e desastres naturais [13], [84], bem como diferentes estudos sociais [85], [86].

A representação computacional de um evento é um desafio em particular [80]. Para representar adequadamente um evento, são considerados atributos que descrevem seus componentes, sendo eles dos tipos temporal (quando aconteceu), textual (o que aconteceu), pessoas e organizações (quem está envolvido) e geográfico (onde aconteceu) [87]. Dessa maneira, o pré-processamento de um evento é realizado de modo a obter, na medida do possível, atributos para melhor descrição desses componentes [80]. O enriquecimento semântico para identificar tais componentes também é crucial para diversas aplicações, principalmente com uso de modelos de linguagem (e.g. *word embeddings*) e entidades nomeadas obtidas com processamento de linguagem natural.

Em muitas situações estão disponíveis apenas o texto das notícias ou *posts* em redes sociais. Desse modo, o enriquecimento semântico da informação é mais complexo, pois além de extrair informação textual, temporal e geográfica (e.g. entidades nomeadas), há o interesse em vincular essas informações em bases de conhecimento externas, como o *Wikipedia*, em um processo denominado *wikification* [88]. Pesquisas mais recentes avançam a mineração de eventos para aplicações que geram predições sobre eventos futuros. Nesse caso, há interesse em identificar outras componentes como o porquê de um evento estar relacionado com outro evento [89], tarefa recente e desafiadora que exige enriquecimento semântico da representação para incorporar conhecimento de especialistas de domínio na aplicação [80].

VII. Considerações Finais

NO cenário atual, percebe-se um aumento no uso de métodos para extração automática de conhecimento de dados textuais, para as mais diversas aplicações. Visto a grande variedade de fontes, domínios, tipos de documentos e objetivos, em muitas dessas aplicações o tratamento adequado de aspectos semânticos pode ser crucial para o sucesso da mineração.

Neste artigo apresentou-se o processo de Mineração de Textos e discutiram-se os desafios inerentes a esse tipo de dados não estruturados. Textos escritos em língua natural apresentam diversas relações semânticas e fenômenos que podem alterar o seu significado composicional, que devem ser considerados durante o processo de mineração. Apesar das diversas pesquisas sendo desenvolvidas neste área, a incorporação da semântica na Mineração de Textos ainda é um desafio em aberto. Não há uma solução que seja adequada à qualquer problema e a seleção da abordagem a ser aplicada deve ser uma decisão criteriosa do analista, considerando as características dos textos disponíveis e dos resultados esperados para o processo de Mineração de Textos.

Agradecimentos

Os autores agradecem os auxílios fornecidos para o desenvolvimento deste trabalho, FAPESP (2019/25010-5

e 19/07665-4) e CNPq (426663/2018-7).

Referências

[1] R. Feldman and I. Dagan, "Knowledge discovery in textual databases (KDT)," in *KDD-95: The First International Conference on Knowledge Discovery and Data Mining*, 1995, pp. 112–117.

[2] A. M. Turing, "Computing machinery and intelligence," *Mind*, pp. 433–460, 1950.

[3] C. C. Aggarwal and C. Zhai, Eds., *Mining Text Data*. Springer, 2012.

[4] S. O. Rezende, Ed., *Sistemas Inteligentes: Fundamentos e Aplicações*. Editora Manole, 2003.

[5] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[6] C. C. Aggarwal and C. Zhai, "A survey of text clustering algorithms," in *Mining Text Data*, C. C. Aggarwal and C. Zhai, Eds., Springer, 2012, ch. 4, pp. 77–128.

[7] C. V. Sundermann, M. A. Domingues, R. A. Sinoara, R. M. Marcacini, and S. O. Rezende, "Using opinion mining in context-aware recommender systems: A systematic review," *Information*, vol. 10, no. 2, 2019, ISSN: 2078-2489. DOI: 10.3390/info10020042. [Online]. Available: http://www.mdpi.com/2078-2489/10/2/42.

[8] R. A. Monteiro, R. L. Santos, T. A. Pardo, T. A. de Almeida, E. E. Ruiz, and O. A. Vale, "Contributions to the study of fake news in portuguese: New corpus and automatic detection results," in *PROPOR 2018: Proceedings of the International Conference on Computational Processing of the Portuguese Language*, 2018, pp. 324–334.

[9] D. G. Vasques, P. S. Martins, and S. O. Rezende, "A semantic approach to uncovering implicit relationships in textual databases," in *CLEI 2018: Proceedings of the XLIV Conferência Latino-americana de Informática*, 2018, pp. 1–10.

[10] R. M. Marcacini, R. G. Rossi, I. P. Matsuno, and S. O. Rezende, "Cross-domain aspect extraction for sentiment analysis: A transductive learning approach," *Decision Support Systems*, 2018.

[11] I. P. Matsuno, R. G. Rossi, R. M. Marcacini, and S. O. Rezende, "Aspect-based sentiment analysis using semi-supervised learning in bipartite heterogeneous networks," *Journal of Information and Data Management*, vol. 7, no. 2, p. 141, 2016.

[12] C. C. Aggarwal and T. Abdelzaher, "Integrating sensors and social networks," in *Social Network Data Analytics*, C. C. Aggarwal, Ed., Springer, 2011, ch. 14, pp. 379–412.

[13] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *WWW'10: Proceedings of the 19th International Conference on World Wide Web*, 2010, pp. 851–860.

[14] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, pp. 1012–1015, Feb. 2009.

[15] M. da Silva Conrado, A. D. Felippo, T. A. S. Pardo, and S. O. Rezende, "A survey of automatic term extraction for brazilian portuguese," *Journal of the Brazilian Computer Society*, vol. 20, no. 1, p. 12, 2014.

[16] W. Zheng, L. An, and Z. Xu, "Dimensionality reduction by combining category information and latent semantic index for text categorization," *Journal of Information and Computational Science*, vol. 10, no. 8, pp. 2463–2469, 2013.

[17] B. M. Nogueira and S. O. Rezende, "Dois novos métodos para seleção não-supervisionada de atributos em mineração de textos," in *CLEI'09: Anais da XXXV Conferencia Latinoamericana de Informática*, 2009, pp. 1–10.

[18] R. A. Sinoara, "Aspectos semânticos na representação de textos para classificação automática," Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional, Ph.D. dissertation, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2018.

[19] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[20] G. Miner, J. Elder, T. Hill, R. Nisbet, D. Delen, and A. Fast, *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, 1st. Academic Press, 2012.

[21] R. P. Oliveira, "Semântica," in *Introdução à linguística: domínios e fronteira, volume 2*, Cortez, 2012, pp. 23–54.

[22] N. Riemer, *Introducing Semantics*, ser. Cambridge Introductions to Language and Linguistics. Cambridge University Press, 2010. DOI: 10.1017/CBO9780511808883.

[23] A. V. S. Pietroforte, "Semântica lexical," in *Introdução à Lingüística II: Princípios de Análise*, Editora Contexto, 2010, pp. 111–135.

[24] A. L. de Paula Müller and E. de Carvalho Viotti, "Semântica formal," in *Introdução à Lingüística II: Princípios de Análise*, Editora Contexto, 2010, pp. 137–159.

[25] D. O. F. do Amaral and R. Vieira, "Nerp-crf: Uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields," *Linguamática*, vol. 6, no. 1, pp. 41–49, 2014.

[26] R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history.," in *COLING 96: Proceedings of the 16th International Conference on Computational Linguistics*, vol. 96, 1996, pp. 466–471.

[27] E. R. Fonseca and J. L. G. Rosa, "An architecture for semantic role labeling on portuguese," in *PROPOR 2012: Proceedings of the 10th International*

Conference on Computational Processing of the Portuguese Language, Springer Berlin Heidelberg, 2012, pp. 204–209.

[28] M. Palmer, D. Gildea, and N. Xue, *Semantic Role Labeling*. Morgan & Caypool Publishers, 2010.

[29] F. A. A. Nóbrega and T. A. S. Pardo, "General purpose word sense disambiguation methods for nouns in portuguese," in *PROPOR 2014: Proceedings of 11th International Conference on Computational Processing of the Portuguese Language*, vol. 8775, Springer International Publishing, 2014, pp. 94–101.

[30] A. Moro, A. Raganato, and R. Navigli, "Entity linking meets word sense disambiguation: A unified approach," *Transactions of the Association for Computational Linguistics (TACL)*, vol. 2, pp. 231–244, 2014.

[31] E. Agirre and P. G. Edmonds, *Word sense disambiguation: Algorithms and applications*. Springer, 2007, vol. 33.

[32] C. Fellbaum, *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[33] R. Vieira, P. N. Gonçalves, and J. G. C. de Souza, "Processamento computacional de anáfora e correferência," *Revista de Estudos da Linguagem*, vol. 16, no. 1, 2008.

[34] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, "Semantic similarity from natural language and ontology analysis," *Synthesis Lectures on Human Language Technologies*, vol. 8, no. 1, pp. 1–254, 2015.

[35] M. T. Pilehvar and R. Navigli, "From senses to texts: An all-in-one graph-based approach for measuring semantic similarity," *Artificial Intelligence*, vol. 228, pp. 95–128, 2015.

[36] E. Fonseca, L. B. dos Santos, M. Criscuolo, and S. Aluísio, "Visão geral da avaliação de similaridade semântica e inferência textual," *Linguamática*, vol. 8, no. 2, pp. 3–13, 2016.

[37] E. Agirre, C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, R. Mihalcea, G. Rigau, L. Uria, and J. Wiebe, "Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability," in *SemEval 2015: Proceedings of the 9th international workshop on semantic evaluation*, 2015, pp. 252–263.

[38] M. D. Lee, B. Pincombe, and M. B. Welsh, "An empirical evaluation of models of text document similarity," in *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, 2005, pp. 1254–1259.

[39] R. A. Sinoara, R. B. Scheicher, and S. O. Rezende, "Evaluation of latent dirichlet allocation for document organization in different levels of semantic complexity," in *CIDM'17: Proceedings of the 2017 IEEE Symposium on Computational Intelligence and Data Mining*, 2017, pp. 2057–2064.

[40] R. G. Rossi, R. M. Marcacini, and S. O. Rezende, "Benchmarking text collections for classification and clustering tasks," Instituto de Ciências Matemáticas e Computação, Universidade de São Paulo, Tech. Rep. 395, 2013, Relatório Técnico 395, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

[41] R. A. Sinoara, J. Antunes, and S. O. Rezende, "Text mining and semantics: A systematic mapping study," *Journal of the Brazilian Computer Society*, vol. 23, no. 9, pp. 1–20, 2017.

[42] B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012.

[43] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining.," in *LREC 2010: Proceedings of the Seventh International Conference on Language Resources and Evaluation*, vol. 10, 2010, pp. 2200–2204.

[44] P. P. Balage Filho, T. A. S. Pardo, and S. M. Aluísio, "An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis," in *STIL 2013: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, 2013, pp. 215–219.

[45] R. B. Scheicher, R. A. Sinoara, J. C. Felinto, and S. O. Rezende, "Sentiment classification improvement using semantically enriched information," in *19th ACM Symposium on Document Engineering*, 2019, pp. 1–4, ISBN: 9781450368872. DOI: 10.1145/3342558.3345410.

[46] R. A. Sinoara, R. G. Rossi, and S. O. Rezende, "Semantic role-based representations in text classification," in *ICPR 2016: Proceedings of the 23rd International Conference on Pattern Recognition*, 2016, pp. 2314–2319.

[47] H.-j. Kim, K.-j. Hong, and J. Y. Chang, "Semantically enriching text representation model for document clustering," in *SAC '15: Proceedings of the 30th Annual ACM Symposium on Applied Computing*, Salamanca, Spain: ACM, 2015, pp. 922–925.

[48] R. A. Sinoara, C. V. Sundermann, R. M. Marcacini, M. A. Domingues, and S. O. Rezende, "Named entities as privileged information for hierarchical text clustering," in *IDEAS'14: Proceedings of the 18th International Database Engineering & Applications Symposium*, Porto, Portugal: ACM, 2014, pp. 57–66.

[49] G. Spanakis, G. Siolas, and A. Stafylopatis, "Exploiting wikipedia knowledge for conceptual hierarchical clustering of documents," *Computer Journal*, vol. 55, no. 3, pp. 299–312, 2012.

[50] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou, "Exploiting wikipedia as external knowledge for document clustering," in *KDD'09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 389–396.

[51] Y. Lu, Q. Mei, and C. Zhai, "Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA," *Information Retrieval*, vol. 14, no. 2, pp. 178–203, 2011.

[52] Z. Liu, M. Li, Y. Liu, and M. Ponraj, "Performance evaluation of latent dirichlet allocation in text mining," in *FSKD 2011: Proceedings of the Eighth International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 4, 2011, pp. 2695–2698.

[53] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[54] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *Journal of Artificial Intelligence Research*, vol. 37, no. 1, pp. 141–188, 2010.

[55] O. Levy and Y. Goldberg, "Dependency-based word embeddings," in *ACL 2014: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 302–308.

[56] M. Baroni, G. Dinu, and G. Kruszewski, "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors," in *ACL 2014: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 238–247.

[57] J. Camacho-Collados, M. T. Pilehvar, and R. Navigli, "NASARI: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities," *Artificial Intelligence*, vol. 240, pp. 36–64, 2016.

[58] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR Workshop*, 2013.

[59] L. Flekova and I. Gurevych, "Supersense embeddings: A unified model for supersense interpretation, prediction, and utilization.," in *ACL 2016: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 2029–2041.

[60] I. Iacobacci, M. T. Pilehvar, and R. Navigli, "Sensembed: Learning sense embeddings for word and relational similarity.," in *ACL 2015: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, 2015, pp. 95–105.

[61] R. Navigli and S. P. Ponzetto, "Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artificial Intelligence*, vol. 193, pp. 217–250, 2012.

[62] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *ICML-14: Proceedings of the 31st International Conference on Machine Learning*, 2014, pp. 1188–1196.

[63] R. A. Sinoara, J. Camacho-Collados, R. G. Rossi, R. Navigli, and S. O. Rezende, "Knowledge-enhanced document embeddings for text classification," *Knowledge-Based Systems*, vol. 163, pp. 955–971, 2019.

[64] S. Wang, W. Zhou, and C. Jiang, "A survey of word embeddings based on deep learning," *Computing*, vol. 102, no. 3, pp. 717–740, 2020.

[65] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.

[66] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[67] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8342–8360.

[68] C. C. Aggarwal, *Recommender Systems, The Textbook*, 1st ed. Springer, 2016.

[69] S. Raza and C. Ding, "Progress in context-aware recommender systems—an overview," *Computer Science Review*, vol. 31, pp. 84–97, 2019.

[70] M. A. Domingues, C. V. Sundermann, M. G. Manzato, R. M. Marcacini, and S. O. Rezende, "Exploiting text mining techniques for contextual recommendations," in *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, IEEE, vol. 2, 2014, pp. 210–217.

[71] G. Adomavicius and A. Tuzhilin, "Context-aware recommender systems," in *Recommender systems handbook*, Springer, 2011, pp. 217–253.

[72] N. Hariri, B. Mobasher, R. Burke, and Y. Zheng, "Context-aware recommendation based on review mining," in *Proceedings of the 9th Workshop on Intelligent Techniques for Web Personalization and Recommender Systems (ITWP 2011)*, 2011, p. 30.

[73] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin, "Incorporating contextual information in recommender systems using a multidimensional approach," *ACM Transactions on Information Systems (TOIS)*, vol. 23, no. 1, pp. 103–145, 2005.

[74] M. Gorgoglione, U. Panniello, and A. Tuzhilin, "The effect of context-aware recommendations on customer purchasing behavior and trust," in *Proceedings of the fifth ACM conference on Recommender systems*, ACM, 2011, pp. 85–92.

[75] A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver, "Multiverse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering," in *Proceedings of the fourth ACM conference on Recommender systems*, ACM, 2010, pp. 79–86.

[76] Y. Zhang and X. Chen, "Explainable recommendation: A survey and new perspectives," *Foundations and Trends in Information Retrieval*, vol. 14, no. 1, pp. 1–101, 2020.

[77] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 2015.

[78] C. C. Aggarwal, "Opinion mining and sentiment analysis," in *Machine Learning for Text*, Springer, 2018, pp. 413–434.

[79] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, e1253, 2018.

[80] X. Chen and Q. Li, "Event modeling and mining: A long journey toward explainable events," *The VLDB Journal*, vol. 29, no. 1, pp. 459–482, 2020.

[81] R. M. Marcacini, R. G. Rossi, B. M. Nogueira, L. V. Martins, E. A. Cherman, and S. O. Rezende, "Websensors analytics: Learning to sense the real world using web news events," in *Anais do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web: Workshops e Pôsteres*, 2017, pp. 169–173.

[82] R. M. Marcacini, J. C. Carnevali, and J. Domingos, "On combining websensors and dtw distance for knn time series forecasting," in *ICPR 2016: Proceedings of the 23rd International Conference on Pattern Recognition*, IEEE, 2016, pp. 2521–2525.

[83] L. S. Rodrigues, S. O. Rezende, M. F. Moura, and R. M. Marcacini, "Agribusiness time series forecasting using perceptually important events," in *2018 XLIV Latin American Computer Conference (CLEI)*, IEEE, 2018, pp. 268–277.

[84] K. Radinsky and E. Horvitz, "Mining the web to predict future events," in *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013, pp. 255–264.

[85] J. Hendler, N. Shadbolt, W. Hall, T. Berners-Lee, and D. Weitzner, "Web science: An interdisciplinary approach to understanding the web," *Communications of the ACM*, vol. 51, no. 7, pp. 60–69, 2008.

[86] H. Peng, J. Li, Y. Song, R. Yang, R. Ranjan, P. Yu, and L. He, "Streaming social event detection and evolution discovery in heterogeneous information networks," *ACM Transactions on Knowledge Discovery from Data*, 2021.

[87] B. N. dos Santos, R. G. Rossi, S. O. Rezende, and R. M. Marcacini, "A two-stage regularization framework for heterogeneous event networks," *Pattern Recognition Letters*, vol. 138, pp. 490–496, 2020.

[88] S. Jabeen, X. Gao, and P. Andreae, "Semantic association computation: A comprehensive survey," *Artificial Intelligence Review*, pp. 1–51, 2019.

[89] F. Hamborg, C. Breitinger, M. Schubotz, S. Lachnit, and B. Gipp, "Extraction of main event descriptors from news articles by answering the journalistic five w and one h questions," in *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, 2018, pp. 339–340.

Roberta A. Sinoara possui doutorado em Ciências da Computação e Matemática Computacional e atualmente é professora do Instituto Federal de São Paulo.

Ricardo M. Marcacini possui doutorado em Ciências da Computação e Matemática Computacional e atualmente é professor da Universidade de São Paulo.

Solange O. Rezende (autor correspondente) possui doutorado em Engenharia Mecânica pela Universidade de São Paulo e atualmente é professora da Universidade de São Paulo. E-mail: *solange@icmc.usp.br*.