

## INTERFACES ENTRE MINERAÇÃO DE TEXTOS E AS MÍDIAS SOCIAIS: EVIDÊNCIAS DE UMA ANÁLISE BIBLIOMÉTRICA

Elias Rocha Gonçalves Júnior (Faculdade Multivix Vitória) E-mail: [eliasrgjunior1@gmail.com](mailto:eliasrgjunior1@gmail.com)

Virgínia Siqueira Gonçalves (Universidade Candido Mendes) E-mail:

[virgíniasiqueiragoncalves@gmail.com](mailto:virgíniasiqueiragoncalves@gmail.com)

Geórgia Regina Rodrigues Gomes (Universidade Federal Fluminense) E-mail:

[georgiargomes@gmail.com](mailto:georgiargomes@gmail.com)

Elias Rocha Gonçalves (Instituto Superior de Educação Professor Aldo Muylaert) E-mail:

[falacomelias@gmail.com](mailto:falacomelias@gmail.com)

**Resumo:** Este trabalho tem como objetivo realizar uma análise bibliométrica abordando a aplicação de técnicas de mineração de textos em mídias sociais, sendo esta baseada em uma busca realizada na Base Scopus, tendo como palavra-chave de pesquisa “text mining AND social media”. O processo analítico teve como base o modelo proposto por Costa (2010), baseado na seleção de um conjunto de referências bibliográfica para servir como apoio o estudo. Os resultados prévios obtiveram 218 registros, sendo estes refinados para apenas publicações em periódicos. Ao final da análise completa dos artigos selecionados, foram considerados 10 trabalhos para compor o núcleo de partida para a pesquisa. Esta análise demonstrou e comprovou as aplicações efetivas e satisfatórias da utilização crescente da mineração de texto relacionadas ao tema mídia social, oportunizando uma discussão positiva da utilização da técnica de mineração de texto na avaliação dos clientes sobre determinados produtos nas comunidades on-line, relação cliente-mercado como influência para novos consumidores e novas vendas, melhoria em mecanismos de buscas de produtos de sites e detecção de relevância de eventos em redes sociais, além de serem utilizadas para pesquisas de intenção de votos em eleições por meio de opiniões on-line e na área da saúde.

**Palavras-chave:** Mineração de textos, mídias sociais, estudo bibliométrico.

## INTERFACES BETWEEN TEXT MINING AND SOCIAL MEDIA: EVIDENCE FROM A BIBLIOMETRIC ANALYSIS

**Abstract:** This paper aims to perform a bibliometric analysis addressing the application of text mining techniques in social media, based on a search carried out in the Scopus Base, having as a search keyword "text mining AND social media". The analytical process was based on the model proposed by Costa (2010), based on the selection of a set of bibliographical references to serve as support for the study. The previous results obtained 218 records, these being refined to only publications in periodicals. At the end of the complete analysis of the selected articles, 10 papers were considered to compose the starting nucleus for the research. This analysis demonstrated and proved the effective and satisfactory applications of the growing use of text mining related to the social media theme, providing a positive discussion of the use of the text mining technique in the evaluation of customers about certain products in the online communities, customer relationship -market as an influence for new consumers and new sales, improvement in search engine products of websites and detection of relevance of events in social networks, as well as being used for polling purposes in elections through online reviews and in the health area.

**Keywords:** Text mining, social media, bibliometric study.

### 1. Introdução

O uso de mídias sociais tem crescido rapidamente, em 2011, quatro entre cinco americanos visitaram sites de mídia social e blogs (BARNES; LESCAULT, 2011). Números relevantes que mostraram o aumento da popularidade e a importância dos

meios de comunicação social. Além disso, a mídia social tornou-se uma mercadoria neutra, em termos de idade, utilizada tanto por homens e mulheres de todas as faixas etárias (STROUD, 2008).

Esse crescimento acelerado estimulou estudos e desenvolvimento de sistemas para avaliação de opiniões automaticamente e, conseqüentemente, extração de informações úteis em textos (HAN; KAMBER; PEI, 2012).

Devido ao aumento do volume de informações disponíveis e aos desafios envolvidos na coleta e análise dessa quantidade de dados e textos, as técnicas de mineração de texto têm sido cada vez mais utilizadas em pesquisas científicas, com o intuito de se acompanhar o que tem sido feito em um campo de pesquisa específico, especialmente no contexto da engenharia (MILANEZ *et al.*, 2013).

A mineração de textos pode ser entendida como o estudo e a prática de extrair padrões, regras e tendências a partir do texto completo de artigos científicos digitais, usando princípios da linguística computacional e métodos analíticos (SULLIVAN, 2011; FAHAZEE *et al.*, 2012).

Existem avanços em práticas metodológicas na mineração de textos não estruturados e ferramentas para realização de análises, como resumo de texto, extração de frase nominal e processos de linguagem natural, agrupamento, representação vetorial e mapeamento (WANNER *et al.*, 2008; YON *et al.*, 2013). Segundo Tseng *et al.* (2007), títulos e resumos são as principais fontes de texto não estruturado, embora todas as partes de documentos possam eventualmente ser exploradas.

Conforme Barion e Lago (2008), o processo de descoberta de conhecimento textual é composto das etapas de pré-processamento, com a finalidade de preparar, transformar, organizar e melhorar a qualidade do texto para a etapa subsequente, o processamento dos dados, que é o objetivo, no qual as técnicas aplicadas variam de acordo com a finalidade, recuperação da informação, indexação, extração da informação, associação de documentos, sumarização, clusterização e classificação/categorização. Feito isso, segue-se a etapa de pós-processamento, que avalia os resultados da mineração.

A partir deste cenário, este artigo tem por objetivo mapear as publicações científicas sobre mineração de textos relacionada a mídias sociais, visando a identificar as características dos trabalhos, como autores que contribuem para o tema e área de pesquisa, período das publicações, países onde o tema é estudado e os periódicos nos quais os trabalhos são publicados

Para isso, serão utilizadas técnicas bibliométricas, que empregam métodos quantitativos na busca por uma avaliação objetiva da produção científica, apontada como primeira contribuição deste artigo. A segunda contribuição é a identificação e análise das tarefas da mineração de textos, as quais os pesquisadores têm mais aplicado de forma prática.

## **2. Referencial Teórico**

Comumente apontadas como sendo manifestações particularmente contemporâneas, as mídias sociais são objetos de estudo relevantes que despontam de forma expressiva na Ciência da Informação, devido ao fato de atuarem como plataformas de representação da informação facilitada por tecnologias interativas e entrepostas por interfaces cada vez mais eficazes (JORENTE; SANTOS, 2014).

Portanto, as informações passam a ser recebidas em tempo real, podendo considerar a penetrabilidade e difusibilidade das tecnologias da informação um dos principais indicadores de desenvolvimento da sociedade, com relação a propagação da informação

(LIMA; FREIRE, 2014), para que seja possível desenvolver uma relação de compromisso com a cultura da comunicação (SOUSA; BORGES, 2019).

O fato de que as mídias sociais possibilitam adquirir conhecimento onde quer que esteja e a qualquer momento vem influenciando o comportamento do indivíduo na busca e no uso da informação, ou, na contemporaneidade, o que se denomina de comportamento infocomunicacional (COSTA *et al.*, 2016). Logo, esta forma de comunicação pode ser definida como uma ferramenta online que estimula a partilha de opiniões e a troca de informações, tal como aplicações baseadas na internet que possibilitam a produção e a circulação de conteúdo (FAORO; ABREU; DEMARCHI, 2017).

A mineração de texto ganhou importância com o crescimento da internet e o aumento no volume de suas informações e dos mecanismos de busca (PINHEIRO, 2008), sendo amplamente utilizada para extrair informações de grandes volumes de documentos textuais, principalmente devido a cerca de 80% do conteúdo presente na Internet estar na forma de texto (CHEN, 2001).

Para Aggarwal e Zhai (2012), analisar as informações para determinar padrões de informação é o principal objetivo da mineração de texto. O objetivo do acesso à informação é conectar as informações corretas com os usuários certos, no momento certo, com menos ênfase em processamento ou transformação de informações de texto, implícitas nas bases de dados (BARION; LAGO, 2008).

Segundo Bezerra e Guimarães (2014), o estabelecimento da Internet como meio de disponibilização de informações, na forma de textos não estruturados impede que se estabeleçam limites das aplicações da mineração de texto. Dentre as aplicações possíveis, existem aquelas que podem ser julgadas, de certa forma, como rotineiras, como por exemplo a sumarização, que consiste na criação de sumários a partir de vários textos do mesmo tópico, o aprendizado ativo, que permite a classificação automática de documentos e aprendizado a partir de dados não classificados (WEISS; INDURKHYA; ZHANG, 2015).

### **3. Material e Métodos**

A concepção da análise bibliométrica proposta por Costa (2010), tem por finalidade aplicar um modelo de mineração de fontes bibliográficas baseados em ferramentas de busca de dados bibliográficos on-line, com o objetivo de a seleção de um conjunto de referências bibliográfica para servir como apoio a um determinado estudo, integrando conceitos de bibliometria (bibliometrics), webmetria (webmetrics), informetria (informetrics) e mineração bibliográfica (biblioming), tendo por objetivo auxiliar na definição de um conjunto inicial de referências bibliográficas sobre o tópico de estudo, no caso as relações entre a mineração de texto e as mídias sociais.

Além disso, proporciona uma investigação entre as referências, o que resulta em um estudo preliminar contendo trabalhos que farão parte do “núcleo de partida”, garantindo ao pesquisador embasamento suficiente para o próprio estudo bibliográfico. O modelo é dividido em seis etapas (COSTA, 2010):

- Definição da amostra de pesquisa;
- Pesquisa da amostra, com as palavras-chave;
- Identificação dos periódicos com maior número de artigos publicados sobre o tema;
- Identificação dos autores com maior número de publicações;

- Levantamento da cronologia de produção, identificando os ciclos de maior produção;
- Seleção dos artigos para a composição do “núcleo de partida” para a pesquisa bibliográfica.

As buscas executadas na base de conhecimento SCOPUS, com base no modelo de Costa (2010) foram realizadas com intuito de encontrar trabalhos relacionados às redes sociais e suas relações, além disso, não foram utilizados filtros de exclusão a fim de evitar a redução dos resultados, optando-se por um retorno com a maior amplitude possível.

#### 4. Resultados e Discussão

Utilizando-se a chave de busca “text mining AND social media” foram encontrados 218 resultados para o período de 2009 a 2017, com destaque para os anos de 2015, 2016 e 2017, que representam mais da metade dos trabalhos analisados, conforme Figura 1.

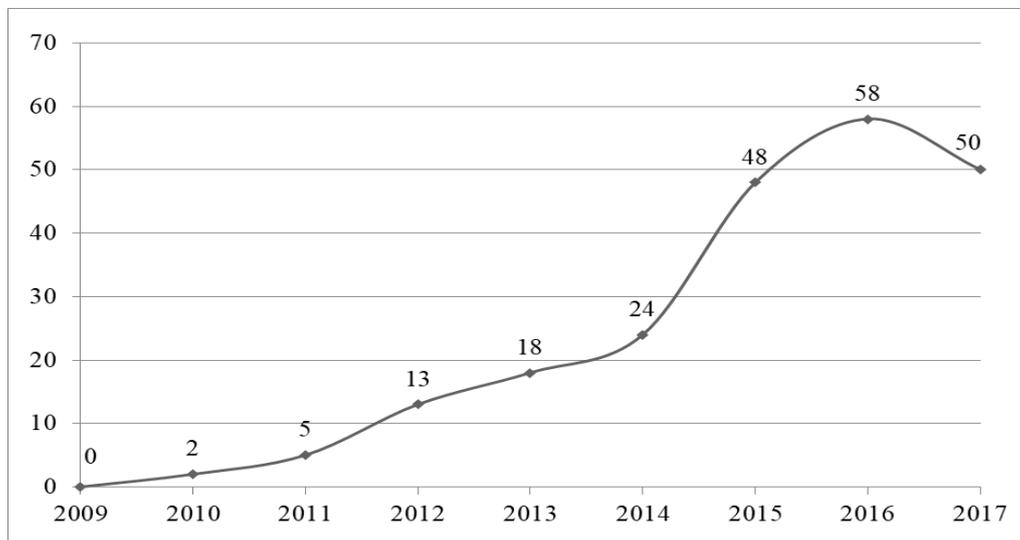


Figura 1 – Publicações identificadas na Base Scopus de 2009 a 2017.

Pode-se observar que as publicações relacionando os dois temas começaram no ano de 2010 e, a partir daí, veio sendo mais explorada, gerando um aumento o número de artigos relacionados, apesar de uma leve queda em 2017.

Além disso, destacaram-se os periódicos nos quais estes artigos foram publicados, autores com maior número de publicações e também os tópicos as quais a pesquisa foi relacionada, sendo que estas informações estão dispostas na Figura 2, 3 e 4, respectivamente.

Foram levados em consideração apenas aqueles periódicos e autores com mais de três registros, devido a considerável quantidade com apenas uma ou duas publicações relacionadas. Já para os tópicos de publicações, todos aqueles informados pela base científica foram relacionados, com especial destaque para o setor de ciências sociais, engenharia e ciência da computação.

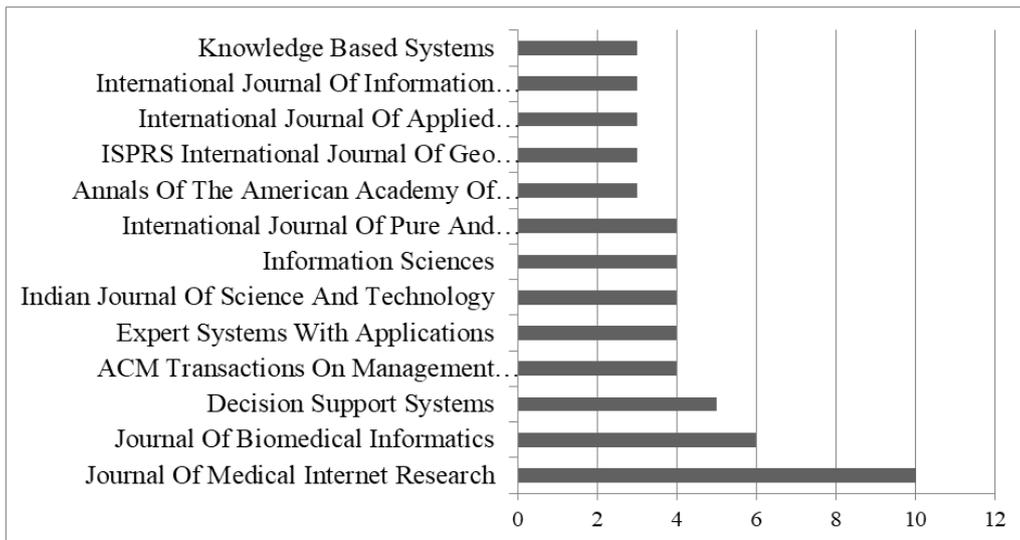


Figura 2 – Títulos dos periódicos dos registros encontrados de 2009 a 2017.

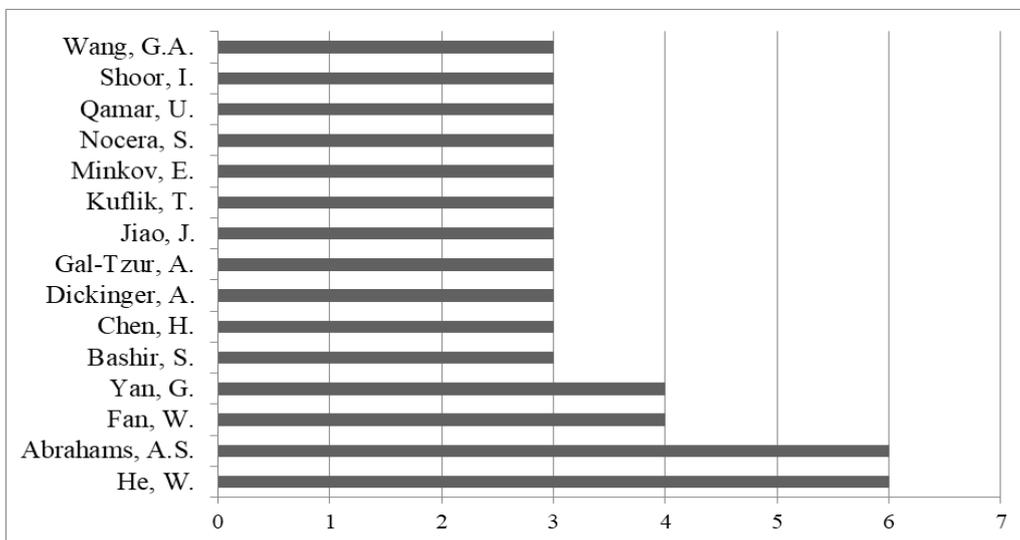


Figura 3 – Distribuição dos autores dos registros encontrados de 2009 a 2017.

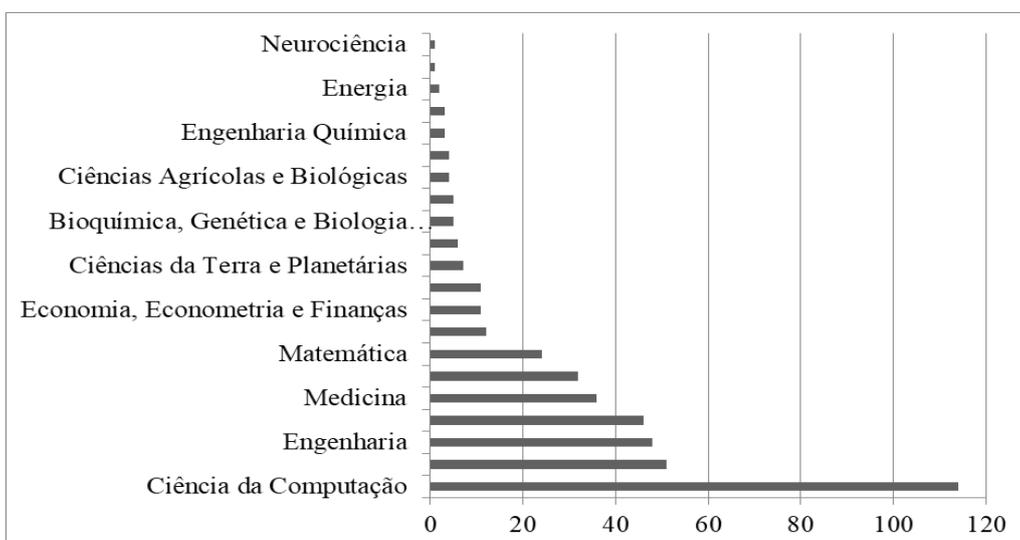


Figura 4 – Relação de tópicos dos registros encontrados de 2009 a 2017.

Por conseguinte, seguiu-se a análise dos países com maior quantidade de publicações, disposta no Quadro 1. Pode ser observado que não há a presença do Brasil na relação dos principais países com publicações relacionadas. O país possui apenas um artigo, sendo este no idioma Inglês e publicado no exterior em 2015.

Quadro 1 – Distribuição por país de origem dos registros encontrados de 2009 a 2017.

País	Publicações
Estados Unidos da América	91
China	28
Índia	23
Reino Unido	16
Itália	13
Austrália	12
Coréia do Sul	10
Áustria	6
Grécia	6
Hong Kong	6

Destacaram-se os trabalhos de maior relevância para o tema proposto com relação aos resultados apresentados anteriormente e também com base no número de citações dos mesmos. Selecionou-se um total de dez artigos para realização da análise, conforme o Quadro 2, por serem considerados os mais aderentes a aplicação das técnicas de mineração de texto e às redes sociais.

Quadro 2 – Relação de publicações mais relevantes dos registros encontrados de 2009 a 2017..

Publicação	Autor	Ano	Número de citações
<i>Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics</i>	Ghose, A.; Ipeirotis, P.G.	2011	324
<i>Social media brand community and consumer behavior: Quantifying the relative impact of user- and marketer-generated content</i>	Goh, K.-Y.; Heng, C.-S.; Lin, Z.	2013	210
<i>Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content</i>	Ghose, A.; Ipeirotis, P.G.; Li, B.	2012	129
<i>Sensing trending topics in twitter</i>	Aiello, L.M.; Petkos, G.; Martin, C.; Corney, D.; Papadopoulos, S.; Skraba, R.; Göker, A.; Kompatsiaris, I.; Jaimes, A.	2013	121
<i>Text and structural data mining of influenza mentions in web and social media</i>	Corley, C.D.; Cook, D.J.; Mikler, A.R.; Singh, K.P.	2010	101

---

<i>A survey of techniques for event detection in Twitter</i>	Atefeh, F.; Khreich, W.	2015	81
<i>Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France</i>	Ceron, A.; Curini, L.; Iacus, S.M.; Porro, G.	2014	81
<i>Vehicle defect discovery from social media</i>	Abrahams, A.S.; Jiao, J.; Wang, G.A.; Fan, W.	2012	63
<i>Marketing activity, blogging and sales "Popularity effect" in user-generated content: Evidence from online product reviews</i>	Onishi, H.; Manchanda, P. Goes, P.B.; Lin, M.; Yeung, C.-M.A.	2012 2014	60 51

---

Ghose e Ipeirotis (2011) é o primeiro estudo que integra técnicas econométricas, de mineração de texto e de modelagem preditiva para uma análise mais completa das informações capturadas pelas avaliações on-line geradas pelos usuários, a fim de estimar sua utilidade e impacto econômico. Foi avaliado o impacto das avaliações sobre os resultados econômicos, como as vendas de produtos e observaram como fatores diferentes afetam os resultados sociais, como a sua utilidade percebida pelos clientes. Foram explorados múltiplos aspectos do texto e de nível de revisão, como níveis de subjetividade, várias medidas de legibilidade e extensão de erros ortográficos para identificar importantes recursos baseados em texto, a utilidade média das revisões passadas e as medidas de identidade auto divulgada dos revisores que são exibidos ao lado de uma revisão. A análise econométrica realizada revela que a extensão da subjetividade, informatividade, legibilidade e correção linguística nas avaliações é um fator importante para influenciar as vendas e a utilidade percebida. Já quando utilizado o classificador Random Forest-based, foi identificado que é possível prever com precisão o impacto gerado pelas revisões dos clientes. Além disso, três principais categorias de recursos: recursos relacionados ao revisor, recursos de subjetividade e recursos de legibilidade, e identificou-se que o uso de qualquer um dos três conjuntos de recursos resulta em um desempenho estatisticamente equivalente, como no caso de usar todos os disponíveis.

Goh, Heng e Lin (2013) realizou um estudo integrando dados qualitativos de conteúdo de interação do comerciante-cliente de uma comunidade de usuários de uma marca no Facebook e dados de transações de consumo para montar um conjunto de dados exclusivo no nível de consumidor individual. Em seguida, quantificou-se o impacto dos conteúdos da comunidade dos consumidores (conteúdo gerado pelo usuário – CGU) e comerciantes (conteúdo gerado pelo comerciante – CGC) nas despesas de compras de vestuário dos consumidores. Um método de análise de conteúdo foi utilizado para construir medidas para capturar a natureza informativa e persuasiva da CGU e da CGC, ao mesmo tempo em que distinguia os modos de comunicação dirigidos e não direcionados na comunidade da marca. Na análise empírica, foram exploradas a adesão da comunidade e as diferenças temporais nas adesões para as estimativas de modelo e estratégias de identificação, além de considerar fatores relevantes, como preços, promoção, atributos de redes sociais e dados demográficos dos consumidores. Feito isso, foi observado que o envolvimento em comunidades de marcas de mídia social leva a um aumento positivo nas despesas de compra. Também foi identificado os diferentes papéis desempenhados pelo CGU e CGC, que variam pelo tipo de modos de comunicação direcionados ou não direcionados pelos consumidores e pelo comerciante.

Especificamente, as elasticidades da demanda em relação à informação obtidas CGU exibem um impacto maior do que a CGC no comportamento de compra do consumidor.

Ghose, Ipeirotis e Li (2012) ilustra como as mídias sociais podem ser incorporadas em um modelo de estimativa de demanda para gerar um novo sistema de classificação nos mecanismos de busca de produtos. Este propõe a geração de um sistema de classificação de produtos que ofereçam, em média, o melhor valor para o consumidor. A ideia-chave é que os produtos que fornecem um lucro maior devem estar mais alto na tela em resposta às consultas do consumidor. Foi utilizado um conjunto de dados exclusivo de reservas de hotéis nos EUA durante um período de três meses através do Travelocity, que foram complementados com dados de várias fontes de mídia social usando técnicas de mineração de texto, classificação de imagem, geotagging social, anotações humanas e mapeamento. Foi proposto um modelo estrutural híbrido de coeficiente aleatório, levando em consideração as duas fontes de heterogeneidade do consumidor, as diferentes ocasiões de viagem e as características do hotel apresentam. Com base nas estimativas do modelo, inferiu-se o impacto econômico de várias características de localização e serviço dos hotéis. Em seguida, apresentou-se um novo sistema de classificação de hotéis com base no benefício de utilização médio que um consumidor recebe de ficar em um determinado hotel, pois assim poderão ser oferecidos aos clientes os hotéis de “melhor valor” no início, destacando assim as estreitas ligações entre o comportamento do usuário em mídias sociais e motores de busca.

Aiello *et al.* (2013) ilustra que detecção de tópicos de tendência é um bloco de construção fundamental para monitorar e resumir informações provenientes de fontes sociais, havendo uma grande variedade de métodos e variáveis e eles afetam muito a qualidade dos resultados. Foram comparados seis métodos de detecção de tópicos em três conjuntos de dados do Twitter relacionados a eventos importantes, que diferem em sua escala de tempo e pelo tópico churn rate (taxa aceitável de cancelamentos). Observou-se a natureza do evento considerado, o volume de atividade ao longo do tempo, o processo de amostragem e o pré-processamento dos dados afetam a qualidade dos tópicos detectados, o que também depende do tipo de método de detecção utilizado. Identificou-se que as técnicas padrão de processamento de linguagem natural podem funcionar bem para fluxos sociais em tópicos muito focados, mas novas técnicas destinadas a explorar a distribuição temporal dos conceitos são necessárias para lidar com fluxos mais heterogêneos contendo várias histórias que evoluem em paralelo. Um dos novos métodos de detecção de tópicos que, com base em redes neurais e no ranking Tf-idf, estatisticamente calculado com base no quão importante e frequente é uma palavra em um documento de uma coleção, atinge consistentemente o melhor desempenho em todas as condições, sendo assim mais confiável do que outras técnicas de ponta.

Corley *et al.* (2010) utiliza técnicas de mineração de texto, por meio dados estruturais da web e redes sociais, para vigilância de doenças e para identificar comunidades on-line de comunicações específicas de saúde pública, com o intuito de assegurar ampla divulgação de informações. A partir da análise de links de comunidades, dados relacionados a gripe foram colhidos ao longo de um período de 24 semanas, de 5 de outubro de 2008 a 21 de março de 2009. A mineração de texto foi empregada para identificar tendências em mensagens sobre gripe que se correlacionam com dados do relatório de pacientes reais com a doença. Também foi aplicada uma técnica de mineração de dados baseada em gráficos para detectar anomalias/divergências com relação às informações de alguns blogs sobre gripe conectados pelo tipo de editor, links e tags dos usuários. Avaliou-se de forma abrangente as postagens de blog contendo

palavras-chave de tópicos de gripe através de texto, link e mineração de dados estruturais. Os resultados da análise mostram uma forte ocorrência de postagens de blogs sobre gripe durante a temporada de gripe dos EUA 2008-2009, ou seja, de 5 de outubro de 2008 a 21 de março de 2009, existe uma alta correlação entre a frequência de postagens, contendo palavras-chave de gripe, por semana e dados de vigilância de organizações relacionados. O algoritmo Girvan-Newman é alavancado para identificar clusters de sites similares como potenciais comunidades alvo para campanhas de informação de saúde on-line. A frequência de mensagens de gripe segue uma distribuição, que identifica, por meio de representação de métodos gráficos, que os blogueiros mais abrangentes não são os mais influentes. Além disso, a mineração de dados baseada em gráficos foi capaz de identificar anomalias significativas em blogs de gripe que não foram identificados através de análise de texto e podem ser marcados para uma investigação mais aprofundada por um analista.

Atefeh e Khreich (2015) elabora uma pesquisa de técnicas para a detecção de eventos a partir de fluxos de mensagens no site Twitter. Em contraste com a mídia convencional, a detecção de eventos de fluxos de Twitter traz novos desafios e contém grandes quantidades de mensagens sem sentido e conteúdo poluído, o que afeta negativamente o desempenho do processo de detecção. Além disso, segundo os mesmos, as técnicas tradicionais de mineração de texto não são adequadas, devido ao curto comprimento de tweets, ao grande número de erros ortográficos e gramaticais e ao uso frequente de linguagem informal e mista. As técnicas de detecção de eventos apresentadas abordam essas questões adaptando técnicas de vários campos para a singularidade do Twitter. Este artigo classifica essas técnicas de acordo com o tipo de evento, a tarefa de detecção e o método de detecção e discute recursos comumente usados. Finalmente, destaca a necessidade da existência de benchmarks públicos para avaliar o desempenho de diferentes abordagens de detecção e vários recursos.

Ceron *et al.* (2014) faz a aplicação de um método recentemente proposto por outros cientistas sociais a três cenários diferentes, analisando, por um lado, a popularidade em linha dos líderes políticos italianos ao longo de 2011 e, por outro lado, a intenção de votação dos usuários de internet franceses tanto na votação para presidente de 2012 e as eleições legislativas subsequentes. Embora os usuários da Internet não sejam necessariamente representativos de toda a população dos cidadãos de um país, a análise mostra uma habilidade notável para as redes sociais para prever os resultados eleitorais, bem como uma correlação notável entre as mídias sociais e os resultados dos levantamentos de massa tradicionais. Também verificou-se que a capacidade de previsão da análise das redes sociais se fortalece à medida que o número de cidadãos que expressam a opinião on-line aumenta, desde que os cidadãos atuem consistentemente sobre essas opiniões.

Abrahams *et al.* (2012) empregou técnicas de mineração de texto em uma meio de mídia social popular usado por entusiastas de veículos: fóruns de discussão on-line. Este afirma que a análise de sentimentos, uma técnica convencional para a detecção de queixas de consumidores, é insuficiente para encontrar, classificar e priorizar defeitos de veículos discutidos em fóruns on-line. Em contrapartida, foi descrito e avaliado um novo sistema de processo e decisão para identificação e priorização de defeitos automotivos, por meio de palavras específicas, retiradas de postagens das montadoras automotivas, para orientação das buscas e validação das mesmas. As descobertas fornecem informações administrativas sobre como a análise das mídias sociais pode melhorar o gerenciamento da qualidade automotiva, mostrando que o gerenciamento da qualidade do veículo pode ser suportado pela análise apropriada de postagens de mídias

sociais.

Onishi e Manchanda (2012) elaborou um conjunto de dados exclusivo do Japão que contém resultados de mercado (vendas) para novos produtos, novos meios de comunicação (blogs) e mídia tradicional (publicidade de TV) na categoria de filmes. Especificou-se um sistema logarítmico-linear de equações simultâneas para resultados de mercado e o volume de blogs. Os resultados sugerem que a mídia social e a tradicional atuam de forma sinérgica, sendo que a publicidade de pré-lançamento na TV promove a atividade de blogs, mas se torna menos efetiva durante o período pós-lançamento e que os resultados do mercado têm um efeito na quantidade de blogs. Os resultados encontrados têm suporte em uma análise de mineração de texto única e inovadora por meio de dados obtidos em uma pesquisa de opinião, buscando palavras relacionadas à satisfação dos clientes, sendo replicadas as descobertas para uma segunda categoria de produto, serviço de celular. Além disso, foram discutidas as implicações gerenciais práticas que os resultados encontrados podem causar.

Goes, Lin e Yeung (2014) tiveram como objeto de estudo o esforço para reunir mais avaliações dos clientes sobre produtos de diversos sites de opinião, que incentivam as interações dos usuários. Foram utilizados dados de um dos maiores sites de revisão de produtos nos quais os usuários podem se inscrever. Ao aplicarem-se os dados obtidos a um método de correspondência flexível, identificou-se que, à medida que os usuários se tornam mais populares, eles produzem mais críticas e avaliações mais objetivas. No entanto, suas classificações numéricas também mudam sistematicamente e se tornam mais negativas e mais variadas. Esse trade-off não foi previamente documentado e tem implicações importantes para a revisão do produto e outros sites de conteúdo gerados pelos usuários.

### **5. Considerações Finais**

Este trabalho apresentou uma pesquisa sobre a aplicação da tecnologia de mineração de texto em mídias sociais, com o objetivo de apresentar uma seleção de artigos como preâmbulo para uma análise de trabalhos mais relevantes referentes ao tema proposto.

O modelo proposto por Costa (2010) mostrou-se adequado, pois oportuniza ao pesquisador uma diretriz para iniciar-se uma pesquisa, sendo esta respaldada em trabalhos científicos com maior relevância e credibilidade, com resultados obtidos a partir de refinamentos nas buscas relacionadas, o que leva a identificação da área de pesquisa com maior número de artigos já publicados, maior número de autores com publicações relacionadas, os países com maior aplicação científica do tema e ainda podendo verificar e explorar a cronologia das produções.

O idioma inglês é aquele mais utilizado para publicação de trabalhos entre os diversos países relacionados na pesquisa, não havendo nenhum trabalho relacionado publicado em português. Isto pode ser identificado pelo número de publicações realizadas com origem nos Estados Unidos, com 91 no total, sendo este o país com mais produções científicas vinculadas ao tema.

A partir dos resultados da análise, foi possível observar que as técnicas de Mineração de Texto têm sido amplamente aplicadas à avaliação dos clientes sobre determinados produtos nas comunidades on-line, relação cliente-mercado como influência para novos consumidores e novas vendas, melhoria em mecanismos de buscas de produtos de sites e detecção de relevância de eventos em redes sociais, além de serem utilizadas para pesquisas de intenção de votos em eleições por meio de opiniões on-line e na área da saúde, com o monitoramento de informações veiculadas em blogs relacionados.

Também foi possível identificar a crescente discussão do tema mídia social com o passar dos anos.

### Referências

- ABRAHAMS, A.S.; JIAO, J.; WANG, G.A. & FAN, W. *Vehicle defect discovery from social media*. Decision Support Systems. Vol. 54, n.1, p.87-97, 2012.
- AGGARWAL, C.C. & ZHAI, C. *Mining text data*. 1. ed. Nova Iorque: Springer-Verlag, 2012. 524 p.
- AIELLO, L. M.; PETKOS, G.; MARTIN, C.; CORNEY, D.; PAPADOPOULOS, S.; SKRABA, R.; GOKER, A.; KOMPATSIARIS, I.; & JAIMES, A. *Sensing trending topics in twitter*. IEEE Transactions on Multimedia. Vol. 15, n.6, p.1268-1282, 2013.
- ATEFEH, F. & KHREICH, W. *A survey of techniques for event detection in Twitter*. Computational Intelligence. Vol. 31, n.1, p.133-164, 2015.
- BARION, E.C.N. & LAGO, D. *Mineração de textos*. Revista de Ciência Exatas e Tecnologias. Vol. 3, n.3, p.123-140, 2008.
- BARNES, N.G. & LESCAULT, A.M. *Social media adoption soars as higher-ed experiments and reevaluates its use of new communications tools*. Center for Marketing Research. University of Massachusetts Dartmouth, North Dartmouth, MA, 2011.
- BEZERRA, C.A. & GUIMARÃES, A.J.R. *Mineração de texto aplicada às publicações científicas sobre gestão do conhecimento no período de 2003 a 2012*. Perspectivas em Ciência da Informação. Vol.19, n.2, p.131-146, 2014.
- CERON, A.; CURINI, L.; IACUS, S.M. & PORRO, G. *Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France*. New Media and Society. Vol. 16, n.2, p.340-358, 2014.
- CHEN, H. *Knowledge management systems: a text mining perspective*. Tucson: Knowledge Computing Corporation, 2001. 64 p.
- CORLEY, C.D.; COOK, D.J.; MIKLER, A.R. & SINGH, K.P. *Text and structural data mining of influenza mentions in web and social media*. International Journal of Environmental Research and Public Health. Vol. 7, n.2, p.596-615, 2010.
- COSTA, H.G. *Modelo para webibliomining: proposta e caso de aplicação*. Revista FAE. Vol. 13, n.1, p.115-126, 2010.
- COSTA, L.F.; ANDRADE, R.L.V.; SILVA, A.C.P.; DUARTE, E.M. & SOUZA, A.C.P. *O uso de mídias sociais por revistas científicas da área da Ciência da Informação para ações de marketing digital*. Revista ACB: Biblioteconomia em Santa Catarina. Vol. 21, n.2, p.338-358, 2016.
- FAHAZEE, H.; AL-HADDAD, S.A.R.; ABDULLAH, R. & SAMSUDIN, K. *Text mining in bioinformatics: past, present and future*. In: International Conference on Information Retrieval and Knowledge Management, 2012, Kuala Lumpur. Anais do evento - p.327-330, 2012.
- FAORO, R.R.; ABREU, M.F. & DEMARCHI, M. *Redes Sociais como Ferramentas de Comunicação: uma síntese teórica*. Ciência da Informação em Revista. Vol. 4, n.3, p.25-39, 2017
- GHOSE, A. & IPEIROTIS, P.G. *Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics*. IEEE Transactions on Knowledge and Data Engineering. Vol. 23, n.10, p.1498-1512, 2011.
- GHOSE, A.; IPEIROTIS, P.G. & LI, B. *Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content*. Marketing Science. Vol. 31, n.3, p.493-520, 2012.
- GOES, P.B.; LIN, M. & YEUNG, C.-A. *"Popularity effect" in user-generated content: Evidence from online product reviews*. Information Systems Research. Vol. 25, n.2, p.222-238, 2014.
- GOH, K.; HENG, C. & LIN, Z. *Social media brand community and consumer behavior: Quantifying the relative impact of user- and marketer-generated content*. Information Systems Research. Vol. 24, n.1, p.88-107, 2013.
- HAN, J.; KAMBER, M. & PEI, J. *Data mining: concepts and techniques*. 3. ed. Waltham: Elsevier, 2012. 703p.

**JORENTE, M.J.V. & SANTOS, P.L.V.A.C.** *Mídias de informação e comunicação e Ciência da Informação*. Perspectivas em Ciência da Informação. Vol.19, n.1, p.190-206, 2014.

**LIMA, A.P.L. & FREIRE, I.M.** *As mídias sociais de olho na CI na perspectiva da disseminação da informação*. Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação. Vol. 19, n.39, p.113-132, 2014.

**MILANEZ, D.H.; AMARAL, R.M.; FARIA, L.I.L. & GREGOLIN, J.A.R.** *Assessing nanocellulose developments using science and technology indicators*. Materials Research. Vol. 16, n.9, p.635-641, 2013.

**ONISHI, H. & MANCHANDA, P.** *Marketing activity, blogging and sales*. International Journal of Research in Marketing. Vol. 29, n.3, p.221-234, 2012.

**PINHEIRO, C.A.R.** *Inteligência Analítica: mineração de dados e descoberta analítica*. 1. ed. Rio de Janeiro: Ciência Moderna, 2008. 414 p.

**SOUSA, G.R. & BORGES, E.M.** *Tecnologias de Informação e Comunicação (TIC) e formação de professores: investigando experiências pedagógicas mídia-educativas no curso de pedagogia a distância da Universidade Federal de Juiz de Fora (UFJF)*. Revista Tempos e Espaços em Educação. Vol. 12, n.28, p.199-216, 2019.

**STROUD, D.** *Social networking: An age-neutral commodity—social networking becomes a mature web application*. Journal of Direct, Data and Digital Marketing Practice. Vol. 9, n.3, p.278–292, 2008.

**SULLIVAN, D.** *Document warehousing and text mining*. New York: John Wiley & Sons, 2011.

**TSENG, Y.H.; LIN, C.J. & LIN, Y.I.** *Text mining techniques for patent analysis*. Information Processing & Management. Vol. 45, n.5, p.1216-1247, 2007.

**WANNER, L; BAEZA-YATES, R; BRÜGMANN, S; CODINA, J; DIALLO, B & ESCORSA, E.** *Towards content-oriented patent document processing*. World Patent Information. Vol. 30, n.1, p.21-33, 2008.

**WEISS, S. M.; INDURKHYA, N. & ZHANG, T.** *Fundamentals of Predictive Text Mining*. 2. ed. Londres: Springer-Verlag, 2015. 239p.

**YOON, J; PARK, H & KIM, K.** *Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis*. Scientometrics. Vol. 94, n.1, p.313-331, 2013.