UNIVERSIDADE ESTADUAL DO MARANHÃO

CENTRO DE CIÊNCIAS TECNOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO E SISTEMAS

MESTRADO PROFISSIONAL EM ENGENHARIA DE COMPUTAÇÃO E SISTEMAS

GUSTAVO NOGUEIRA DE SOUSA

ANÁLISE INTELIGENTE DE MÍDIAS SOCIAIS PARA POTENCIALIZAR GESTÃO DO RELACIONAMENTO COM CLIENTES

São Luís

2021

UNIVERSIDADE ESTADUAL DO MARANHÃO

CENTRO DE CIÊNCIAS TECNOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO E SISTEMAS

MESTRADO PROFISSIONAL EM ENGENHARIA DE COMPUTAÇÃO E SISTEMAS

GUSTAVO NOGUEIRA DE SOUSA

ANÁLISE INTELIGENTE DE MÍDIAS SOCIAIS PARA POTENCIALIZAR GESTÃO DO RELACIONAMENTO COM CLIENTES

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia da Computação. Centro de Ciências Tecnológicas. Universidade Estadual do Maranhão.
Orientador: Prof. Dr. Fábio Manoel França Lobato.
Coorientador: Prof. M.Sc. Antonio Fernando Lavareda Jacob Jr.

São Luís

2021

Sousa, Gustavo Nogueira de.
Análise inteligente de mídias sociais para potencializar gestão de relacionamento com clientes / Gustavo Nogueira de Sousa. – São Luís, 2021.

124 f.

Dissertação (Mestrado) – Curso de Engenharia de Computação e Sistemas, Universidade Estadual do Maranhão, 2021.

Orientador: Prof. Dr. Fábio Manoel França Lobato.

1.Mídias sociais. 2.Mineração de texto. 3.Aprendizagem de máquina. 4.Gestão de relacionamento com clientes. 5.Social CRM. I. Título.

CDU: 004.738.5:658.89

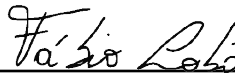
Elaborado por Giselle Frazão Tavares - CRB 13/665

Gustavo Nogueira de Sousa

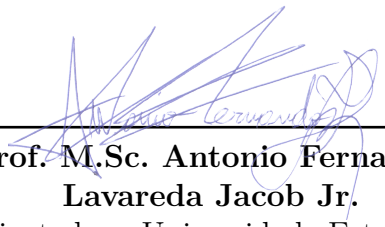
Análise inteligente de Mídias Sociais para potencializar gestão do relacionamento com clientes

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Computação e Sistemas da Universidade Estadual do Maranhão, como parte das exigências para a obtenção do título de Mestre em Engenharia de Computação e Sistemas.

Trabalho aprovado. São Luís, 09 de Setembro de 2021:

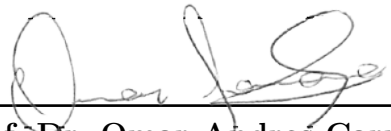


Prof. Dr. Fábio Manoel França Lobato
(Orientador - Universidade Federal do Oeste do Pará)



Prof. M.Sc. Antonio Fernando Lavareda Jacob Jr.
(Coorientador - Universidade Estadual do Maranhão)

Profª. Drª. Carla Bonato Marcolin
(Universidade Federal de Uberlândia)



Prof. Dr. Omar Andres Carmona Cortes
(Instituto Federal de Educação, Ciência e Tecnologia do Maranhão)

Aos meus pais, Jeconias e Josefa.

AGRADECIMENTOS

Agradeço a Deus por todas as bênçãos durante essa jornada, pela saúde, cuidado e por toda a força para superar os momentos mais difíceis.

Agradeço aos meus pais, Jeconias e Josefa, heróis que sempre me deram todo apoio, sustento e incentivo para a conclusão desta jornada. E também, agradeço às minhas irmãs, Késsia, Karine e Isabelly por todo apoio e incentivo.

Agradeço a minha amada esposa, Rebeca, por todo apoio, paciência e compreensão em todos os momentos difíceis desta caminhada.

Agradeço ao meu orientador Prof. Dr. Fábio Lobato, pelo companheirismo e oportunidades dadas desde a graduação, por todo conhecimento repassado, paciência, suporte, correções e incentivo para a conclusão deste projeto.

Agradeço ao meu coorientador Prof. Msc. Antonio Fernando Lavareda Jacob Jr., pela recepção na universidade, contribuição e suporte a este trabalho.

Agradeço ao SCRC (Social CRM Research Center), por todo suporte e auxílio na execução deste projeto.

Agradeço a UEMA, a todos os docentes, direção e administração por todo o suporte.

Agradeço aos meus irmãos na fé da Igreja Batista Nova Jerusalém (Santarém), em especial ao Pr. Donizete, Gregório Mateus e Jones. Que através de seus exemplos de vida, foram fundamentais para minha formação e para compreensão do real sentido de fazer o que faço.

RESUMO

A popularização das plataformas de mídias sociais tem transformado a relação entre empresas e consumidores. Por meio destas plataformas, consumidores e empresas podem interagir, colaborar, criar e compartilhar conteúdos de forma simples, rápida e barata. Esta produção interativa de conteúdos possui grande influência na tomada de decisões de compras por consumidores, pois cerca de dois terços deles checam as opiniões e avaliações de produtos e serviços antes de adquiri-los. Negligenciar este cenário pode impactar negativamente a operação diária de uma empresa, porém quando bem administrado, tal cenário representa uma rica fonte de conhecimento sobre clientes, produtos e serviços. Neste contexto, o Social CRM se apresenta como uma estratégia de negócios apoiada por processos e tecnologias que permitem a integração das mídias sociais a sistemas tradicionais de CRM. Nesse sentido, a utilização do Social CRM é essencial para a integração e utilização do conhecimento extraído em todos os processos e áreas operacionais da empresa, o que na prática permite reter e tornar os clientes mais satisfeitos. No entanto, a tarefa de extração de conhecimento a partir de conteúdos se estabelece como uma atividade não trivial, os desafios de utilizar grandes volumes de dados na forma não estruturada e de múltiplas fontes distintas, são exemplos de pontos que limitam a plena utilização por empresas. Diante disto, esta dissertação apresenta um compêndio de quatro artigos que analisam pontos de melhorias em sistemas Social CRM, por meio da otimização da análise da efetividade das comunicações empresariais em publicações no Facebook e da utilização de dados de plataformas de reclamações online para extração de conhecimentos. Estes estudos permitiram a exploração e a definição da melhor abordagem para de automação de análises de publicações de empresas, bem como a demonstração da viabilidade, possibilidades, restrições e aplicabilidade em um cenário real do uso de dados de reclamações para o aperfeiçoamento de estratégias de Social CRM.

Palavras-chaves: Mídias Sociais, Mineração de Texto, Aprendizagem de Máquina, Gestão de Relacionamento com Clientes, Social CRM.

ABSTRACT

The popularization of social media platforms has transformed the relationship between companies and consumers. Through these platforms, consumers and companies can interact, collaborate, create and share content in a simple, fast and cheap way. This interactive content production greatly influences purchasing decisions, as around two-thirds of consumers check the opinions and evaluations of products and services before purchasing them. Neglecting this phenomenon can negatively impact the daily operation of a company, but when managed well, this scenario represents a rich source of knowledge about customers, products, and services. In this context, Social CRM presents itself as a business strategy supported by processes and technologies that allow social media integration into traditional CRM systems. In this sense, the use of Social CRM is essential for integrating and using the knowledge extracted in all of the company's processes and operational areas, which in practice allows for customer loyalty. However, the task of extracting knowledge from content is established as a non-trivial activity; the challenges of using large volumes of data, unstructured and from multiple sources, are examples of points that limit the full use by companies. In view of this, this dissertation presents a compendium of four articles that analyze points of improvement in Social CRM systems by optimizing the analysis of the effectiveness of business communications in publications on Facebook and the use of data from online complaints platforms to extract knowledge. These studies allowed the exploration and definition of the best approach for automating the analysis of company publications, as well as demonstrating the feasibility, applicability, and restrictions of using claims data to improve Social CRM strategies.

Key-words: Social Media, Text Mining, Machine Learning, Customer Relationship Management, Social CRM.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANATEL	*Agência Nacional de Telecomunicações*
API	*Application Programming Interfaces*
ASL	*Average Sentence Length*
ASW	*Average number of Syllables per Word*
BERT	*Bidirectional Encoder Representations from Transformers*
BRASNAM	*Brazilian Workshop on Social Network Analysis and Mining*
CF	*Feedback do consumidor*
CGE	Conteúdos Gerados por Empresas
CGOV	*Consumidor.gov*
CNN	*Convolutional Neural Network*
CP	*Exposição do produto corrente*
CRISP-DM	*Cross Industry Standard Process for Data Mining*
CRM	*Customer Relationship Management*
CSV	*Comma-Separated-Values*
DAAD	*Deutscher Akademischer Austauschdienst*
ECD	*Engajamento em Conteúdo Digital*
eWoM	*Eletronic-Word-of-Mouth*
FAPEMA	*Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão*
FRES	*Flesch Readability Ease Score*
HEIs	*Higher Education Institutions*
HTC	*Hierarchical Text Classification*
IBGE	*Instituto Brasileiro de Geografia e Estatística*
ICRM	*International Workshop on Integrated Social CRM*

IES	*Instituições de Ensino Superior*
IT	*Informação e entretenimento*
INPI	Instituto Nacional da Propriedade Industrial
KNN	*K-Nearest Neighbors*
LDA	*Latent Dirichlet Allocation*
LSA	*Latent Semantic Analysis*
MJSP	*Ministry of Justice and Public Security*
NMF	*Non-Negative Matrix Factorization*
NP	*Anúncio de novo produto*
OB	*Marca da organização*
PIB	*Produto Interno Bruto*
PMI	*Pointwise Mutual Information*
PROCCE	*Pró-Reitoria da Cultura, Comunidade e Extensão*
PRS	*Publicidade em Redes Sociais*
RA	*ReclameAqui*
RF	*Random Forest*
RISTI	*Revista Ibérica de Sistemas e Tecnologias de Informação*
SA	*Vendas*
SCRC	*Social CRM Research Center*
Social CRM	*Social Customer Relationship Management*
SVM	*Support Vector Machine*
SW	*Sorteios e concursos*
TF-IDF	*Term Frequency-Inverse Document Frequency*
UEMA	*Universidade Estadual do Maranhão*
UFMG	*Universidade Federal de Minas Gerais*
UFOPA	*Universidade Federal do Oeste do Pará*

UFPA	*Universidade Federal do Pará*
UGC	*User-Generated Content*

SUMÁRIO

1 INTRODUÇÃO

A popularização e a difusão das plataformas de mídias sociais têm alterado e influenciado as ações e hábitos de indivíduos e empresas (BELLO-ORGAZ; JUNG; CAMACHO, 2016; SHIAU; DWIVEDI; LAI, 2018; BERTHON et al., 2012; KAPLAN; HAENLEIN, 2010). Este cenário tem transformado a forma como os consumidores e empresas se relacionam, uma vez que os consumidores passaram a compartilhar opiniões nas mídias sociais sobre os produtos e serviços. Esta relação entre consumidores e empresas é o objeto principal deste estudo, e neste capítulo serão apresentados temas importantes que nortearam a execução, tais como a contextualização, desafios, objetivos, contexto de execução e, por fim, a organização deste manuscrito.

1.1 CONTEXTUALIZAÇÃO E DESAFIOS

Em paralelo à expansão do acesso à internet no mundo, as mídias sociais têm ganhado cada vez mais destaque no cenário global por meio de um crescimento significativo e constante nos últimos anos (BELLO-ORGAZ; JUNG; CAMACHO, 2016; SHIAU; DWIVEDI; LAI, 2018). Nesse cenário de crescimento acelerado, diversos meios de interação entre as pessoas surgiram e se popularizaram, tais como blogs, *wiki*, serviços de microblog (e.g. Twitter[1]), sites de redes sociais (e.g. Facebook[2], LinkedIn[3]), entre outros tipos. É comum que diferentes plataformas foquem em estratégias distintas de interação entre os usuários, porém, em sua essência, a interação ocorre por meio do envio e compartilhamento de conteúdo textuais e/ou visuais (CARR; HAYES, 2015). Dentre as diversas finalidades e funcionalidades que as plataformas de mídias sociais oferecem, encontra-se também a finalidade pessoal de cada usuário na utilização dessas mídias, baseado em Whiting e Williams (2013) cada um dos tipos é definido a seguir:

- **Interação social** — Utilizar as mídias sociais para se comunicar e interagir com outros, através da conexão de pessoas com os mesmo interesses;
- **Busca de informações** — Utilizar as mídias sociais para se manter informado sobre os diversos temas;
- **"Passar o tempo"** — Utilizar as mídias sociais para ocupar o tempo livre e aliviar o tédio;

[1] https://twitter.com/
[2] https://facebook.com/
[3] https://www.linkedin.com/

- **Entretenimento** — Utilizar as mídias sociais para se entreter e se satisfazer emocionalmente;
- **Relaxar** — Utilizar as mídias sociais para aliviar o estresse do dia a dia;
- **Compartilhar informações** — Utilizar as mídias sociais como um facilitador de comunicação;
- **Por conveniência** — Utilizar as mídias sociais por considerá-las adequadas e de fácil utilização.

Em todos os diversos propósitos pessoais de utilização das mídias sociais, o conteúdo gerado por usuário, *do inglês - User Generated content* (UGC) está presente, seja no consumo ou na criação (FAASE; HELMS; SPRUIT, 2011). UGC pode ser definido como qualquer forma de conteúdo criado, divulgado e consumido por usuários (KIM; JOHNSON, 2016). Considerando a perspectiva de análise, alguns autores apresentam e utilizam dois subgrupos de UGC, os conteúdos gerados por consumidores e os conteúdos gerados por empresas (DUNN; HARNESS, 2019).

Os conteúdos gerados por consumidores em ambientes virtuais recebem a classificação de boca a boca virtual, do inglês, *Electronic Word of Mouth* (eWoM) e são definidos como opiniões, revisões e avaliações a respeito de marcas, empresas e serviços (SCHMÄH; WILKE; ROSSMANN, 2017). Na literatura, não existe um termo específico para definir o conteúdo gerado pela empresa em ambientes virtuais. Neste trabalho, tais postagens serão denominadas Conteúdos Gerados por Empresas (CGE), e são definidos como conteúdos gerados por empresas através da divulgação de produtos, sorteios, respostas a clientes e divulgação de informações relacionadas a companhia (GAVILANES; FLATTEN; BRETTEL, 2018).

Para os consumidores em geral, os conteúdos de eWoM e CGE podem representar uma rica fonte de informações para tomada de decisões de compra; já para empresas, o eWoM e as interações em CGE podem representar uma fonte de dados para análise e extração conhecimento sobre os seus consumidores e produtos (WANG; YU, 2015; LOBATO et al., 2017). Assim, quando bem utilizadas pelos consumidores, as mídias sociais auxiliam na tomada de decisões de compras e incrementam a possibilidade de satisfação com produtos e serviços adquiridos (WANG; YU, 2015). Já por parte das empresas, essas plataformas podem gerar vantagens competitivas por meio de melhorias de produtos, serviços e processos internos (KUBINA; LENDEL, 2015; CONSTANTINIDES; HOLLESCHOVSKY, 2016).

Diante disso, a gestão de relacionamento com clientes usando redes sociais, conhecido pelo seu acrônimo em inglês, o *Social Customer Relationship Management* (Social CRM) se destaca como uma ferramenta para a integração das mídias sociais aos processos

operacionais das empresas. O Social CRM é definido como uma estratégia de negócios apoiada por processos e tecnologias que permitem às empresas integrarem dados sociais às suas estratégias, processos e sistemas de CRM (ORENGA-ROGLÁ; CHALMETA, 2016; WITTWER; REINHOLD; ALT, 2017; REINHOLD; ALT, 2013). Além das tarefas realizadas nos CRMs tradicionais, no Social CRM há a integração e a utilização da análise de mídias sociais em cada uma destas tarefas e funcionalidades com o objetivo obter informações relevantes sobre produtos, clientes e tendências de mercado de modo a melhorar todos os processos relacionados (ALT; REINHOLD, 2012; REINHOLD; ALT, 2013; FAASE; HELMS; SPRUIT, 2011). A Figura 1 apresenta a integração entre os sistemas tradicionais de gestão de relacionamento com os clientes e as mídias sociais.

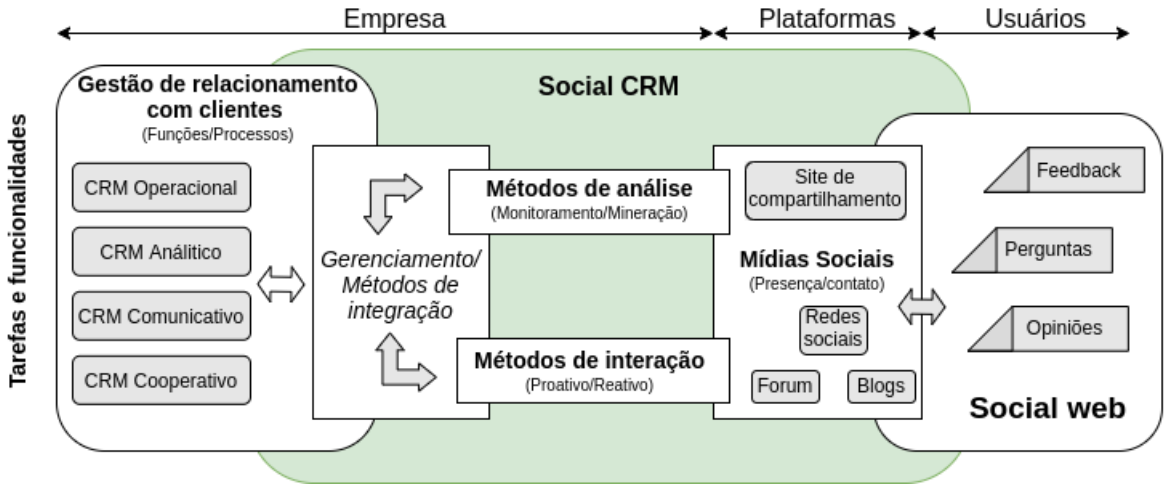


Figura 1 – Visão integrada das funcionalidades do Social CRM. Figura traduzida e adaptada de Reinhold e Alt (2013)

A extração de conhecimento de dados provenientes de mídias sociais é um processo muito importante na execução de Social CRM (MAROLT; ZIMMERMANN; PUCIHAR, 2018). Para que a integração dos dados ocorra de forma eficiente, as várias tarefas e áreas operacionais que compõem o Social CRM devem ser integradas (ALT; REINHOLD, 2012; REINHOLD; ALT, 2013). Destaca-se que processos inerentes à gestão de relacionamento com clientes poderão utilizar os resultados obtidos a partir de técnicas e métodos de monitoramento e análise do comportamento dos clientes, tal como apresentado nos trabalhos de (ORENGA-ROGLÁ; CHALMETA, 2016; LOBATO et al., 2017; ROSENBERGER, 2015; MARSHALL; MUECK; SHOCKLEY, 2015; WITTWER; REINHOLD; ALT, 2017). Tais tarefas e áreas listadas na Figura 1 encontram-se descritas na Tabela 1.

Tarefas/ áreas operacionais	Descrição e objetivos principais
Mídia Social	São serviços de mídia social, como fóruns, wikis, grupos para compartilhamento de informações e criação colaborativa.
Análises	Técnicas analíticas para análise e mineração, tal como filtragem, pesquisa, agregação, classificação e previsão.
Gerenciamento	Funcionalidades de gerenciamento, tal como moderação, gerenciamento de processos, gerenciamento de reputação, integração de dados, gerenciamento de privacidade e coordenação.
CRM	Processo de CRM e integração de dados e interligação com funções de CRM, como *lead*, contato, campanha ou gerenciamento de serviço.
Interação	Técnicas de interação, tal como entrega de conteúdo, desenvolvimento de diálogo, publicação, disseminação, recomendação e coordenação de interação.

Tabela 1 – Tarefas e áreas operacionais do Social CRM. Tabela traduzida e adaptada de Reinhold e Alt (2013).

Ao analisar de diferentes maneiras o uso de ferramentas de Social CRM na literatura relacionada, há *insights* importantes quanto às formas mais adequadas de utilização das mídias sociais exposta nos estudos de (ALMEIDA; LOBATO; CIRQUEIRA, 2017; RAMAN; MENON, 2018; COOPER; STAVROS; DOBELE, 2019; OLIVEIRA; CASAIS, 2018; SASHI; BRYNILDSEN; BILGIHAN, 2019; ZHANG; YANG, 2019). Por exemplo, percebe-se que uma implementação bem-sucedida tem o potencial de contribuir com o desenvolvimento de ações preventivas para mitigar falhas e ganhar mais competitividade por meio de clientes mais satisfeitos e de fácil retenção (ORENGA-ROGLÁ; CHALMETA, 2016).

Neste contexto, o Brasil se apresenta com um imenso potencial para as companhias utilizarem o Social CRM. Há no país cerca de 210 milhões de habitantes, sendo a região Sudeste com aproximadamente 42%, a região Sul com 14%, a região Centro-Oeste com 8%, a região Norte com 8% e a região Nordeste com 27% da população brasileira (IBGE, 2019a). Referente a telecomunicações, de acordo com dados da Agência Nacional de Telecomunicações (ANATEL) (2020), no mês de abril de 2020 o Brasil atingiu o número de 33 milhões de conexões a banda larga fixa, o que se traduz em uma densidade de acesso de 47,4 a cada 100 domicílios, e conta com cerca de 225,6 milhões de acessos à telefonia móvel, com uma densidade de acesso de 95,9 a cada 100 habitantes. Além disso, em 2019 aproximadamente 143,5 milhões de pessoas têm acesso à internet de alguma maneira, seja por *smartphone* (98.6%), computadores (46,2%), televisão(31,9%) ou *tablets* (10,9%) (IBGE, 2019b).

Em 2020, o Brasil cresceu em 6.4% o número de usuários de internet, os quais permanecem conectados diariamente por cerca de 10 horas e 8 minutos, divididas em 5 horas e 17 minutos na internet móvel e 4 horas e 43 minutos na internet do computador. Além disso, as plataformas de mídias sociais tem bastante relevância no país, com 70.3% da população presente em alguma plataforma neste período, representando um crescimento de 7,1% em relação ao ano anterior (Simon Kemp, 2021). Neste mesmo ano, um brasileiro gastou diariamente cerca 3 horas e 42 minutos nessas plataformas (a terceira posição global (KEMP, 2021)). O reflexo da popularização do acesso à internet também é visto no crescimento significativo do comércio eletrônico no Brasil, que somente em 2020 cresceu 41% em relação ao ano anterior e faturou cerca de 87 bilhões de reais. Ainda neste ano, este segmento cresceu cerca de 29% no número de clientes e passou a ter aproximadamente 79,7 milhões, os quais realizaram mais de 194 milhões de pedidos (EBIT, 2021).

Além das vantagens e oportunidades que o crescimento da internet e das mídias sociais proporciona para diversos segmentos do mercado, novos desafios são impostos e potencializados. Este cenário tem transformado a maneira como os consumidores tomam as decisões de compras, pois cerca de dois terços deles verificam as avaliações de produtos, serviços e marcas antes de decidirem adquiri-los (AHMAD; LAROCHE, 2017; CONSTANTINIDES; HOLLESCHOVSKY, 2016). Desta forma, a percepção de um possível consumidor tende a seguir a percepção majoritária de outros consumidores nos diversos meios disponíveis para a divulgação e acesso a conteúdo de eWoM, tais como *Twitter*[4], *Facebook* [5], *TripAdvisor*[6], *Booking.com*[7], *Reclame Aqui*[8], *Consumidor.gov*[9], entre outras (TIRUNILLAI; TELLIS, 2012; CONSTANTINIDES; HOLLESCHOVSKY, 2016).

Neste contexto, a tarefa de construção de conhecimento a partir de conteúdo de eWoM se estabelece como uma atividade não trivial, pois existem desafios que limitam sua plena utilização por empresas, com destaque para os seguintes desafios:

- **Múltiplas fontes** - Um cidadão brasileiro tem cerca de 9 perfis em diferentes plataformas de mídias sociais (GLOBALWEBINDEX, 2020). Desta forma, há a possibilidade de que o conteúdo de eWoM seja publicado em qualquer uma destas plataformas, o que requer a fusão e o tratamento de múltiplas fontes de dados distintas para a execução do processo de análise e extração de conhecimento (FARSEEV; CHUA, 2017; WANG et al., 2018);
- **Dados não-estruturados** - Estima-se que cerca de 80% dos dados na internet não são estruturados e, devido à natureza das mídias socais, os dados são essencialmente

[4] https://twitter.com/
[5] https://pt-br.facebook.com/
[6] https://www.tripadvisor.com.br/
[7] https://www.booking.com/
[8] https://www.reclameaqui.com.br/
[9] https://consumidor.gov.br/

não estruturados em diferentes tipos, tal como textos, imagens, vídeos *etc* (KUMAR; DABAS; HOODA, 2018; AGHASIAN; GARG; MONTGOMERY, 2020). Esta característica das mídias sociais torna os conjuntos de dados complexos e de difícil análise (CHEN; H.L.CHIANG; C. Storey, 2018; AGHASIAN; GARG; MONTGOMERY, 2020).

- **Dados Esparsos e Ruidosos** - Os dados provenientes das plataformas geralmente contêm incongruências e aspectos que não são úteis para as análises, por exemplo, em conteúdos textuais do português brasileiro é comum encontrar as palavras *kkkkk*, *rsrsrsrs*, *hahahaha* que indicam risos, porém, há tantas outras palavras que merecem um tratamento especial, tais como *vc*, *blz*, *tbm* e *tmj*, por exemplo (CIRQUEIRA et al., 2018). Assim, do tempo necessário para a extração de conhecimento, estima-se que 80% é utilizado somente na preparação dos dados para as análises (PRESS, 2016);
- **Grande quantidade de Dados** - Devido à inserção das plataformas de mídias sociais no cotidiano da sociedade, a quantidade de dados gerados diariamente é gigantesca, seja em fotos, textos, vídeos *etc*. Por exemplo, a cada minuto de 2019, cerca de 511.200 tweets foram publicados no *Twitter*, 55.140 fotos foram publicadas no *Instagram* e 92.340 *posts* foram publicados no *Tumblr* (DOMO, 2019). Assim, a grande quantidade de dados nessas plataformas é um dos principais desafios para a extração de conhecimento.

1.2 OBJETIVOS

1.2.1 OBJETIVO PRINCIPAL

Propor e avaliar diferentes abordagens para análises de dados de mídias sociais que auxiliem na melhoria de sistemas de Social CRM.

1.2.2 OBJETIVOS ESPECÍFICOS

1. Automatizar a classificação de conteúdo de marketing veiculados em mídias sociais em categorias de acordo com o modelo proposto por Gavilanes, Flatten e Brettel (2018), bem como correlacionar as categorias de conteúdo com métricas/padrão de engajamento. Exposto no Capítulo 3;
2. Analisar o potencial uso de dados de reclamações para melhorar sistemas de Social CRM. Exposto no Capítulo 4;
3. Identificar oportunidades e desafios que plataformas de reclamação online proveem para aperfeiçoamento de gestão de relacionamento com o cliente. Exposto no Capítulo 5;

4. Demonstrar a aplicabilidade da análise de reclamações publicadas nas mídias sociais em um cenário real de uso do Social CRM. Exposto no Capítulo 6;

5. Mapear o estado da arte e o estado da prática em relação à análise de mídias sociais, para identificar as bases, métodos e ferramentas mais utilizadas pelos pesquisadores em suas análises. Exposto no Apêndice A;

6. Analisar o uso de Social CRM por Micro e pequenas empresas, a fim de identificar oportunidades de intervenção. Exposto no Apêndice B.

No Capitulo 2, serão apresentados os pontos principais dos estudos que compõe o cerne desta dissertação, associados aos objetivos específicos 1, 2, 3 e 4. Desta forma, os veículos de publicação, problemas, justificativas, objetivos, impactos, e as contribuições técnico-científicas serão destacadas. Além disso, os objetivos específicos 5 e 6 são considerados estudos marginais e, portanto, estão expostos os artigos completos nos apêndices deste trabalho.

1.3 METODOLOGIA

No desenvolvimento deste estudo, foi utilizada a metodologia de análise de dados denominada *Cross Industry Standard Process for Data Mining* (CRISP-DM) (CHINCHILLA; FERREIRA, 2016; ROLLINS, 2015; SCHAFER et al., 2019; WIRTH, 2000). Esta metodologia foi escolhida devido à sua maleabilidade e aplicabilidade em diversos cenários de análise de dados, que se adapta bem aos diversos tipos de dados disponíveis nas mídias sociais. Além disso, devido à natureza cíclica, o processo de mineração de dados pode não ser finalizado quando uma modelagem é realizada. Se for verificado que os resultados não são satisfatórios, o processo de análise pode ser reiniciado para que sejam feitas melhorias com base nas lições aprendidas em cada etapa (WIRTH, 2000).

O CRISP-DM possui um processo hierárquico através de seis etapas encadeadas que formam um ciclo de análise (CHINCHILLA; FERREIRA, 2016; ROLLINS, 2015; SCHAFER et al., 2019; WIRTH, 2000). Na Figura 2 contém a organização de cada etapa no fluxo de execução das análises, iniciando na fase de “Entendimento do Negócio”, seguida por “Entendimento dos dados”, “Preparação dos dados”, “Modelagem”, “Avaliação” (após a avaliação também é possível o retorno para a fase inicial) e “Entrega”. Nas próximas subseções, cada uma das etapas que compõem a metodologia será descrita.

1.3.1 ETAPA 1: ENTENDIMENTO DO NEGÓCIO

Nesta etapa inicial é a fase central de toda a execução das análises, pois todos os aspectos que envolvem o projeto serão analisados e compreendidos para que seja gerada uma boa definição do problema, dos objetivos e dos requisitos da solução que deve ser

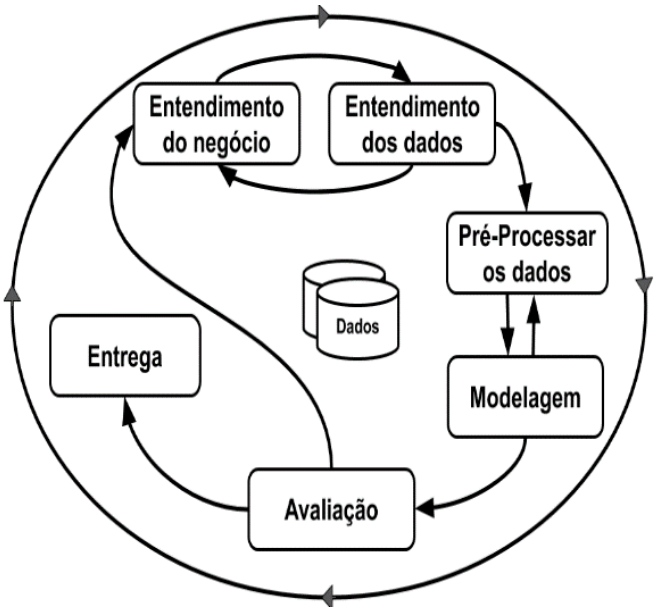


Figura 2 – Diagrama de funcionamento do modelo CRISP-DM (Adaptada de (WIRTH, 2000)).

implementada. Para que seja possível definir, com base no problema, objetivos e requisitos, um plano preliminar para a implementação da análise de dados (WIRTH, 2000; SCHAFER et al., 2019).

1.3.2 ETAPA 2: ENTENDIMENTO DOS DADOS

Esta etapa começa com a coleta e descrição inicial dos dados necessários à realização das análises propostas. A partir da conclusão da coleta, será realizada a exploração e verificação da qualidade dos dados coletados para familiarizar-se com esse conjunto de dados e identificar problemas de qualidade que inviabilizam a execução das análises. Além disso, a execução destas tarefas podem possibilitar a descoberta dos primeiros *insights* sobre os dados e novas possibilidades de análise (WIRTH, 2000; SCHAFER et al., 2019; CHINCHILLA; FERREIRA, 2016).

Segundo Wirth (2000), existe uma ligação estreita entre o "Entendimento do negócio" e o "Entendimento dos dados", pois a formulação do problema e o plano de execução do projeto requerem uma compreensão básica dos dados disponíveis.

1.3.3 ETAPA 3: PREPARAÇÃO DOS DADOS

Os dados serão preparados para análise de acordo com os requisitos dos parâmetros de entrada dos algoritmos de aprendizado de máquina que serão utilizados nas análises. Dependendo do plano de execução do projeto, nesta fase, os dados textuais podem ser pré-processados para remover todas as informações irrelevantes, diferentes conjuntos de dados podem ser fundidos e entre outras tarefas (WIRTH, 2000; ALLAHYARI et al., 2017;

NGUYEN et al., 2016; LI et al., 2019; CHANEY; BLEI, 2012; CHEN et al., 2019).

1.3.4 ETAPA 4: MODELAGEM

Com uma primeira versão dos dados já pré-processados e preparados, nesta etapa, o enfoque será no desenvolvimento de modelos descritivos e preditivos de acordo com os objetivos definidos (ROLLINS, 2015). O processo de modelagem é altamente iterativo, dada uma abordagem de análises definida, a seleção e avaliação de algoritmos são realizada e, à medida que os primeiro resultados são gerados, também é possível avaliar e ajustar os conjuntos de dados produzidos nas etapas anteriores (WIRTH, 2000).

1.3.5 ETAPA 5: AVALIAÇÃO

Nesta etapa, o modelo e os resultados serão avaliados para verificar sua qualidade e garantir que atende de forma adequada e completa aos objetivos definidos (WIRTH, 2000). Este processo permite identificar a qualidade e a eficácia do modelo através da utilização de métricas mais adequadas e de acordo com as especificidades de cada modelo. Caso seja verificado que o modelo e os resultados não atendam aos objetivos definidos, o fluxo retorna para o estágio de "Entendimento do negócio" para reiniciar o processo, e caso contrário o processo segue para a "Entrega" (ROLLINS, 2015).

1.3.6 ETAPA 6: ENTREGA

Em todos os casos de construção de modelos de análise de dados, é importante entender o que deve ser entregue no final do processo. Pois segundo Wirth (2000), normalmente os resultados gerados precisará ser organizado e apresentado de uma forma, e dependendo dos requisitos, poderá ser tão simples quanto gerar um relatório ou tão complexo quanto implementar um processo de mineração de dados repetível[10].

1.4 CONTEXTO DE EXECUÇÃO DO PROJETO

Este projeto é executado no contexto do grupo de estudo e pesquisa em computação aplicada[11] da Universidade Federal do Oeste do Pará - UFOPA, o qual conta com a cooperação da Universidade Federal do Pará (UFPA), Universidade Estadual do Maranhão (UEMA); e do Instituto de Sistemas de Informação da Universidade de Leipzig, Alemanha.

Este estudo também foi executado em parceria como o *Social CRM Research Center* (SCRC)[12]. O SCRC é um instituto de pesquisa sem fins lucrativos sediado em

[10] *Disclaimer: É importante destacar que os códigos implementados nos processos de análise são um artefato importante, pois farão parte de um programa de computador a ser registrado no INPI. Por esse motivo, eles não serão entregues a junto a este trabalho.*

[11] http://dgp.cnpq.br/dgp/espelhogrupo/312564

[12] https://scrc-leipzig.de/

Leipzig na Alemanha. Nele são estudados tópicos de CRM com o enfoque na integração com Social CRM. Além disso, conta por parcerias com algumas universidades brasileiras com o objetivo de prover treinamentos práticos para profissionais e estudantes sobre negócios e a Social CRM.

1.5 ORGANIZAÇÃO DO TRABALHO

No *Capítulo 2* serão apresentados os pontos principais de cada um dos trabalhos que compõem este compêndio, destacando os veículos de publicação, problemas e justificativa de cada trabalho, objetivos, impacto e contribuições técnico científicas. Os *Capítulo 3*, *Capítulo 4*, *Capítulo 5* e *Capítulo 6* conterão os manuscritos completos dos artigos que foram publicados em eventos e periódicos científicos. E por fim, no *Capítulo 7* estão as conclusões e considerações finais.

2 APRESENTAÇÃO DOS TRABALHOS

Neste capítulo os quatro trabalhos que compõe o cerne desta dissertação são apresentados.

2.1 TRABALHO "GERENCIAMENTO DE PUBLICIDADE NAS PLATAFORMAS DAS REDES SOCIAIS DE ACORDO COM CATEGORIAS DE CONTEÚDO"- *SODEBRAS (2019)*

Autores: Gustavo Nogueira de Sousa, Isabelle da Silva Guimarães, Antônio Fernando Lavareda Jacob Jr, Fábio Manoel França Lobato

Veículo de publicação: *Publicado em Revista Sodebras, Volume 14 (2019)*[1]

Texto completo: O manuscrito publicado deste trabalho está no capítulo 3 desta dissertação.

Problema e Justificativa: É notável que a análise de conteúdo de eWoM tem o potencial de gerar importantes ganhos e vantagens competitivas. Saber a efetividade da comunicação empresarial junto aos consumidores é muito importante, pois desta forma, é possível medir o impacto das soluções implementadas com base na análise de conteúdo. No entanto, a quantidade de dados presentes nessas plataformas torna o processo de avaliação manual dos resultados das publicações de negócios muito caro e ineficiente. Diante disto, este trabalho tem como objeto de estudo o crescimento da utilização de mídias sociais como meio para o relacionamento entre empresas e consumidores, em que avalia formas de otimização do processo de avaliação da efetividade, em termos de engajamento, de publicações empresariais.

Objetivo: O objetivo principal deste trabalho é automatizar o processo de mensuração da efetividade de publicações empresariais, por meio da classificação e correlação automática das categorias de conteúdos com métricas/padrões de engajamento, de acordo com o modelo proposto por Gavilanes, Flatten e Brettel (2018). Além disso, alguns objetivos secundários também estão implícitos no trabalho, tais como:

- Avaliar o desempenho de algoritmos de aprendizado de máquina em tarefas de classificação de conteúdos de mídias sociais;

- Eliminar o viés pessoal na classificação dos conteúdos das publicações;

[1] http://sodebras.com.br/

- Prover um meio para a classificação e avaliação da efetividade de grandes volumes de dados;

- Tornar o processo de avaliação da efetividade de publicações menos custoso para empresas.

Impacto e contribuições técnicos científicas: Este trabalho propõe e valida uma abordagem para as avaliações de publicações de empresas em mídias sociais. Os resultados são pertinentes, pois elimina o viés da classificação e avaliação manual dessas publicações, tornando o processo mais eficiente e menos oneroso, o que otimiza a tomada de decisões sobre a forma que está sendo conduzida o gerenciamento do relacionamento com os clientes. Além disto, neste trabalho foi necessário anotação manual de uma base de dados de treinamento com publicações empresariais categorizadas de acordo com as categorias de conteúdos proposta por Gavilanes, Flatten e Brettel (2018), esta base de dados foi tornada pública (SOUSA; JUNIOR; LOBATO, 2021) e pode ser usada como ponto de partida para a utilização em outras análises pela comunidade.

2.2 TRABALHO "ANÁLISE DO SETOR DE TELECOMUNICAÇÃO BRASILEIRO: UMA VISÃO SOBRE RECLAMAÇÕES"- *RISTI (2020)*

Autores: Gustavo Nogueira de Sousa, Isabelle da Silva Guimarães, Julio Augusto Nogueira Viana, Olaf Reinhold, Antonio Fernando Lavareda Jacob Junior, Fábio Manoel França Lobato.

Veículo de publicação: *Publicado em periódico Revista Ibérica de Sistemas e Tecnologia da Informação (RISTI) (2020)*[2]

Texto completo: O manuscrito publicado deste trabalho está no capítulo 4 desta dissertação.

Problema e Justificativa: Os conteúdos de eWoM representam uma rica fonte de informação para clientes, visto que cerca de dois terços dos consumidores verificam as avaliações de produtos e serviços antes de decidirem adquiri-los (AHMAD; LAROCHE, 2017; CONSTANTINIDES; HOLLESCHOVSKY, 2016). No entanto, pesquisas na literatura revelaram que há uma grande variedade de trabalhos que analisam dados de eWoM de diversas plataformas de mídias sociais (KIM; JOHNSON, 2016; VERMEER et al., 2019; MCILROY et al., 2016; VU et al., 2016), porém são poucos os trabalhos que usam conteúdo eWoM publicado em plataformas específicas para reclamações online. No Brasil, , as plataformas de reclamação online têm bastante relevância e influência, sendo que a principal plataforma do tipo está entre os 25 sites mais acessados do país. Diante disto,

[2] http://www.risti.xyz/

este trabalho tem como objeto de estudo a utilização dessas plataformas como fonte de dados para a extração de conhecimento.

Objetivo: O objetivo principal deste trabalho é o de analisar o potencial uso de dados de reclamações para melhorar os sistemas de Social CRM. Para isso, o objetivo principal está segmentado nos objetivos específicos a seguir:

- Identificar os principais termos presentes nas reclamações;
- Identificar como os principais tópicos estão relacionados entre si;
- Conhecer aspectos específicos das reclamações;
- Analisar a distribuição das reclamações considerando dimensões geo-temporais;
- Identificar quais as implicações práticas dos resultados das análises conduzidas para os negócios.

Impacto e contribuições técnicos científicas: Através das análises realizadas, este trabalho demonstra a análise de reclamações como uma forma viável de aprimorar os sistemas de Social CRM. A abordagem utiliza técnicas de aprendizado de máquina, que permitiram identificar e mensurar o grau de associação entre os principais temas nas reclamações dos consumidores. Além disso, neste trabalho iniciou-se a construção de um *pipeline* de análise com todos os métodos utilizados, com objetivo de reduzir o tempo requerido para as análises. Os resultados deste trabalho foram avaliados por um especialista em desenvolvimento de negócios do *Social CRM Research Center*[3], que atestou a viabilidade de uso para a melhoria do relacionamento com clientes.

2.3 TRABALHO "ANÁLISE COMPARATIVA DAS PRINCIPAIS PLATAFORMAS DE RECLAMAÇÕES ONLINE: IMPLICAÇÕES PARA ANÁLISE DE MÍDIA SOCIAL EM NEGÓCIOS"- *BRASNAM (2020)*

Autores: Gustavo Nogueira de Sousa; Isabelle Guimarães; Antonio F. L. Jacob Jr.; Fábio M. F. Lobato.

Veículo de publicação: *Publicado em anais do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM) (2020)*[4]

Texto completo: O manuscrito publicado deste trabalho está no capítulo 5 desta dissertação.

[3] https://scrc-leipzig.de/

[4] http://www2.sbc.org.br/csbc2020/ix-brazilian-workshop-on-social-network-analysis-and-mining/

Problema e Justificativa: Através da inserção das plataformas de mídias sociais no cotidiano da sociedade, os consumidores passaram a ser responsáveis tanto pelo consumo quanto pela produção de conteúdo sobre um produto/serviço nas mídias sociais (ROY; DATTA; MUKHERJEE, 2019). Devido a facilidade de propagação e de engajamento nessas plataformas, tais conteúdos têm a capacidade de influenciar a tomada de decisões de compra de muitos clientes (CONSTANTINIDES; HOLLESCHOVSKY, 2016). Por meio de pesquisas na literatura, percebeu-se que há uma escassez de estudos que envolvem plataformas específicas para reclamações online, portanto, este trabalho tem por objeto de estudo um comparativo entre duas das principais plataformas de reclamações online do Brasil.

Objetivo: O objetivo principal deste trabalho é identificar oportunidades e desafios que plataformas de reclamação proveem para aperfeiçoamento de gestão de relacionamento com o cliente. Para isso, alguns objetivos específicos também foram definidos, tal como:

- Determinar de forma objetiva a quantidade mais adequada de tópicos na modelagem;
- Analisar através da modelagem de tópicos as diferenças entre os conjuntos de reclamação de cada empresa nas plataformas;
- Analisar as diferenças dos grupos de consumidores nas duas plataformas.

Impacto e contribuições técnicos científicas: Este trabalho apresenta as possibilidades e restrições de análises de diferentes plataformas de reclamação online para a melhoria de sistemas de Social CRM. Desta forma, apresenta para o português brasileiro a aplicação de um método automático para determinar de forma objetiva a melhor quantidade de tópicos para modelagem, bem como as diferenças entre as duas principais plataformas e as diferenças entre os grupos de consumidores presentes nessas plataformas. Os métodos utilizados no processo de extração de conhecimento possibilitaram que melhorias e novas funcionalidades fossem adicionadas no *pipeline* de análise elaborado no artigo descrito na Seção 2.2.

2.4 TRABALHO "*GAINING INSIGHTS ON STUDENT SATISFACTION BY APPLYING SOCIAL CRM TECHNIQUES FOR HIGHER EDUCATION INSTITUTIONS*"- *ICRM (2021)*

Autores: Gustavo Nogueira de Sousa, Fabio Lobato, Julio Viana, Olaf Reinhold.

Veículo de publicação: *Publicado em anais do International Workshop on Integrated Social CRM (ICRM 2021)*[5]

[5] https://bisconf.org/2021/icrm/

Texto completo: O manuscrito publicado deste trabalho está no capítulo 6 desta dissertação.

Problema e Justificativa: Por meio de pesquisas na literatura, foi possível verificar que existem poucos estudos sobre gestão de reclamações por Instituições de Ensino Superior (IES), e também que o potencial de construção de relacionamentos e gestão do percurso do aluno é apenas parcialmente examinado na literatura. Diante disso, este trabalho tem como objeto de estudo a utilização de uma plataforma de reclamações online para melhorar o percurso acadêmico dos alunos nas IES.

Objetivo: O objetivo principal deste trabalho é demonstrar a aplicabilidade, em um cenário real de uso do Social CRM, da análise de reclamações publicadas nas mídias sociais. Para isso, a relevância das mídias sociais é explorada e analisada como canais para tratamento de reclamações como parte do processo de melhoria da trajetória acadêmica de estudantes em IES brasileiras. Para tanto, foram definidos os seguintes objetivos:

- Explorar a utilização das mídias sociais para reclamações por alunos;
- Avaliar se há uma ligação entre o número de reclamações e a gestão ativa das redes sociais;
- Identificar os principais temas das reclamações;
- Identificar a relação entre os diferentes problemas das universidades.

Impacto e contribuições técnicos científicas: Por meio das análises realizadas, este trabalho demonstra a aplicabilidade da análise de mídias sociais na melhoria da trajetória acadêmica em IES. A abordagem utiliza e valida as técnicas de aprendizado de máquina implementadas no *pipeline* de análise nos trabalhos apresentados nas Seções 2.2 e 2.3 e permitiu o tratamento e a extração de conhecimentos das reclamações coletadas. Os resultados deste trabalho foram avaliados por um especialista em desenvolvimento de negócios e por um especialista em Social CRM, ambos do *Social CRM Research Center*[6], que atestaram a viabilidade da utilização de dados de reclamações para melhorar a trajetória acadêmica.

2.5 CONSIDERAÇÕES

Neste capítulo, foram apresentados os pontos principais de cada um dos quatro estudos que compõem o cerne desta dissertação. Cada estudo está associado a um dos objetivos específicos descritos na Seção 1.2 e abordaram diferentes perspectivas para a análise de dados de mídias sociais voltada para negócios. Foi utilizada como base

[6] https://scrc-leipzig.de/

metodológica o CRISP-DM, tal como o processo descrito na Seção 1.3, o que possibilitou a execução eficiente de diferentes perspectivas e cenários de análise. Estes processos de análise e os resultados obtidos mostraram-se relevantes e tem o potencial para aprimorar a utilização do Social CRM.

Desta forma, os estudos resultaram nos artigos, os quais foram descritos neste capítulo. Com propósito de provê uma visão detalhada dos resultados obtidos, nos Capítulos 3, 4, 5 e 6 contém o texto completo de cada um dos artigos.

3 ARTIGO - GERENCIAMENTO DE PUBLICIDADE NAS PLATAFORMAS DAS REDES SOCIAIS DE ACORDO COM CATEGORIAS DE CONTEÚDO

RESUMO

O uso de mídias sociais está se expandindo por diferentes setores da sociedade, consequentemente, uma grande quantidade de conteúdos gerados pelos usuários é produzida todos os dias. Devido aos diferentes efeitos gerados nos usuários, a gestão de conteúdo é essencial para a publicidade comercial nessas plataformas. No entanto, o grande volume de dados faz com que os custos para a medição dos efeitos que os conteúdos têm sobre os usuários sejam elevados. Este artigo investiga o uso de técnicas de aprendizado de máquina para automatizar o processo de análise, aumentando a eficiência dos processos e a confiabilidade dos resultados. Mais especificamente, avalia-se o uso de um classificador de texto para categorizar as publicações de acordo com o seu conteúdo, como estudo de caso, adotou-se publicações de empresas no Facebook. Os resultados mostram que o classificador obtido apresenta potencial para analisar uma quantidade significativa de conteúdo com eficiência. O classificador tem implicações práticas, uma vez que permite uma extensa análise dos concorrentes e também é capaz de influenciar as campanhas de marketing em mídias sociais.

Palavras-chave: Rede Social, Propaganda, Engajamento de usuários, Aprendizado de Máquina, Tomada de Decisão Baseada em Dados.

ABSTRACT

Social media usage is expanding in different sectors of society; consequently, a large amount of User-Generated-Content is produced every day. Due to its various effects on users, content management is essential for business advertising on these platforms. However, the massive social media's data volume, increase the costs for analyzing the content effects on users. This paper examines the use of machine learning techniques to reduce the cost and effort of this kind of analysis. More specifically, an automatic document classification to identify content categories is evaluated. As a case study, we adopted some Facebook companies' posts. The results show that the machine learning classifier obtained has the potential to analyze a significant amount of content. The classifier has practical implications since it allows an extensive competitor analysis to be conducted and is also able to influence social media campaigns.

Keywords: Social Network, Advertising, User Engagement, Machine Learning, Data-Driven Decision Making.

3.1 INTRODUÇÃO

O uso de mídias sociais está crescendo constantemente (BELLO-ORGAZ; JUNG; CAMACHO, 2016). Em 2019, estima-se que essas plataformas tenham cerca de 2,77 bilhões de usuários, com o Facebook como plataforma líder, com 2,3 bilhões de usuários (SHIAU; DWIVEDI; LAI, 2018). Além disso, outras plataformas apresentam um grande número de usuários, como o Youtube com 1,9 bilhão, o Twitter com 330 milhões e o Instagram com 1 bilhão ((STATISTA, 2019). Devido à facilidade de uso, os usuários podem criar, interagir, colaborar e compartilhar conteúdo com outras pessoas por meio dessas mídias (MAIZ; ARRANZ; JUAN, 2016). Este fenômeno resultou em uma melhoria significativa na comunicação e interação social, impactando diretamente na relação entre as empresas e seus clientes (PRADIPTARINI, 2011; ALMEIDA; LOBATO; CIRQUEIRA, 2017; LOBATO et al., 2017; NOGUEIRA DE SOUSA et al., 2018).

Devido a facilidade para compartilhar informações nessas plataformas, iniciou-se um fenômeno chamado de Boca a Boca Virtual (electronic Word-of-Mouth - eWoM), que transformou os consumidores em atores ativos (BARRETO, 2014). O eWoM pode ser entendido como o ato de criar e compartilhar informações sobre marcas, produtos e serviços nas mídias digitais (SCHMÄH; WILKE; ROSSMANN, 2017). Isso significa que a mídia social se tornou um importante meio para compartilhar esses tipos de conteúdo (AHMAD; LAROCHE, 2017), podendo representar uma importante fonte de informações sobre as preferências das pessoas (ROSSOW, 2019).

As informações derivadas do eWoM permitem uma comparação entre marcas, produtos ou serviços (HUSSAIN et al., 2018). Ao aplicar métodos de detecção de comunidades em dados de mídia social, é possível fornecer insights úteis sobre algumas das dinâmicas e fenômenos que ocorrem nesses sistemas (SILVA et al., 2017). Além disso, este fenômeno está diretamente relacionado ao setor de turismo, já que um grande número de turistas seleciona seu destino, hotel, passeios e restaurantes com o auxílio de conteúdos de eWoM, tais como fotos, vídeos, avaliações e feedbacks (HARRIGAN et al., 2017; OLIVEIRA; CASAIS, 2018). Isso pode ser descrito como Turismo Inteligente, que é definido como atendimento ao cliente de forma onipresente por meio de informações turísticas relevantes, e é caracterizado pela provisão, gerenciamento e compartilhamento de serviços e experiências durante a jornada dos turistas (GRETZEL et al., 2015; LI et al., 2017).

As transformações e mudanças trazidas pelo uso do Turismo Inteligente podem ser potencializadas através da integração de sistemas de CRM (*Customer Relationship Management*) e planejamento de estratégias de atuação nas mídias sociais (COLOMO-PALACIOS et al., 2017). Essa estratégia reduz os riscos envolvidos na tomada de decisões e leva à transparência e confiança nos contatos com os clientes (COLOMO-PALACIOS et al., 2017; VECCHIO et al., 2018).

Tendo em vista a importância do eWoM para o mercado, neste artigo focamos na Publicidade em Redes Sociais (PRS) no que diz respeito a: A) As formas de conteúdo que são criadas pelas marcas e disseminadas pelas mídias sociais; e B) o Engajamento em Conteúdo Digital (ECD), que pode ser definido como o estado psicológico induzido pelas interações com a identidade da marca em um ambiente digital. Sete categorias de conteúdo PRS e três níveis de ECD são definidos por Gavilanes, Flatten e Brettel (2018), que podem ser correlacionados pelos analistas de marketing para determinar sua eficácia. Dada a grande quantidade de conteúdo PRS nas mídias sociais, são necessários altos custos e um esforço considerável para avaliar seu impacto no ECD (LIU et al., 2017a). Neste contexto, a seguinte questão de pesquisa foi definida:

- É possível a classificação automática de posts de acordo com as categorias de conteúdo de PRS apresentadas por Gavilanes, Flatten e Brettel (2018)?

Para responder a essa questão de pesquisa, testamos diversos métodos de aprendizado de máquina, a fim de desenvolver um classificador automático de conteúdo de PRS com base nas categorias supramencionadas. O método de classificação automática de dados foi desenvolvido e testado em publicações sobre turismo no Facebook. Além disso, os profissionais validaram o modelo e discutiram suas implicações práticas. Por exemplo, é possível usar o classificador na automação da análise de concorrentes, na implementação de estratégias de marketing, avaliando o conteúdo de PRS e seu engajamento. É possível também o desenvolvimento de um sistema de suporte à decisão para prever o engajamento do usuário com base no conteúdo da postagem.

O restante deste artigo está estruturado como segue. Na Seção 2 descreve-se a metodologia empregada no estudo. Os resultados são analisados e discutidos na Seção 3. Finalmente, as conclusões e sugestões de trabalhos futuros são apresentadas na Seção 4.

3.2 METODOLOGIA

Nesta seção, a questão de pesquisa é respondida por meio de uma descrição do conjunto de dados, seguida pelas categorias de PRS e níveis de ECD, e pelo estabelecimento da Estrutura Experimental.

3.2.1 DESCRIÇÃO DO CONJUNTO DE DADOS

Neste trabalho, usamos dados que foram extraídos de publicações no Facebook [1] de diversas empresas do setor de turismo. Seguindo os seguintes critérios de seleção: i) as empresas precisavam ter perfil no Facebook; e ii) as empresas deveriam estar ativas nos

[1] *Disclaimer: Todos os dados coletados no Facebook serão utilizados exclusivamente para fins de prova de conceito. Os autores não tem qualquer interesse no uso comercial desses dados.*

últimos seis meses, em outras palavras, ter postagens no respectivo período. Esta abordagem foi adaptada de NOGUEIRA DE SOUSA et al. (2018). O Facebook foi escolhido devido à sua popularidade, grande número de usuários e consequente relevância mercadológica. A extração de dados foi realizada por meio da API oficial do Facebook de janeiro a junho de 2018.

Os critérios de seleção foram definidos com base em Maiz, Arranz e Juan (2016). No total, dados de 93 empresas foram coletados de uma ampla gama de empreendimentos correlatos ao turismo, como bares, restaurantes, hotéis, pousadas e afins. Ao todo, houve um total de 10.925 publicações durante o período de extração, sendo os dados para cada publicação estão descritos na Tabela 2.

Tabela 2 – Description of the data extracted.

Dados	**Formato**	**Descrição**
Post ID	Numérico	Identificação de cada publicação na rede social.
Texto	String	Conteúdo textual de cada publicação.
Tipo	String	Tipo de publicação - "foto", "vídeo", "status"e/ou "link".
Link	URL[2]	Post link.
Data da publicação	Data	Publication date in the social network.
Reação	Emoticons	As reações dos usuários no post - essas reações são: "Curtir", "hahas", "amei", "wows", "Triste", "Raiva", "Especial".
Compartilhamento	Numérico	Número de vezes que a publicação foi compartilhada.

3.2.2 CATEGORIAS DE PRS E NÍVEIS DE ECD

Este artigo foca na publicidade em redes sociais e nos níveis de engajamento de conteúdo digital descritos por Gavilanes, Flatten e Brettel (2018). A Tabela 3 mostra as categorias de conteúdo com suas respectivas descrições. Da mesma forma, Tabela 4 mostra os níveis de ECD com uma avaliação de seu grau de influência sobre sua eficácia e a descrição de cada nível.

3.2.3 ESTRUTURA EXPERIMENTAL

A estrutura experimental adotada para o estudo é descrita na Figura 3 e é composta das seguintes etapas: 1) Aquisição de dados (dados brutos); 2) Pré-processamento de dados; 3) Anotação manual de uma amostra significativa dos dados; 4) Classificação do restante das publicações por meio de um algoritmo de aprendizado de máquina; 5) Medição da precisão e validação do classificador obtido no passo 4; 6) Correlação das categorias PRS com os níveis de ECD; 7) Validação dos resultados obtidos.

Tabela 3 – Categorias de PRS adotadas de Gavilanes, Flatten e Brettel (2018).

Categoria	Rótulo	Descrição
Nenhuma	-	Publicações que não pertencem a nenhuma das outras categorias
Anúncio de novo produto	NP	Publicações destacando o anúncio de novos produtos e/ou serviços.
Exposição do produto corrente	CP	Publicações que destacam o produto atual ou o retorno de um produto.
Sorteios e concursos	SW	Publicações com informações sobre sorteios, regras e regulamentos.
Vendas	SA	Publicações que anunciam vendas ou promoções de um produto, incluindo informações de descontos e vouchers.
Feedback do consumidor	CF	Publicações solicitando que os clientes forneçam informações, como avaliação do produto, avaliação ou problemas.
Informação e entretenimento	IT	Publicações que fornecem informações novas, úteis, educativas ou interessantes.
Marca da organização	OB	Publicações que destacam a organização ou marca (por meio de logotipos, legendas, informações gerais da empresa, atributos organizacionais, rede de lojas e funcionários).

Tabela 4 – Categorias de ECD descritas por Gavilanes, Flatten e Brettel (2018).

Categoria	Grau	Descrição	Métricas
Filtragem positiva	Moderado	Resposta mostrando atitudes emocionais positivas em relação ao conteúdo	Reações (Curtir, amei, wows, haha, tristes, raiva, especial).
Processamento cognitivo e afetivo	Moderado para forte	Co-criação no ambiente da marca	Comentário.
Apoio	Forte	Forte investimento cognitivo e emocional, co-criação de valor, publicação, autoexpressão	Compartilhamentos.

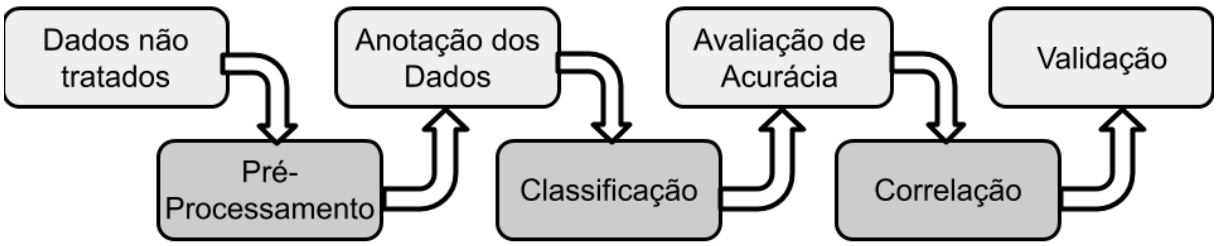


Figura 3 – Fluxo de trabalho dos experimentos.

Tabela 5 – Lista de algoritmos e sua melhor parametrização considerando a precisão da classificação.

Algoritmos	**Parâmetros**	**Acurácia**
KNN	k = 9, distance = 'cosine'	73 %
Gaussian Naive Bayes	Priors=None, Var_smoothing=10^{-9}	71 %
SVM	C = 10000, Kernel='sigmoid'	80 %
Multinomial Naive Bayes	Alpha=1.0, Class_prior=None, Fit_prior=True	69 %
Random Forest	n_estimators=7, min_samples_split=9	79 %

A primeira etapa, descrita na Figura 3, é a aquisição de dados, a qual foi realizada por meio da API do Facebook. O processo de classificação automatizado envolve preparar e anotar os dados para construir um modelo de classificação. A segunda etapa foi a aplicação de métodos de pré-processamento nos textos de cada publicação visando reduzir o ruído. Este passo seguiu o *workflow* descrito por Cirqueira et al. (2017b). O ruído, neste contexto, era composto por URLs, palavras irrelevantes, números, acentuação, emoticons e caracteres especiais.

A terceira etapa do fluxo de trabalho refere-se à anotação de dados manual de acordo com as categorias de PRS mencionadas anteriormente. A anotação foi realizada por dois avaliadores independentes utilizando-se de um sistema adaptado para tal (CIRQUEIRA et al., 2017a). A confiabilidade das anotações foi avaliada por meio do Coeficiente Kappa de Cohen e apenas os dados com concordância entre os anotadores foram retidos, em outras palavras, somente quando os dois avaliadores atribuíram o mesmo rótulo a uma publicação, a publicação era incluída ao conjunto de dados de treinamento.

Para o quarto passo do processo descrito na Figura 1, alguns algoritmos de aprendizado de máquina usados para classificação foram testados para responder às questões de pesquisa previamente definidas. Os algoritmos testados foram: K-Nearest Neighbors (KNN), Gaussian Naïve-Bayes, Support-Vector Machine (SVM), Multinomial Naïve-Bayes e Random Forest (RF). Esses algoritmos estão disponíveis no framework scikit-learn (PEDREGOSA et al., 2011). A parametrização do algoritmo foi realizada usando GridSearch para SVM, KNN e RF. Para os outros algoritmos os parâmetros padrão foram usados. Na Tabela 5 os algoritmos e parâmetros adotados são apresentados.

Foram utilizadas as seguintes medidas de desempenho: Acurácia, Precisão, *Recall* e *F1-measure* (ponderada), e os dados foram estratificados por meio de validação cruzada (10 vezes). A *F1-measure* (ponderada) foi adotada em vez de micro / macro, uma vez que leva em conta o desequilíbrio do rótulo. À luz das melhores medições de desempenho, o algoritmo SVM foi adotado (Tabela 5). A correlação das categorias de PRS com as métricas que determinam os níveis de ECD é realizada para determinar a eficácia do conteúdo. A validação foi realizada considerando um cenário real relacionado ao turismo inteligente, no qual dois profissionais que atuam nesse setor realizaram uma avaliação qualitativa dos resultados obtidos nas etapas anteriores.

O processo de avaliação consistiu em uma seleção aleatória de algumas publicações classificadas para serem usadas no processo de validação. Além disso, foi aplicado um algoritmo de modelagem de tópico para todo o conjunto de dados classificados, apresentando os resultados (por classe) para os profissionais. Eles analisaram se a correlação do tópico era consistente com as classes propostas. No entanto, devido a restrição no tamanho do artigo, a apresentação deste processo foi suprimida dos resultados.

3.3 RESULTADOS E DISCUSSÕES

Neste trabalho foram coletados um total de 10.925 publicações de 93 empresas atuantes no setor de turismo, como bares, restaurantes, hotéis, pousadas e afins. Considerando a quantidade de publicações coletadas, foram necessárias 628 publicações para obter um intervalo de confiança de 99% com uma margem de erro de 5%. No entanto, como era possível que os anotadores não correspondem ao mesmo rótulo para os mesmos dados, 1.020 publicações foram extraídas para anotação manual.

Como mencionado na Seção de Metodologia, as anotações foram avaliadas usando o coeficiente Kappa de Cohen e obtiveram o valor de 0,5, o que significa que há um nível moderado de concordância entre as anotações. Esses dados inicialmente representavam 9% do número total de publicações, embora apenas os dados classificados com o mesmo rótulo fossem mantidos, o que resultou em 680 publicações, ou cerca de 6 %, mais do que o requerido pelo intervalo de confiança e erro previamente definido. Todos os dados foram extraídos, anotados e processados aleatoriamente.

Vários estágios de pré-processamento foram executados no conjunto de dados de treinamento, para limpar e remover todas as informações desnecessárias, por exemplo, URLs e palavras irrelevantes. No classificador SVM, as publicações que foram anotadas manualmente foram submetidas ao treinamento do algoritmo, e os resultados obtiveram uma Acurácia de 80,87%, Precisão de 77,03%, Recall de 80,87% e F1-measure de 78,17%. As medições de desempenho, também, podem ser observadas na Figura 4, que foi normalizada para fornecer uma melhor visualização de dados.

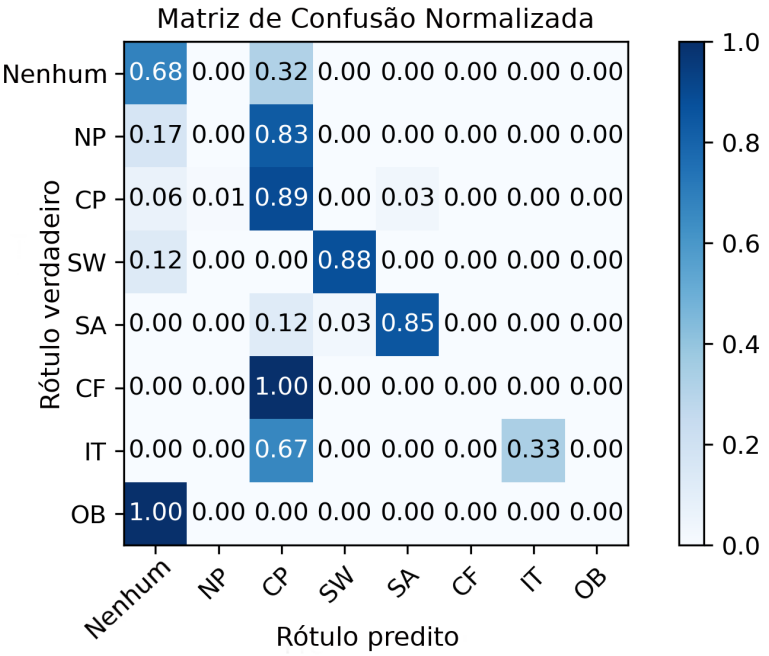


Figura 4 – Matriz de Confusão relacionada às categorias de ECD consideradas.

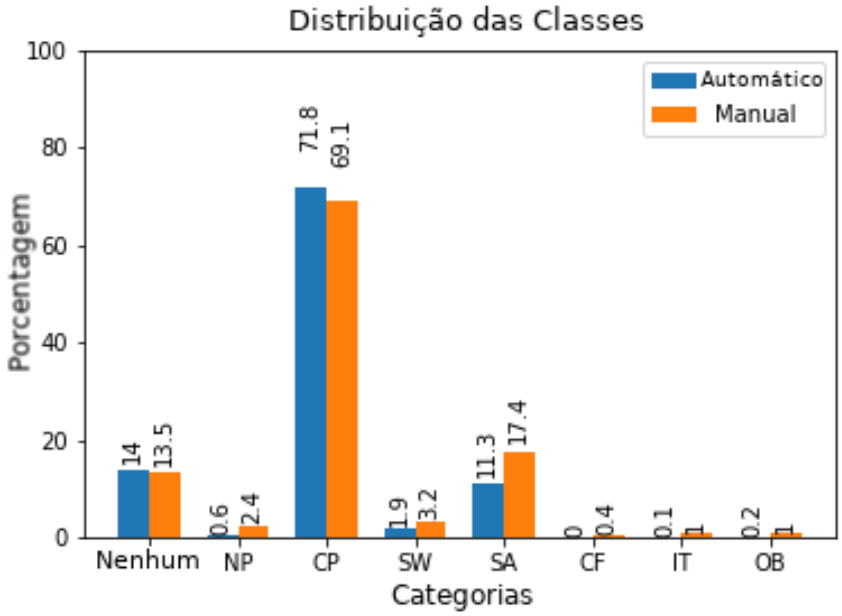


Figura 5 – Os resultados fornecidos por anotações automáticas e manuais por categorias.

Como forma de provar que a classificação automática de publicações pode obter resultados semelhantes a uma classificação manual, o classificador treinado foi aplicado ao restante do conjunto de dados (cerca de 94 % dos dados). Os resultados foram promissores, com a distribuição das categorias nos resultados observados na Figura 5 é possível observar que o uso de técnicas de aprendizado de máquina é uma maneira eficaz de reduzir o esforço da tarefa de rotulagem para medir a eficácia do PRS no ECD. Além disso, permite analisar

grandes quantidades de dados e demonstrar como é a melhor envolver os usuários por meio de marketing em redes sociais (GAVILANES; FLATTEN; BRETTEL, 2018).

A Tabela 6 foi montada com base nos resultados da classificação automática, e uma correlação pode ser observada entre as categorias de PRS e as métricas dos níveis de ECD propostas por Gavilanes, Flatten e Brettel (2018). Nesta e na Figura 6, as categorias de Sorteios e Concursos têm um apelo maior ao ECD de acordo com a média de curtidas, comentários e compartilhamentos em cada publicação.

Tabela 6 – Relação do PRS com o ECD em posts do Facebook. Nota: AV = Média e F = Frequência.

	Filtragem positiva		Processamento cognitivo e afetivo		Apoio	
	F	Av	F	Av	F	Av
Nenhum	87175	54.4	3820	2.3	5508	3.4
Anúncio de novo produto	23523	136.7	947	5.5	472	1.7
Exposição do produto corrente	604910	81.5	39367	5.3	26194	3.5
Sorteios e concursos	66005	244.4	42617	157.8	7455	157.8
Vendas	128148	100.9	3571	2.8	1742	1.3
Feedback do consumidor	331	16.5	27	1.35	4	0.2
Informação e entretenimento	1977	34	160	2.75	99	1.7
Marca da organização	7406	68.5	459	4.2	219	2

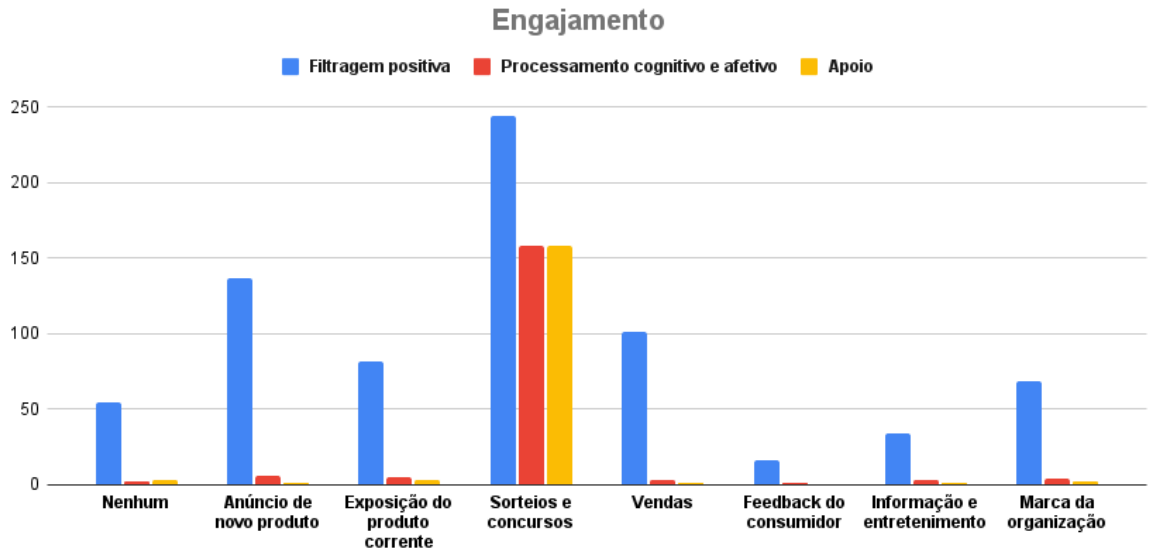


Figura 6 – Métricas de engajamento associadas às categorias de conteúdo obtidas a partir da classificação automática.

3.4 CONCLUSÃO

Esta pesquisa investigou o uso de técnicas de aprendizado de máquina para reduzir o custo e o esforço de avaliar estratégias de Publicidade em Redes Sociais. Para isso, adotamos as categorias de Engajamento de Conteúdo Digital propostas por Gavilanes, Flatten e Brettel (2018) e dados do Facebook. Com base em experimentos extensivos, avaliamos vários classificadores de aprendizado de máquina, a saber, K-Nearest Neighbors (KNN), Gaussian Naïve Bayes, Support Vector Machine (SVM), Multinomial Naïve Bayes e Random Forest (RF). Os resultados mostram que o classificador SVM tem um excelente potencial para realizar a classificação de conteúdo, a qual permite uma posterior correlação entre categorias e escala de engajamento. Sendo capaz de analisar uma grande quantidade de conteúdo com maior eficiência e menor custo/esforço. Assim, em termos proporcionais, o classificador automático pode alcançar resultados semelhantes aos obtidos pela anotação manual. No entanto, como é possível observar na Figura 4, algumas classes de conteúdo obtiveram uma classificação com erro inaceitável, a saber a NP, CF, IT e OB. Isto ocorreu devido ao desbalanceamento na quantidade de dados anotados manualmente para treino do classificador para estas classes, como exemplificado na Figura 5.

Os resultados do classificador foram avaliados e validados por profissionais. Com base nos resultados, algumas implicações práticas foram visualizadas: i) é possível acompanhar os concorrentes analisando automaticamente o conteúdo publicitário das mídias sociais; ii) o classificador pode ser incorporado em um sistema de suporte à decisão, auxiliando na mensuração do engajamento do usuário e no desenvolvimento de novas estratégias de marketing; iii) o classificador também pode apoiar a análise do engajamento do usuário de acordo com as demandas atuais do mercado por gerenciamento comunitário nos setores de energia e outros. Estes são serviços bem conhecidos fornecidos por agências de marketing digital (BARATA et al., 2018), que podem ser aprimorados pelo classificador obtido.

No entanto, existem algumas limitações em nosso estudo que precisam ser abordadas em estudos futuros. Primeiro, nosso conjunto de dados incluía apenas dados de turismo e, assim, resultou em um conjunto de dados desbalanceado. Em segundo lugar, como apenas dois anotadores foram usados, houve um baixo intervalo de confiança nos resultados. Em vista disso, em trabalhos futuros, gostaríamos de expandir nosso conjunto de dados para outros setores de mercado e incluir um terceiro anotador para melhorar a confiabilidade da pesquisa. Além disso, estamos planejando incluir técnicas que lidam com conjuntos de dados desbalanceados ou para melhorar o desempenho da classificação.

3.5 AGRADECIMENTOS

Os autores agradecem o apoio financeiro a essa pesquisa da Universidade Federal do Oeste do Pará (UFOPA), do Serviço Alemão de Intercâmbio Acadêmico (DAAD, SCRM-SPECS) e Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA).

4 ARTIGO - ANÁLISE DO SETOR DE TELECOMUNICAÇÃO BRASILEIRO: UMA VISÃO SOBRE RECLAMAÇÕES

RESUMO

Mídias digitais estão cada vez mais presentes no cotidiano do ser humano. Este fato contribui para que o volume de conteúdo gerado por usuário aumente consideravelmente. De um ponto de vista prático, as análises desses dados requerem diferentes perspectivas e métodos para se obter resultados satisfatórios. Essas análises podem subsidiar a tomada de decisão por gestores por meio da identificação de necessidades e problemas, guiando o processo de melhoria continuada de produtos e serviços. Diante disso, este trabalho propõe uma análise de reclamações postadas em uma plataforma online de reclamações, a fim de identificar pontos que orientem a tomada de decisões das empresas e, consequentemente, melhorar o relacionamento com clientes. Os resultados obtidos permitem a identificação de uma cadeia de problemas relacionados. A principal contribuição deste estudo está na provisão de uma abordagem que auxilia no planejamento estratégico de corporações, levando em consideração situações reportadas pelos consumidores.

Palavras-chave: Mídias Sociais, Mineração de Texto, Gestão de Relacionamento com Clientes, CRM Social, Reclamações.

ABSTRACT

Digital media are increasingly present in the daily life of human beings. This fact contributes to the increasing volume of user-generated content. From a practical point of view, the analysis of these data requires different perspectives and methods to obtain consistent results. These analyzes can support managers' decision-making by identifying needs and problems, guiding the process of continuous improvement of products and services. Therefore, this work proposes an analysis of complaints posted on an online complaints platform to identify points that guide companies' decision making and, consequently, improve the relationship with customers. The results obtained allow the identification of problems and their relationship. This study's main contribution is the provision of an approach that helps in the corporation's strategic planning, taking into account situations reported by consumers.

Keywords: Topic Modeling, Complaints, eWOW, Text Mining, Social Media.

4.1 INTRODUÇÃO

As mídias sociais oferecem uma ampla gama de funcionalidades que permitem ao usuário o compartilhamento e o consumo de conteúdo online (CARR; HAYES, 2015; SILVA et al., 2018). Este fenômeno impacta diretamente no relacionamento entre empresas e consumidores. Isso ocorre devido à popularização da internet e o aumento no acesso a essas plataformas sociais que, consequentemente, tornam os consumidores mais engajados com marcas e na troca de informações sobre produtos e serviços (LOBATO et al., 2017). Consequentemente, percebe-se um aumento substancial na quantidade de conteúdos gerados pelos usuários (User Generated Content – UGC) (BAHTAR; MUDA, 2016; LOBATO et al., 2017; NUSAIR et al., 2017).

UGC pode ser definido como qualquer forma de conteúdo criado, divulgado e consumido por usuários (KIM; JOHNSON, 2016); podendo incluir também dados relacionados a marcas, produtos ou serviços publicados em mídias sociais, os quais constituem uma subcategoria chamada de boca-a-boca virtual (Eletronic Word of Mouth - eWoM (ALMEIDA; CIRQUEIRA; LOBATO, 2017; SCHMÄH; WILKE; ROSSMANN, 2017). Este cenário impõe em um grande desafio para negócios, pois os usuários não apenas criam e compartilham conteúdo pessoal em seus perfis, mas também, recomendações, opiniões, reclamações e impressões sobre produto e serviço (ALT; REINHOLD, 2012). eWoM é visto como um forte determinante na decisão de compra, haja vista que cerca de dois terços dos consumidores verificam as avaliações de produto, serviços e marcas antes de decidirem adquiri-los (AHMAD; LAROCHE, 2017; CONSTANTINIDES; HOLLESCHOVSKY, 2016). A análise de dados relacionados ao eWoM tem o potencial de auxiliar na tomada de decisões por gestores, gerando respostas e melhorias significativas a partir da identificação de necessidades e problemas a serem resolvidos (GAVILANES; FLATTEN; BRETTEL, 2018; EINWILLER; STEILEN, 2015).

Pesquisas na literatura revelaram que há poucos trabalhos que realizam a análise e extração de conhecimento de eWoM expressas em plataformas de reclamações online. Encontra-se o uso massificado de diversas plataformas de eWoM como fonte de dados para a geração de conhecimento, tais como: Facebook (BAHTAR; MUDA, 2016; KIM; JOHNSON, 2016; LIU et al., 2017b; VERMEER et al., 2019), Twitter (CHAKRABORTY et al., 2017; EINWILLER; STEILEN, 2015; VERMEER et al., 2019), e reviews em loja de aplicativo (ALI; JOORABCHI; MESBAH, 2017; MCILROY et al., 2016; VU et al., 2016).

No Brasil, plataformas de reclamações online têm bastante relevância e influência. Segundo o site (ALEXA, 2019), no período de 01/11/2019 a 01/02/2020 o ReclameAqui esteve entre os 25 sites mais acessados do país, no qual, diariamente cada usuário fez em média três visitas na página com um tempo médio de três minutos nas interações. Além disto, percebeu-se que as empresas de telecomunicações são as mais mal avaliadas de acordo com o ranking que considera as 120.000 cadastradas na plataforma . Diante do

contexto apresentado, com destaque para a lacuna na literatura e do potencial de análise e de geração de conhecimento que eWoM apresenta, as seguintes perguntas de pesquisa nortearam o presente trabalho:

- PP1: Quais são os principais tópicos presentes em reclamações envolvendo empresas de telecomunicação?
- PP2: Como esses tópicos estão relacionados entre si?
- PP3: Quais os padrões distributivos das reclamações considerando dimensões geo-temporais?
- PP4: Quais as implicações práticas do resultado das análises conduzidas para os negócios?

Para responder às perguntas da pesquisa, usou-se dados de reclamações extraídos do site ReclameAqui de quatro empresas do setor de telecomunicações atuantes no Brasil. Este setor foi escolhido devido sua importância na garantia do desenvolvimento de uma sociedade, pois traz consigo a inclusão digital, igualdade de oportunidade, facilidade de transações e comunicação entre indivíduos(BANKOLE; OSEI-BRYSON; BROWN, 2015; MUJAHID et al., 2018; SHARMA et al., 2014). Devido às características deste trabalho, as empresas foram selecionadas considerando o número de clientes e a presença em todo o território nacional, e de acordo com a Agência Nacional de Telecomunicações (ANATEL) (2020), o *Market share* das empresas escolhidas neste trabalho representam juntas 97% de telefonia móvel, 72% da banda larga fixa, 96,9% de TV por assinatura e 94,4% de telefonia fixa.

O restante do artigo encontra-se organizado como segue. Na Seção 2 são apresentados os trabalhos relacionados. Na Seção 3 a metodologia utilizada é descrita. Os resultados são discutidos na Seção 4. Por fim, as conclusões do estudo e sugestões de trabalhos futuros são apresentadas na Seção 5.

4.2 TRABALHOS RELACIONADOS

Para que o conteúdo de mídia social seja útil para geração de conhecimentos que embasam a tomada de decisões, é necessário a definição e o uso de estratégias para a extração dos dados. Considerando que alguns dados são de difícil acesso, Olmedilla, Martínez-Torres e Toral (2016) define uma arquitetura para um framework com diretrizes e abordagens a serem seguidas para a extração de dados. Com isto, mostrou-se como um *Webcrawler* pode ser extremamente eficaz no processo de reunir e identificar grandes quantidades de conteúdo gerado pelo usuário. Além disso, mostram a importância das ciências sociais e

da computação no processo de análise de dados sociais. Estudos semelhantes baseiam-se neste framework, a citar (NETTO et al., 2019; RODRIGUES; JR; LOBATO, 2019).

Devido a informalidade dos textos e a necessidade de ajustes gramaticais, o pré-processamento se faz um processo fundamental para a análise de conteúdo de mídias sociais para se garantir a confiabilidade dos resultados. Cirqueira et al. (2018) realizaram uma análise na literatura para identificar os principais métodos utilizados para tratar conteúdos escritos em Português-Brasileiro. Foram reunidos um total de 62 artigos relevantes, os quais possibilitaram a listagem dos principais métodos e etapas necessárias.

Em relação aos tipos de análises, a modelagem de tópicos merece destaque, uma vez que permite a identificação de certos padrões nos dados, nos quais seriam difíceis a descoberta manual. Para isso, Ernala et al. (2018) utilizaram técnicas de modelagem de tópicos para identificar o nível de engajamento de usuários. Foram reunidos mais de 1,9 milhões de tweets de 146 usuários e, a partir das análises desses dados, os autores determinaram que menções, apoio emocional e discussões em torno da vida pessoal são fortes preditores de um ambiente em que é possível a divulgação dos mais variados tipos. Outros trabalhos apresentam a modelagem de tópicos como uma forma de extrair os termos importantes no texto, tal como (CIRQUEIRA et al., 2017a; ALDOUS; AN; JANSEN, 2019; ALMEIDA; CIRQUEIRA; LOBATO, 2017; LI et al., 2019).

Sob o ponto de vista da análise de eWoM, Rhee e Yang (2015) apresenta uma análise do setor de turismo a partir de reviews publicados em plataformas online. Os autores avaliaram seis aspectos relevantes em hotéis, com cinco tipos de viagens pré-definidos e viajantes de dois países diferentes. Como resultado, os autores verificaram que o "valor" e o tipo do "quarto" são os elementos que mais influenciam em uma avaliação. Assim sendo, a análise de conteúdos de mídias sociais permite que as ações nestas plataformas sejam mais efetivas e eficientes (CHIRUMALLA; OGHAZI; PARIDA, 2018).

No trabalho de Tang e Guo (2015) destacam a viabilidade e eficácia do uso de técnicas de mineração de texto para processar o conteúdo. No mesmo sentido, Carrascal et al. (2019) apresenta a técnica de mineração de texto como forma eficaz para a descoberta de conhecimento e conteúdos textuais. Os autores realizaram a identificação das palavras mais relevantes em conjuntos textuais e com potencial de realizar provisões futuras com base nos resultados.

Para o desenvolvimento deste trabalho optou-se pela metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM). A metodologia CRISP-DM é implementada a partir de um processo hierárquico, consistindo em um conjunto de tarefas que descrevem quatro níveis de abstração (CHINCHILLA; FERREIRA, 2016). Considerando os níveis de abstração propostos, os trabalhos de Rollins (2015), Schafer et al. (2019), Wirth (2000) apresentam descrições detalhadas dos processos que compõe o CRISP-DM. Estes processos formam o ciclo de um projeto de mineração de dados composto de seis

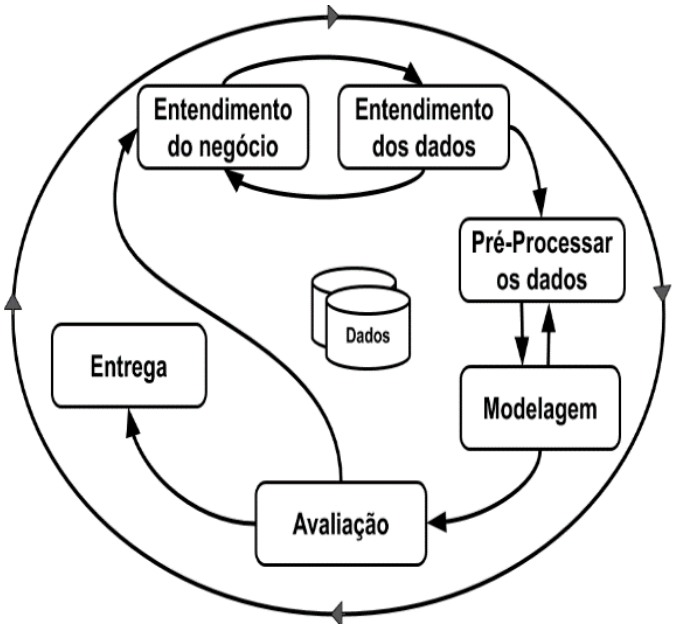


Figura 7 – Diagrama de funcionamento do modelo CRISP-DM (Adaptada de Wirth (2000)).

fases, conforme pode ser observado na Figura 7.

4.3 METODOLOGIA

Para o desenvolvimento deste trabalho optou-se pela metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM). A metodologia CRISP-DM é implementada a partir de um processo hierárquico, consistindo em um conjunto de tarefas que descrevem quatro níveis de abstração (CHINCHILLA; FERREIRA, 2016). Considerando os níveis de abstração propostos, os trabalhos de Rollins (2015), Schafer et al. (2019), Wirth (2000) apresentam descrições detalhadas dos processos que compõe o CRISP-DM. Estes processos formam o ciclo de um projeto de mineração de dados composto de seis fases, conforme pode ser observado na Figura 7.

Devido a característica cíclica deste método, o processo de mineração de dados não é finalizado quando uma solução é implementada. Neste caso, as lições aprendidas durante cada etapa podem gerar novas possibilidades de análises e novos resultados (WIRTH, 2000). Nas próximas subseções são descritas as etapas de Entendimento do Negócio, Entendimento dos Dados, Pré-Processamento dos dados e Modelagem. A etapa de Avaliação é descrita na Seção 5 e a Entrega é feita por meio da apresentação dos resultados aos *stakeholders*.

4.3.1 ENTENDIMENTO DO NEGÓCIO

Tomando em consideração a maleabilidade que a metodologia CRISP-DM possibilita, nesta etapa é realizado o entendimento do contexto em que as análises poderiam

ser aplicadas, a partir da definição dos objetivos da mineração de dados. Neste contexto, verificou-se que as reclamações podem representar um importante meio para obtenção de uma grande quantidade de informações autênticas sobre produtos e serviços feito de forma voluntária. Com base nisto, os seguintes objetivos foram definidos:

- Identificar os principais assuntos/problemas reportados nas reclamações;
- Identificar aspectos específicos das reclamações;
- Analisar a distribuição das reclamações considerando dimensões geo-temporais;
- Identificar as implicações práticas das reclamações nas empresas.

4.3.2 ENTENDIMENTO DOS DADOS

Nesta etapa são realizadas a coleta, descrição, exploração e verificação da qualidade dos dados coletados. Devido as dificuldades de utilização da API fornecida, os dados foram extraídos da plataforma de reclamações online chamada ReclameAqui [1] por meio de um *WebCrawler* escrito em Python, o qual utiliza a biblioteca *requests*[2]. Esta plataforma foi escolhida devido sua popularidade. Dentre todos os sites acessados no Brasil, o ReclameAqui é o 25º site com maior número de acessos (ALEXA, 2019), sendo o site mais popular na categoria de reclamações.

Como alvo da extração, foram selecionadas as quatro maiores empresas do setor de telecomunicações brasileiro. Estas empresas foram escolhidas devido a abrangência nacional de prestação de serviço, as quais atendem milhões de clientes em todas as regiões do país. Além disso, essas empresas são apontadas como as piores empresas no ranking fornecido pelo ReclameAqui, o qual contém mais de 120.000 empresas cadastradas. Cada reclamação extraída era composta pelos dados detalhados na Tabela 7.

Campos	Descrição
Empresa ID	Identificação da empresa relacionada na reclamação
Reclamação ID	Identificação da reclamação
Título da Reclamação	Título dados pelo o usuário a reclamação
Reclamação	Relato do problema reportado
Estado e cidade	Estado e cidade do consumidor que realizou a reclamação
Data/hora	Data e horário do relato

Tabela 7 – Descrição dos dados coletados

A fim de verificar a qualidade dos dados extraídos, foi realizada uma comparação manual com os dados da plataforma. Para isso, foi utilizado um conjunto amostral que

[1] *Disclaimer: É importante ressaltar que todos os dados coletados no ReclameAqui serão utilizados exclusivamente para fins de prova de conceito. Os autores não tem qualquer interesse no uso comercial desses dados.*

[2] https://requests.readthedocs.io/en/master/

representa um grau de confiança de 95% e uma margem de erro de 4% considerando o total de reclamações coletadas na plataforma por este estudo. Por fim, foi verificado que, com o uso da ferramenta de extração, os dados mantiveram o padrão de qualidade observado no site da plataforma.

4.3.3 PRÉ-PROCESSAMENTO DOS DADOS

Nesta etapa foi realizada a aplicação de técnicas de pré-processamento nos textos de cada reclamação. Em todos os dados de reclamações foram realizadas as seguintes tarefas de remoção: de saudações, de URLs, *stopwords*, números, acentuação e de caracteres especiais (CIRQUEIRA et al., 2018).

A remoção de saudações e de URLs significa que todo conjunto de caractere que representa uma saudação (e.g. "Olá", "Oi") ou um endereço de algum site (e.g. "www.site.com.br") foi removido. Da mesma forma, números, acentuação e caracteres especiais foram retirados, visto que são desnecessários para as análises. Palavras que são consideradas *stopwords*, ou seja, palavras que não contribuem para o significado do texto (e.g. "e", "de", "em") foram eliminadas.

4.3.4 MODELAGEM

Esta etapa foi dividida em duas fases: a aplicação da extração de tópicos; e a correlação entre os tópicos encontrados. Na primeira fase da modelagem foram aplicadas técnicas de extração de termos relevantes sobre os dados. Alguns algoritmos, tal como o *Non-Negative Matrix Factorization* (NMF), *Latent Dirichlet Allocation* (LDA) e *Latent Semantic Analysis* (LSA) foram utilizados nesta fase.

Durante o processo de avaliação qualitativa e anotação dos tópicos percebeu-se que os termos obtidos pelo NMF estavam mais relacionados entre si e que representavam tópicos mais coerentes e diversos (CHEN et al., 2019). Devido a isto, o *Term Frequency-Inverse Document Frequency* (TF-IDF) juntamente com o NMF foram adotados neste trabalho com o objetivo de classificar todas as palavras por ordem de importância no conjunto de textos, tal como demonstrado por Salminen et al. (2018), Trstenjak, Mikac e Donko (2014).

A modelagem e análise de todas as reclamações foi realizada de acordo com a empresa sob uma perspectiva nacional e regional. Os tópicos foram determinados pelos autores de forma manual com base nos termos obtidos na modelagem. Na segunda fase foi realizado o estudo de correlação dos tópicos obtidos. Vale destacar que um tópico representa um conjunto de termos relacionados, sendo que cada termo é associado há um conjunto de reclamações. Neste caso, para realizar a correlação, os dados foram modelados da seguinte forma: cada tópico é representado por um nó; as reclamações foram convertidas em arestas, as quais ligam os diferentes tópicos os quais tem concorrência.

Neste sentido, o conjunto de dados foi então transformado em registros contendo pares de arestas, por meio de uma combinação simples denotada por:

$$r = \frac{n!}{p!(n-p)!)} \tag{4.1}$$

Sendo que r é a quantidade de registros resultantes; n é o número de palavras-chave do trabalho; e p foi definido como 2 (dois), pois as arestas são formadas aos pares. Como resultado desta modelagem, foi criado um arquivo *Comma-Separated-Values* (CSV), o qual que pode ser visualizado com o suporte de uma ferramenta de análise de redes. No presente estudo o software Gephi[3] foi utilizado para este fim.

4.4 RESULTADOS

Nesta seção serão apresentados os resultados obtidos a partir das análises descritas anteriormente. Os resultados são divididos em três subseções. A primeira apresenta os dados coletados e a relação do número das reclamações com a distribuição da população brasileira. A modelagem de tópicos em termos nacionais e regionais é mostrada na segunda subseção. E, por fim, na terceira subseção são apresentadas as análises geo-temporal das reclamações obtidas.

4.4.1 EXTRAÇÃO DOS DADOS

O processo de coleta dos dados resultou em um total de 397.950 reclamações. Considerando a dinamicidade do setor de telecomunicação e sua propensão a mudanças (STONE, 2015), nestas análises foram utilizadas as reclamações do período de 09/09/2018 a 09/09/2019, resultando em 225.593 reclamações. Na Tabela 8 é apresentado as quantidades de reclamações coletadas em nível regional e nacional de cada empresa.

Tabela 8 – Distribuição dos dados pelas regiões do país

	Norte	Nordeste	Centro-Oeste	Sudeste	Sul	Total
Tim	1.377	10.100	3.176	47.339	9.282	72.174
Vivo	1.114	6.836	4.610	59.678	9.304	81.542
Oi	1.257	6.401	2.765	19.459	4.779	34.661
Claro	974	3.601	2.753	27.056	3.732	38.116

O Brasil possui aproximadamente 210 milhões de habitantes distribuídos em 5 regiões (IBGE, 2019a). As reclamações analisadas foram estratificadas de acordo com a operadora e a respectiva região de origem, conforme pode ser observado na Tabela 8. A região Sudeste tem um destaque frente as demais. Esta região possuí aproximadamente

[3] https://gephi.org/

Tabela 9 – Número de tópicos encontrados e resultantes

	Tópicos encontrados	**Tópicos resultante**
Tim	31	17
Vivo	29	20
Oi	26	17
Claro	26	19

42% da população brasileira e apresenta uma taxa superior a 65% do total de reclamações coletadas. Este fato pode ser justificado, uma vez que a região sudeste é a mais rica do país, com aproximadamente 53% de todo o Produto Interno Bruto (PIB) produzido. Enquanto, por exemplo, a região Norte representa aproximadamente 6% do total (IBGE - Instituto Brasileiro de Geografia e Estatistica, 2020).

4.4.2 MODELAGEM DE TÓPICOS

Nesta subseção é apresentada a modelagem de tópicos sob duas perspectivas: 1) tópicos de reclamações de todo o país; 2) a partir de uma distribuição regional. Devido às características da plataforma analisada, na qual o usuário seleciona a categoria do problema, optou-se por utilizar a mesma quantidade de categorias de problemas como a quantidade de tópicos a serem modelados. Isto permite dimensionar o espaço de busca. Maiores detalhes são descritos a seguir.

4.4.2.1 PANORAMA NACIONAL

O total de reclamações de cada empresa (apresentados na Tabela 8) foram utilizados para compor os tópicos de reclamação no panorama nacional. A partir da limitação do número de tópicos frente as categorias de problemas da plataforma, obteve-se a seguinte distribuição por empresa: 31 tópicos da Tim; 29 na Vivo; 26 tópicos na Oi; e 26 na Claro. Após uma análise inicial, alguns tópicos semelhantes (termos sinônimos, por exemplo) foram combinados. Além disso, foram considerados apenas tópicos únicos e relevantes. O resultado encontra-se sintetizado na Tabela 9.

Na Tabela 10 são expostos os principais tópicos das reclamações de acordo com a quantidade observados na Tabela 9. Ao observar os dados é possível reconhecer os problemas específicos de cada empresa no país. Estes dados podem ser utilizados como fundamento para o início de um processo de melhoria do serviço, fidelizar clientes e conquistar vantagens competitivas em relação aos demais concorrentes (GAVILANES; FLATTEN; BRETTEL, 2018; EINWILLER; STEILEN, 2015).

Para aprimorar a discussão dos resultados, alguns detalhes sobre o conhecimento do domínio fazem-se necessário, isto é, características inerentes às empresas analisadas. Por exemplo, a Tim oferece um plano exclusivo para seus clientes, o qual é acessível somente

por meio de um convite enviado por outros usuários desse plano[4]. A partir da identificação do tópico relacionado é possível verificar que reclamações relacionadas a este serviço são frequentes e podem ser relacionados este processo de negócio utilizado pela empresa.

Tabela 10 – Modelagem de tópicos para avaliação do panorama nacional.

Tim	Vivo	Oi	Claro
	Atendimento		Atendimento
Atendimento	Ativação chip	Atendimento	Ativação linha
Ativação chip	Cancelamento	Cancelar serviço	Cancelamento
Cancelamento	Cobertura móvel	Cancelamento	Cobertura móvel
Cobertura móvel	Cobrança indevida	Cobrança indevida	Cobrança indevida
Cobrança indevida	Internet residencial	Instalação de equipamento	Internet móvel
Convite para plano	Ligações telemarketing	Internet móvel	Ligações telemarketing
Franquia de dados	Linha cancelada	Ligações telemarketing	Linha cancelada
Ligações telemarketing	Loja física	Linha cancelada	Loja física
Loja física	Mudança de endereço	Loja física	Multa por fidelidade
Nome negativado	Mudança de plano	Multa por fidelidade	Nome negativado
Linha cancelada	Multa por fidelidade	Nome negativado	Número cancelado
Pagamento	Nome negativado	Plano controle	Pacotes de dados
Plano	Número cancelado	Planos	Pagamento
Plano Controle	Pagamento	Portabilidade	Planos
Portabilidade	Portabilidade	Qualidade do serviço	Portabilidade
Recarga	Qualidade do serviço	Recarga	Recarga
Valor serviço	Recarga	Telefone fixo	Senha de acesso
	Visita técnica		Serviço de música

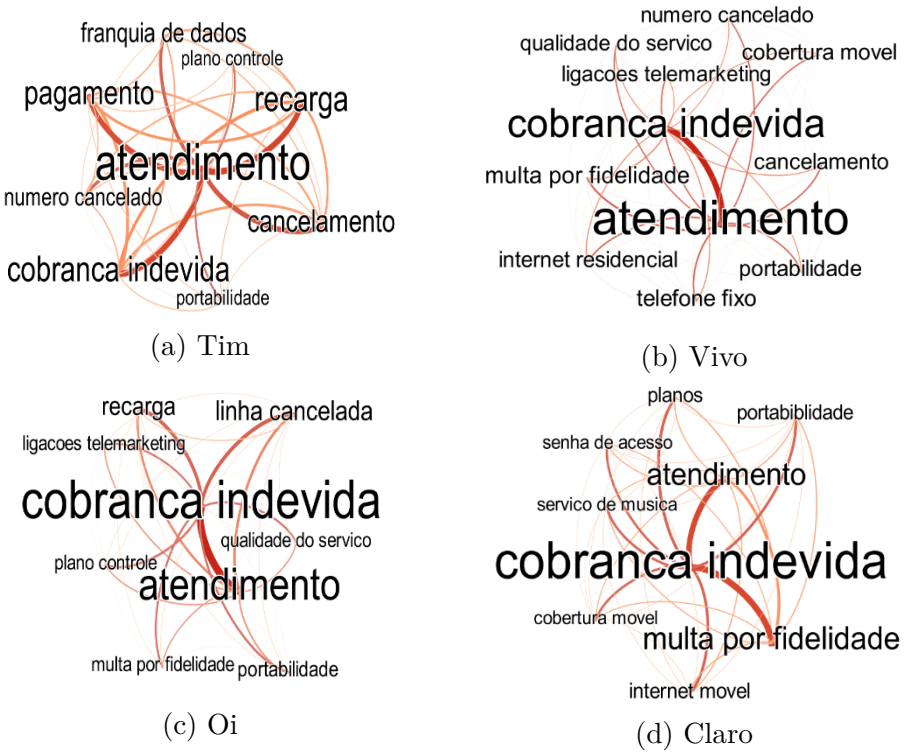
De forma geral, as reclamações encontram-se relacionadas a mais de um tópico. A correlação entre os tópicos das reclamações dos consumidores propicia a verificação da distribuição desta cadeia de insatisfação. Na Figura 8, são expostas as relações entre os tópicos, sendo que tanto o tamanho das palavras quanto a espessuras e variação dos tons das arestas representam a importância e o peso da relação.

Ao se analisar os dados dispostos Figura 8, três pontos merecem destaques em relação às companhias estudadas:

- Tim - tem como seu principal problema o atendimento. Na Figura 8a é visível que há associações entre diversos tópicos o que indica que a empresa tem dificuldade de oferecer soluções para seus clientes por meio do atendimento. Além disso, problemas como "pagamento", "cobrança indevida"e "recarga"que estão fortemente associados e indica falhas recorrentes na maneira que a empresa trata o processo de cobrança dos clientes;

- Vivo e Oi apresentam como seus principais gargalos o atendimento e a cobrança indevida. Na Figura 8b (Vivo) e na Figura 8c (Oi) há uma grande associação entre os dois problemas, sendo que para a empresa Vivo o atendimento tem peso maior.

[4] https://www.timbeta.com.br/timbeta/como_ser_beta



(a) Tim

(b) Vivo

(c) Oi

(d) Claro

Figura 8 – Relação entre tópicos das reclamações

Já para a empresa Oi, cobranças indevidas o fruto de maior frustração entre clientes. Isto indica que os processos de reclamação sobre cobrança indevida não são tratados satisfatoriamente pelos canais de atendimento, gerando descontentamento para seus clientes;

- Claro - tem um conjunto de 3 tópicos fortemente associados, que são a "cobrança indevida", “atendimento” e “multa por fidelidade”, como observado na Figura 8d. Essa associação de tópicos indica uma forte relação entre os três assuntos e facilita a análise do teor da reclamação.

4.4.2.2 ANÁLISE REGIONAL

A modelagem também foi realizada considerando uma perspectiva regional. Os tópicos obtidos para cada região são apresentados na Tabela 11. Os tópicos que iniciam com “+” são específicos para a empresa no contexto regional. Já os que iniciam com “-” são tópicos que não faz parte do contexto da região relacionada.

Contrastando-se as Tabelas 10 e 11, cenário nacional e regional, respectivamente, é possível evidenciar as particularidades de cada empresa por região. Por exemplo, na Tim é possível notar que as regiões com mais problemas que diferem do contexto nacional são a Nordeste e Centro-Oeste. Para a empresa Vivo, a região Norte apresenta a maior diferença em relação aos problemas nacionais. Nela não há oito tópicos de problemas nacionais e há quatro novos tipos de problemas que são específicos da região.

Para a empresa Oi, o tópico "renegociação" é um problema específico das regiões Sul, Nordeste e Centro Oeste. Por fim, a Claro não apresenta problemas específicos regionais. Contudo, alguns problemas nacionais que não fazem parte dos contextos regionais, como é o caso da região Sudeste que não apresenta dentro os principais tópicos de problema o "número cancelado" e o "pacote de dados".

Diante disso, empresas de telefonia podem analisar quais os tópicos mais reclamados em uma determinada região, bem como os tópicos que não são relevantes (no período estudado) para os clientes em cada área do país. Tal informação permite que áreas menos populosas e com menos reclamações sejam tratadas de acordo com a sua especificidade, eliminando o possível viés que as áreas mais populosas podem causar em análises que consideram reclamações de todo país e elencam somente problemas nacionais.

É importante ressaltar que, a análise de relação entre tópicos de reclamação exemplificado na subseção 4.4.2.1, é perfeitamente aplicável nos contextos regionais, estaduais e municipais, no entanto por questões de espaço estas análises foram suprimidas do presente trabalho.

Tabela 11 – Tópicos por regiões das empresas estudadas.

	Norte	Sul	Sudeste	Nordeste	Centro-Oeste
Tim	+ acesso a serviços + multa por fidelidade + renovação de pacote + serviço de música	+ internet móvel + promoção + sms indevido	+ acréscimo nos valores + multa por fidelidade + serviço de música	+ mudança de plano + multa por fidelidade + promoção + serviço de música + sms indevido	+ internet móvel + multa por fidelidade + promoção + qualidade do serviço + serviço de música
	- plano controle	- cobertura móvel - plano controle	- cancelamento - plano controle	- loja física - número cancelado - plano controle	- cancelamento - plano controle
Vivo	+ celular bloqueado + instalação de equipamento + plano + sms ilimitado		+ planos	+ entrega de compras	+ cancelamento
	- internet residencial - ligações telemarketing - linha cancelada - mudança de endereço - multa por fidelidade - qualidade da internet - qualidade do serviço - telefone fixo - visita técnica	- ativação chip - cancelamento	- cancelamento - qualidade da internet - telefone fixo	- mudança de endereço - qualidade da internet - relação - telefone fixo	- internet residencial - mudança residencial - telefone fixo
Oi	+ contrato + pagamento	+ manutenção + renegociação	+ manutenção	+ pagamento + renegociação	+ renegociação
	- multa por fidelidade - plano controle - telefone fixo	- cancelamento - plano controle	- plano controle - telefone fixo	- cancelamento - cobertura móvel - ligações telemarketing - plano controle - qualidade do serviço - telefone fixo	- cancelamento - plano controle - qualidade do serviço - telefone fixo
Claro					
	- cancelamento - linha cancelada - loja física - pacotes de dados - serviço de música	- linha cancelada - loja física - pacotes de dados - serviço de música	- número cancelado - pacotes de dados	- cancelamento - linha cancelada - pacotes de dados - serviço de música	- linha cancelada - loja física - pacotes de dados - recarga - senha de acesso - serviço de música

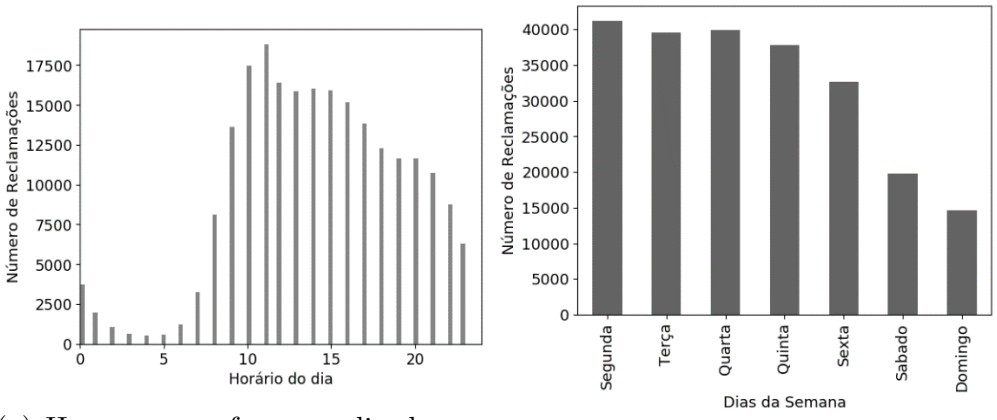
4.4.2.3 DISTRIBUIÇÃO "GEO-TEMPORAL"

As reclamações coletadas estão distribuídas em todo território do Brasil. São distribuídas de forma heterogênea e desproporcional a taxa populacional de cada região. Na Tabela 12 é possível observar como se dá a distribuição geográfica das reclamações no país através da comparação das taxas populacionais, PIB, Linhas ativas e de reclamações em cada região. Os dados são apresentados de forma a relacionada ao PIB, taxa populacional e de linhas móveis ativas em cada região.

Tabela 12 – Proporção Populacional (IBGE, 2019a), PIB (IBGE - Instituto Brasileiro de Geografia e Estatistica, 2020), Linhas Ativas (ANATEL, 2020) e de reclamações por Região

	População	**PIB**	**Linhas ativas**	**Reclamações**
Norte	9%	6%	7%	2%
Nordeste	27%	10%	24%	12%
Centro-Oeste	14%	17%	8%	6%
Sudeste	42%	53%	46%	68%
Sul	14%	14%	15%	12%

Na Figura 9, são expostas as quantidades de reclamações por horário e dia da semana, sendo que na Figura 9a contém a distribuição de todas as reclamações em 24 horas e na Figura 9b contém a distribuição de todas as reclamações nos dias da semana.



(a) Hora em que foram realizadas as reclamações

(b) Número de reclamações por dia da semana.

Figura 9 – Número de reclamações por dia da semana.

Observando-se a Figura 9 é possível perceber que a distribuição em relação aos dias da semana e horários seguiram também estáveis com variações para baixo nos finais de semana e entre onze horas da noite e oito horas do dia seguinte. No entanto, essas informações podem auxiliar as empresas na definição de tarefas e processos internos para solucionar os problemas levantados nessa plataforma online de reclamações de forma ágil e eficiente.

A análise dos dados dispostos na Tabela 12 e figura 9 nos revela alguns *insights* importantes sobre a distribuição das reclamações no país, a saber:

- Região Norte é a região como menor PIB do país, menor número de habitantes e em quantidade de linhas de celulares ativas. Há uma certa discrepância em relação à proporção de reclamações de outras regiões. Todas as outras tiveram a taxa de reclamação mais próxima do número de linhas ativas;
- Região Centro-Oeste tem cerca de 8% do total de linhas ativas e 6% das reclamações coletadas são referentes a essa região;
- Região Sudeste tem a maior taxa populacional e o maior número de reclamações no país. Sendo assim, a taxa de reclamação desta região é bastante elevada, ficando cerca de 22 pontos percentuais acima do número de linhas ativas;
- Todas as outras regiões, exceto a região sudeste, tiveram taxas de reclamações inferiores a taxa de linhas ativas. O que pode indicar que os clientes da região Sudeste são mais propensos a utilizar plataformas de reclamações online para realizar suas queixas.

4.5 CONSIDERAÇÕES FINAIS

Neste artigo, foram analisadas reclamações de quatro empresas do setor de telecomunicações brasileiro objetivando determinar quais eram os principais tópicos nas reclamações, como elas se relacionam entre si e qual a sua distribuição geo-temporal. Para tal, utilizou-se a metodologia CRISP-DM, sendo que na fase de extração de conhecimento foram aplicados métodos de modelagem de tópicos a nível nacional e regional. Os resultados mostram que o uso de técnicas de mineração de texto combinado com uma análise geo-temporal permite a antecipação e a correção de problemas, a análise de grandes volumes textuais de forma rápida e análises estratégicas, além de prover bases sólidas para a tomada de decisão por gestores.

As análises conduzidas permitem responder as perguntas de pesquisa, relevando importantes *insights* sobre as empresas analisadas. Por exemplo, em relação à PP1 – verificou-se que dentre os principais tópicos das reclamações o “atendimento” e “cobrança indevida” são os tópicos que mais se destacam nas empresas analisadas. Para PP2, - identificou-se que a relação entre os tópicos é diferente em cada empresa e indicam que a cadeia de problemas a serem solucionados variam entre os concorrentes. Acerca da PP3, verificou-se que existe uma diferença regional em relação aos tópicos das reclamações analisadas, porém não foram identificadas diferenças significativas no número de reclamações durante o ano. Por fim, por meio da resposta da PP4, os resultados auxiliam as empresas de telecomunicação no processo de tomada de decisão, uma vez que essas empresas podem

direcionar seus esforços para solucionar os problemas de seus clientes de acordo com o contexto em cada região.

Diante disto, a principal contribuição deste estudo está na provisão de uma abordagem de análise de reclamações que identifique as reais necessidades dos usuários de telefonia, auxiliando as empresas identificadas no estudo na implementação de soluções personalizadas por tópico e por região.

No entanto, este estudo apresenta algumas limitações que precisam ser tratadas em trabalhos futuros. A primeira está relacionada a forma como os tópicos foram determinados os quais podem conter viés dos autores. A segunda limitação é relacionada às análises e resultados obtidos, pois não incluem a aplicação prática dos resultados nas empresas. Como trabalhos futuros pretendemos incluir a validação cruzada em todas as etapas da metodologia, e expandir as análises incluindo o uso prático em empresa dos resultados encontrados.

5 ARTIGO - ANÁLISE COMPARATIVA DAS PRINCIPAIS PLATAFORMAS DE RECLAMAÇÕES ONLINE: IMPLICAÇÕES PARA ANÁLISE DE MÍDIA SOCIAL EM NEGÓCIOS

ABSTRACT

New forms of relationship between companies and customers have been introduced through the massive use of social media, and have transformed the way in which customers make purchasing decisions. This new reality explained the importance of content analysis related to brands and products published by consumers on social media platforms. In this sense, this article presents a comparative analysis of the two largest online complaint platforms in Brazil, *ReclameAqui* and *Consumidor.gov*. The analyzes use content from textual data mining of consumer complaints from five major companies in the Brazilian ecommerce sector, in order to provide a basis for understanding the challenges and opportunities in the analysis of these social media in business. The results show that the way in which companies operate on these platforms must be specific to each platform, since each platform has its particularities, such as consumer group, content and main types of problems.

Keywords: Social Media, Data Mining,d Topic Modeling.

RESUMO

Novas formas de relacionamento entre empresas e clientes foram introduzidas através do uso massivo de mídias sociais, e têm transformado a forma com a qual clientes tomam decisões de compras. Esta nova realidade explicitou a importância da análise do conteúdo relacionado a marcas e produtos publicados por consumidores em plataformas de mídias sociais. Nesse sentido, este artigo apresenta uma análise comparativa das duas maiores plataformas de reclamação online no Brasil, o *ReclameAqui* e o *Consumidor.gov*. Nas análises são utilizados os conteúdos provenientes de mineração de dados textuais de reclamações de cinco grandes empresas do setor de *ecommerce* brasileiro, com o objetivo de fornecer uma base para compreender os desafios e oportunidades nas análises destas mídias sociais em negócios. Os resultados mostram que a forma de atuação das empresas nessas plataformas deve ser especifica para cada plataforma, pois existem particularidades, tais como grupo de consumidores, conteúdo e principais tipos de problemas.

Palavras-chave: Mídias Sociais, Mineração de Dados, Modelagem de Tópicos.

5.1 INTRODUÇÃO

O surgimento de novas ferramentas de comunicação modificaram a forma com a qual as pessoas tomam suas decisões de compra (ORENGA-ROGLÁ; CHALMETA, 2016; LOBATO et al., 2017). Os canais de compartilhamento de informações online facilitam o diálogo entre consumidores de todos os lugares, independentemente de suas posições geográficas (CONSTANTINIDES; HOLLESCHOVSKY, 2016). E, assim como o boca-a-boca tradicional se faz presente na sociedade, o ambiente virtual tornou-se um dos meios de disseminação e troca de experiências a respeito de produtos e serviços (KIM; JOHNSON, 2016). Como consequência, surgiu o conceito de boca-a-boca virtual (*Eletronic Word-of-Mouth, eWoM*), definido como o ato de compartilhar opiniões sobre produtos e serviços em mídias sociais (SCHMÄH; WILKE; ROSSMANN, 2017).

O eWoM tem se mostrado ainda mais eficaz que o ”boca-a-boca” tradicional, devido a sua facilidade de propagação e de engajamento (CONSTANTINIDES; HOLLESCHOVSKY, 2016). Por este motivo, os consumidores passaram a ser chamados de *prosumers*, uma vez que são responsáveis tanto pelo consumo quanto pela produção de conteúdo (ROY; DATTA; MUKHERJEE, 2019). Ao estudar conteúdos de eWoM, as empresas têm uma ricas fontes de conhecimento sobre seus clientes e quais são suas preferências (ALDOUS; AN; JANSEN, 2019). Porém, a sobrecarga de conteúdo torna o entendimento individual dos clientes um grande desafio (LOBATO et al., 2017).

Na literatura, há diversos estudos que utilizam dados de revisões e de reclamações online obtidos de diferentes plataformas de mídias sociais tais como *Twitter*, *Facebook*, *TripAdvisor*, *Booking.com* e entre outras (BAHTAR; MUDA, 2016; SILVA et al., 2017; XIANG et al., 2017; VERMEER et al., 2019). Geralmente, estes estudos utilizam em suas análises um conjunto amostral de dados para extrair informações que possibilitam a detecção, descrição ou previsão de padrões que influenciam na tomada de decisões teóricas e/ou práticas. Entretanto, há poucos estudos tal como o de Almeida, Lobato e Junior (2019) que utilizam em suas análises plataformas específicas para publicações de reclamações online de marcas e/ou produtos.

Com isto em mente, este estudo analisa comparativamente o conteúdo textual das reclamações de duas plataformas de reclamações online[1], a saber o ReclameAqui (RA) e o Consumidor.gov (CGOV). O objetivo é prover uma base para a compreensão dos desafios e oportunidades que estas plataformas proporcionam para o desenvolvimento de análises de mídias sociais que orientem a tomada de decisões em diversos segmentos do mercado.

O restante do artigo encontra-se organizado como segue. Na Seção 5.2 são apresentados os trabalhos relacionados, na Seção 5.3 é descrito processo de condução da pesquisa,

[1] *Disclaimer: É importante ressaltar que todos os dados coletados no ReclameAqui e no Consumidor.gov serão utilizados exclusivamente para fins de prova de conceito. Os autores não tem qualquer interesse no uso comercial desses dados.*

na Seção 5.4 são apresentados os resultados decorrentes das análises e por fim na Seção 5.5 o autor apresenta suas considerações finais ao desenvolvimento do estudo.

5.2 TRABALHOS RELACIONADOS

O engajamento de consumidores em múltiplas plataformas de mídias sociais é um grande desafio que empresas têm enfrentado. Neste sentido, Aldous, An e Jansen (2019) buscou entender como o engajamento de usuários ocorre em diferentes plataformas de mídias sociais e como isto afeta o conteúdo produzido, através da modelagem de tópicos em dados de publicações, comentários e curtidas extraídos de perfis de empresas no *Facebook*, *Instagram*, *Twitter*, *YouTube* e *Reddit*.

Quattrone et al. (2018) propôs a junção de vários métodos para analisar as revisões em plataformas de economia compartilhada, com o objetivos de gerar conhecimento e valor para o setor por meio da classificação e análise semântica de milhares de revisões. E ao analisar mídias sociais em busca de padrões que determinam o comportamento do usuário, Vydiswaran et al. (2018) examinou a viabilidade dos uso dos dados textuais presentes no *Twitter* para determinar padrões alimentícios de usuários. Através da coleta e classificação de dados textuais relacionados a alimentação, verificou-se que é possível utilizar as mídias sociais para determinar tendências, tal como determinar o nível de consumo de alimentos saudáveis e não saudáveis. Outro ponto que merece destaque quanto a busca de padrões é o apresentado por Fang et al. (2016c), que estudaram os fatores que influenciam na percepção de valor em *reviews*. Através de análises de perspectiva dos textos e de seus autores, verificou-se que a legibilidade e sentimento do texto exerce influência significativa na importância que as pessoas atribuem a um *review*.

Há diferentes tipos de plataformas de mídias sociais com diferentes formas de interagir. Xiang et al. (2017) propõem um estudo que examina comparativamente três plataformas de *reviews* de consumidores em termos de qualidade das informações relacionadas às avaliações *online* nesses sites. O objetivo deste trabalho foi fornecer uma base para a compreensão dos desafios metodológicos e para identificar oportunidades para o desenvolvimento da hotelaria e turismo. Foram extraídos 1.491 *reviews* em três das principais plataformas *online* no mundo da área, a saber: *TripAdvisor*, *Expedia* e a *Yelp*. Com as análises das características dos dados, foi verificado que, apesar de grande parte da literatura apresentar os *reviews* como fontes primárias de conhecimento para consumidores, há, no entanto, diferenças significativas nos *reviews* em diferentes plataformas.

Assim como Aldous, An e Jansen (2019), este trabalho apresenta técnicas de mineração de texto provindos de mídias sociais, juntamente com classificações e análises semânticas presentes em Quattrone et al. (2018). E em consonância com o trabalho de Fang et al. (2016a), também são verificados fatores que influenciam na percepção de valor

dos *reviews*, além de medir a viabilidade de determinação de padrões e tendências a partir da coleta e classificação de dados textuais originados de mídias sociais (VYDISWARAN et al., 2018). E os resultados corroboram com a pesquisa de Xiang et al. (2017) que observou a partir de características dos dados de *reviews* que existem diferenças notáveis entre as plataformas de clientes.

5.3 MATERIAIS E MÉTODOS

Nesta seção são apresentados os processos de coleta e análise de dados, baseados nos trabalhos de Xiang et al. (2017), Fang et al. (2016a). Os trabalhos foram escolhidos devido sua relevância em publicações na área de mineração de dados textuais. Na primeira etapa, os dados de reclamações foram coletados. Posteriormente, foram pré-processados e extraídas as métricas textuais. Por fim, foram conduzidos os cálculos de coerência e de modelagem de tópicos.

5.3.1 COLETA DE DADOS E PRÉ-PROCESSAMENTO

Devido à sua importância econômica, foi escolhido para a mineração dos dados o setor de *ecommerce* brasileiro (ABCOMM, 2019). Neste caso, selecionou-se as cinco empresas com maiores números de acessos em suas lojas virtuais (NETRICA, 2020). Os dados coletados incluem nome da empresa, relato da reclamação, tempo de resposta, estado, cidade e data da reclamação.

Todos os dados de reclamações foram pré-processados utilizando os procedimentos de remoção de saudações, de URLs, *stopwords*, números, acentuação e caracteres especiais (CIRQUEIRA et al., 2018). A remoção de saudações e de URLs significa que toda *string* que representa uma saudação (e.g. "Olá", "Oi") ou um endereço de algum site (e.g. "www.site.com.br") foram removidos. Da mesma forma, ocorreu com números, acentuações e caracteres especiais, visto que são desnecessários para as análises. Palavras que são consideradas *stopwords*, que não contribuem para o significado do texto, também, foram removidas (e.g. "e", "de", "em"). Este processo é importante, uma vez que garante uma maior qualidade dos dados para executar as análises propostas.

5.3.2 EXTRAÇÃO DE CARACTERÍSTICAS TEXTUAIS

As características textuais são importantes para detectar padrões que possibilitem a identificação e comparação de autores de textos em diferente locais (HIRSCH et al., 2017). Os dados textuais das reclamações foram processados e avaliados, levando em consideração a quantidade de caracteres de cada reclamação e do grau de legibilidade através do *Flesch Readability Ease Score (FRES)*. O FRES é uma fórmula desenvolvida por Rudolph Flesch em 1948 utilizada para calcular o grau de legibilidade de um texto (FLESCH, 1948). O

valor de FRES é obtido por meio da seguinte fórmula (OTHMAN et al., 2012):

$$FRES = 206.835 - (1.015 \times ASL) - (84.6 \times ASW) \tag{5.1}$$

A variável *ASL (Average Sentence Length*, em português "tamanho médio da sentença") é obtida a partir da divisão do número de palavras pelo número de sentenças. O *ASW (Average number of Syllables per Word*) é obtido a partir do número médio de sílabas pelo número de palavras. Os valores de FRES representam o grau de legibilidade de um texto e pode variar dentro do intervalo de zero a cem, sendo que quanto maior for o valor mais fácil é a leitura de um texto (OTHMAN et al., 2012). Esses valores são classificados de acordo com a Tabela 13, construída a partir dos trabalhos de (PORTO et al., 2014; OTHMAN et al., 2012; HIRSCH et al., 2017).

Tabela 13 – Classificação dos graus de legibilidade

Classificação	Grau
1º ao 5º ano	75 à 100
6º ao 9º ano	50 à 75
Ensino Médio e Nível Superior	25 à 50
Textos Acadêmicos	0 à 25

5.3.3 COERÊNCIA E MODELAGEM DE TÓPICOS

A tarefa de determinar o melhor número de tópicos em modelagens de tópicos têm se mostrado muito subjetiva em diversos trabalhos na literatura. Visando evitar este viés, foi utilizado um método apresentado por Damani (2013), Fang et al. (2016b), Aldous, An e Jansen (2019) que possibilita a avaliação da modelagem de tópicos por meio do cálculo da coerência de cada tópico. A etapa inicial na análise de coerência consiste em calcular para cada tópico o *Pointwise Mutual Information (PMI)*, também conhecido como similaridade semântica (Equação 5.2). A etapa final na análise de coerência ocorre por meio do cálculo da média dos resultados do PMI, Equação 5.3, através do qual é obtido o valor da coerência do tópico analisado *C(t)* .

$$PMI(w_i w_j) = log \frac{p(w_i w_j)}{p(w_i) \times p(w_j))} \tag{5.2}$$

$$C(t) = \frac{1}{\sum_{m=1}^{n-1}} \sum_{i=1}^{n} \sum_{j=i+1}^{n} PMI(w_i w_j) \tag{5.3}$$

A coerência média de um modelo é dada a partir da pontuação de coerência em todos os tópicos (ALDOUS; AN; JANSEN, 2019). De acordo com (FANG et al., 2016b), usuários finais normalmente estão interessados apenas nos tópicos mais coerentes de um modelo, em vez do modelo inteiro com todos os tópicos. Para isso, dado um modelo com um total de k tópicos, seleciona-se uma quantidade n (*coerência@n*) de tópicos de maior

coerência e calcula-se a média destes valores. A fim de evitar o viés neste cálculo, o ideal é comparar as médias de diferentes valores de n e ampliar o espaço de busca pelos tópicos mais coerentes do modelo.

A identificação de tópicos nas reclamações foi realizada através do uso do algoritmo *Latent Dirichlet Allocation (LDA)* (LI et al., 2019). A implementação deste algoritmo foi realizado por meio da biblioteca *scikit-learn*[2] da linguagem Python.

O LDA é um modelo bastante utilizado por estudos relacionados na literatura (GENC-NAYEBI; ABRAN, 2017), sendo que diversos trabalhos demonstraram a eficácia do modelo na extração de tópicos em diferentes conjuntos de dados textuais (CHARLES-SMITH et al., 2015; LI et al., 2019; FANG et al., 2016a). Na execução, o algoritmo assume que existem estruturas de tópicos que não estão visíveis no conjunto de dados e usa a coocorrência de palavras observadas em diferentes instâncias (GÓMEZ et al., 2015). Dado um conjunto de dados textuais, a execução deste algoritmo retorna uma lista de tópicos associados a um conjunto de palavras. Este conjunto de palavras são associados ao tópico por meio de pesos e uma lista de registros (no caso deste trabalho, reclamações) com um vetor de valores de peso, que definem a probabilidade de haver um documento que contém um tópico específico (XIANG et al., 2017).

5.4 RESULTADOS

Foram coletados dados de empresas que atuam no setor de *ecommerce* e que tinham reclamações de consumidores nas duas plataformas. No total foram coletadas, no período de 26/01/2019 a 26/01/2020, 289.566 reclamações nas duas plataformas. Na Figura 11 são apresentadas as taxas de reclamações proporcionais ao tamanho da população de cada estado. Sendo os estados da região sudeste tem maiores taxas de reclamações por habitante.

Na Tabela 14 são expostos a quantidade de reclamações coletadas, juntamente com a taxa de resposta e de resoluções fornecidas pelas plataformas no período de 6 meses. Ao observar a tabela, é possível constatar que o CGOV ter menor quantidade de reclamações quando comparado com o RA, porém, proporcionalmente no CGOV há uma alta taxa de respostas às reclamações por parte das empresas analisadas.

As plataformas analisadas são similares em termos de propósitos e conteúdo dos textos. Entretanto, há particularidades em cada plataforma que as diferem e influenciam as reclamações e taxa de resposta. O RA é aberto para reclamação de todas as empresas disponíveis no mercado, qualquer cliente pode acessar a plataforma e enviar seu relato. Já o CGOV só aceita reclamações de empresas que aderiram voluntariamente e se cadastraram

[2] https://scikit-learn.org/stable/

Tabela 14 – Número de reclamações por empresa em cada plataforma.

	Consumidor.gov			ReclameAqui		
	Reclamações	Respondidas	Resolvidas	Reclamações	Respondidas	Resolvidas
Americanas.com	16.456	99,9%	67,8%	33.297	95,9%	85,8%
Magazine Luíza	5.694	98,7%	79,8%	50.493	94,2%	90,2%
Mercado Livre	26.346	100%	77,9%	71.018	0%	0%
Netshoes	4.377	99,8%	84,5%	47.237	0%	0%
Submarino	5.169	99,9%	75,5%	29.479	93,8%	84,2%

na plataforma. Isto tem impacto na quantidade de empresas presentes nas plataformas e é o principal motivo de haver reclamações sem resposta e resolução no RA, Tabela 14.

As métricas textuais são pontos que destacam as diferenças entre as plataformas, tal como a *readability* e tamanho das reclamações. Na Figura 10 são apresentadas as distribuições do tamanho das reclamações do CGOV e RA. A partir dos dados expostos, foi possível verificar que os relatos de problemas no CGOV (em azul) tem maior variação no tamanho dos relatos no RA (em vermelho), indicando a presença de duas bases de usuários com características distintas nas duas plataformas.

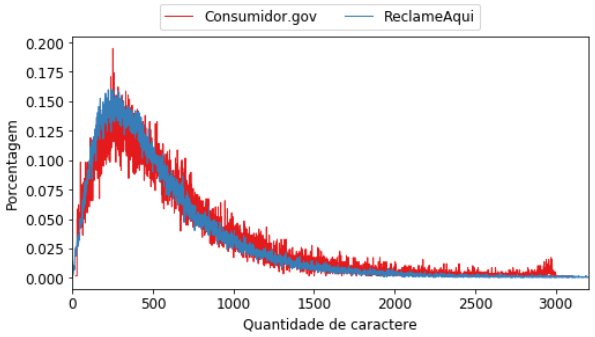


Figura 10 – Comparação do tamanho das reclamações nas duas plataformas.

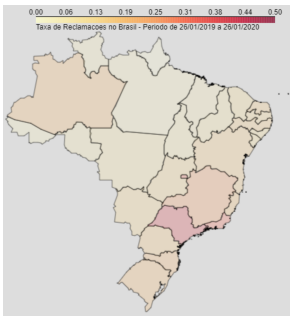


Figura 11 – Taxa de distribuição das reclamações no Brasil.

Esta diferença fica mais evidente na Figura 12, que representa todas as reclamações coletadas das empresas nas duas plataformas divididas em inferidos níveis de escolaridade. A segregação das reclamações em níveis de escolaridade nos revelam um pouco mais sobre perfis dos consumidores de cada plataforma de reclamação online. A seguir são descritas as caraterísticas encontradas. No caso do CGOV identificou-se como principal grupo as pessoas cuja escolaridades é de 6º à 9º ano do ensino fundamental, seguido por "Ensino Médio e Nível Superior", "Textos Acadêmicos" e, por fim, pessoas com 1º à 5º ano do ensino fundamental. No caso do RA, encontrou-se um público de maior nível de escolaridade, sendo o principal grupo de pessoas identificados como que cursaram o "Ensino Médio e Nível Superior", seguido por textos de nível "Textos Acadêmicos", 6º à 9º ano e 1º à 5º ano do ensino fundamental. Pode-se destacar que nas duas plataformas, o menor grupo de pessoas que realizam reclamações são os que tiveram poucos anos de ensino formal, de 1º à 5º ano.

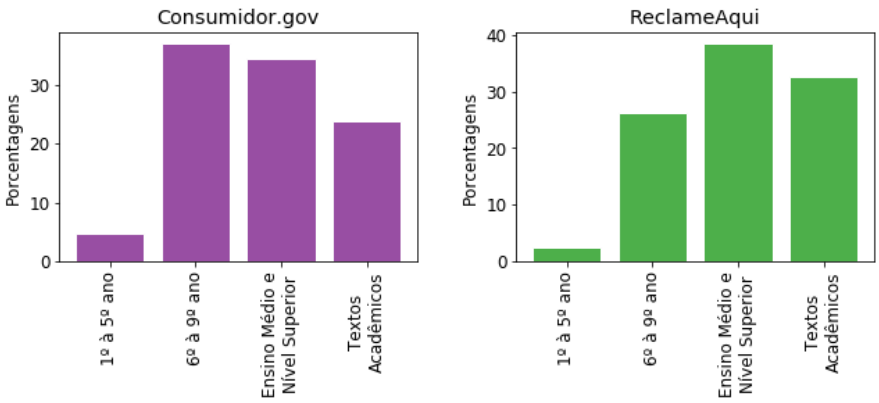


Figura 12 – Comparação no nível de escolaridade entre as plataformas.

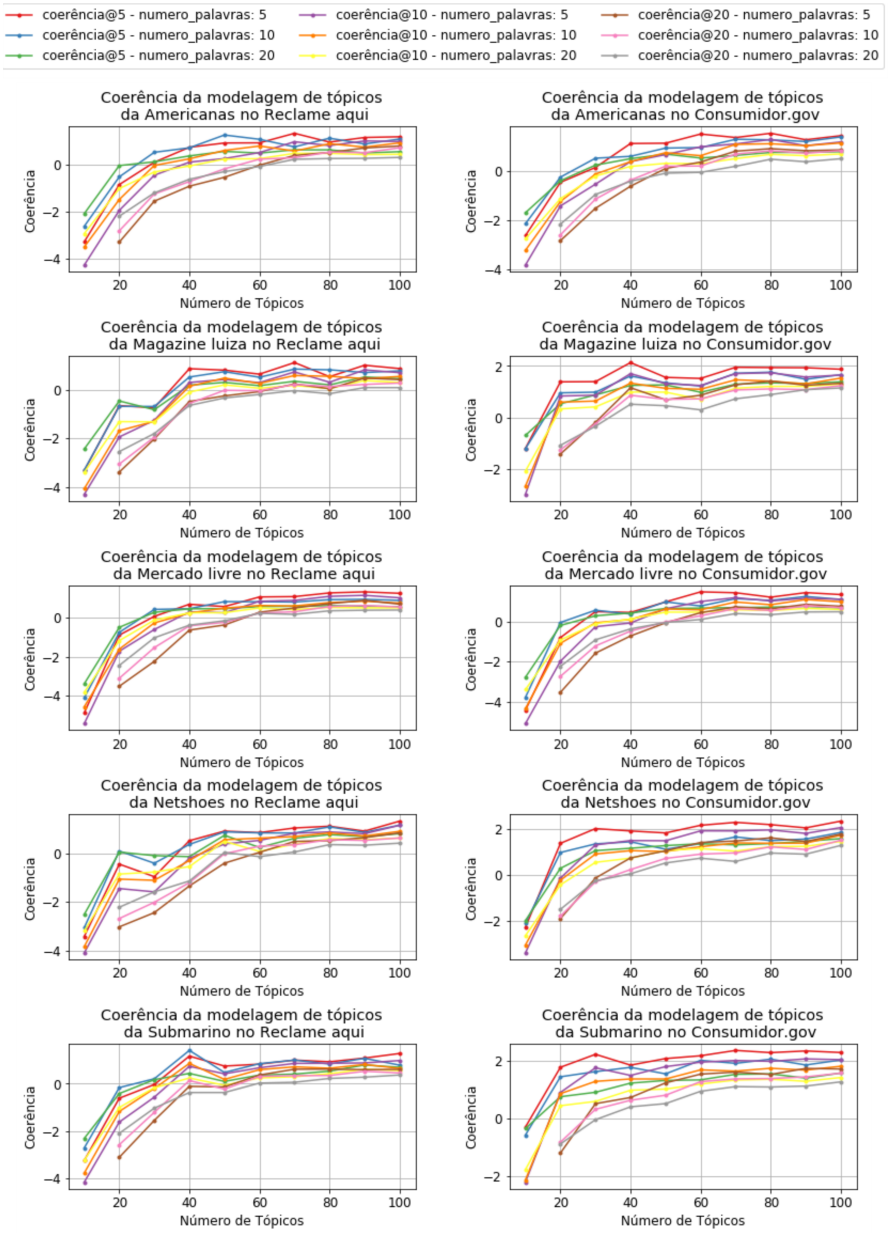


Figura 13 – Coerência da modelagem de tópicos.

Os cálculos de coerência das empresas em diferentes cenários são necessários para se ter uma modelagem de tópicos mais acurada e com maior nível de confiabilidade para determinar os principais tipos de problemas. Na Figura 13 são apresentados os resultados dos cálculos de coerência para cada empresa a partir de diferentes cenários que mesclam: i) plataforma; ii) quantidade de tópicos; iii) número de palavras por tópicos. Os maiores valores de coerência são considerados melhores e mais aptos a serem utilizados. Cada linha dos gráficos indica os resultados para um cenário diferente *(e.g. "coerência@5: numero_palavra:5"indica que o cálculo foi realizado para os 5 principais tópicos que continham 5 palavras)*. A fim de determinar os melhores parâmetros a serem utilizados, os tópicos foram modelados considerando os seguintes valores de número de tópicos de *[10, 20, 30, 40, 50, 60, 70, 80, 90, 100]* variando a quantidades de palavras por tópico em *[5, 10, 20]*.

Ao final do processo, foi observado que houve variação nas quantidades ideais de tópicos modelados, tópicos principais e números de palavras para cada empresa e plataforma analisada. Na Tabela 15 são apresentados os valores obtidos para cada empresa, sendo: "Nº tópicos" - representa a quantidade de tópicos modelados pelo algoritmo; "Nº tópicos principais" - a quantidade de tópicos com maior coerência média que deverão ser utilizados nos resultados; "Nº palavras" - representa a quantidade de palavras para um tópico ter a melhor coerência.

Conforme pode ser observado, todas as empresas nas duas plataformas obtiveram *5* como número de tópicos e de palavras ideais para serem utilizados na modelagem, com exceção do "Submarino" que tem *10* como número de palavras ideais.

Tabela 15 – Melhores valores para modelagem de tópicos nas plataformas analisadas.

	Reclame Aqui			Consumidor.gov		
	Nº tópicos Modelados	Nº tópicos Principais	Nº palavras	Nº tópicos Modelados	Nº tópicos Principais	Nº palavras
Americanas.com	70	5	5	60	5	5
Magazine Luíza	70	5	5	40	5	5
Mercado Livre	90	5	5	60	5	5
Netshoes	100	5	5	100	5	5
Submarino	40	5	10	70	5	5

A Tabela 16 apresenta os principais tópicos das reclamações nas duas plataformas, que representam os cinco principais tipos de problemas com maior valor e importância dentre os demais, considerando cada plataforma.

As células T1 e T5 do RA foram mescladas pois deram origem ao mesmo tópico "Entrega". É importante ressaltar que cada tópico representa um elemento importante para a análise de reclamações, pois fornece bases para o entendimento do contexto da empresa. Os dados estão organizados na tabela de forma a representar a importância através da ordem sequencial dos valores de coerência, sendo que na linha T1 estão os tópicos

Tabela 16 – Tópicos modelados a partir das reclamações nos sites Consumidor.gov e ReclameAqui.

	Consumidor gov				
	Americanas.com	Magazine Luiza	Mercado Livre	Netshoes	Submarino
T1	Cashback	Assinatura	Resolução de conflitos	Produtos	Produtos
T2	Atendimento	Expectativas	Publicidade	Entrega	Promoção
T3	Cadastro	Transporte	Custos	Atendimento	Devolução
T4	Solicitações fundamentadas	Atendimentos	Produtos	Chuteira Infantil	Entrega
T5	Fraude	Defeitos	Serviço	Desistência	Compra

	Reclame Aqui				
	Americanas.com	Magazine Luiza	Mercado Livre	Netshoes	Submarino
T1	Busca	Entrega	Resolução de conflitos	Entrega	Entrega
T2	Produto	Cobrança	Venda	Devolução	Reembolso
T3	Entrega	Produto	Cancelamento	Acesso	Produto
T4	Atendimento	Retorno produto	Promoção	Cobrança	Desistência
T5	Reembolso	Atendimento	Cobrança	Enganação	

mais representativos e na linha T5 estão os menos representativos dentre os principais. Os principais tipos de reclamações variam de acordo com a plataforma, e a seguir são exemplificadas essas diferenças obtidas a partir dos tópicos mais coerentes:

- A empresa Americanas.com tem o ”CashBack” como problema principal no CGOV, o que indica que os usuários da plataforma têm mais problemas com este produto da empresa. Porém, no RA a principal reclamação é sobre a ”Busca”, que pode estar relacionado com os mais variados serviços da empresa, o que significa que é necessário investigar quais e como os serviços lidam com a busca dos consumidores.

- A empresa Magazine Luiza tem o tópico ”Assinatura” como problema principal no CGOV, onde o grupo de clientes que são usuários do CGOV tendem a ter mais problemas com os produtos com assinatura. Para o grupo de usuários do RA a principal reclamação é sobre a ”Entrega”, e devido as características da empresa essas reclamações estão extremamente relacionadas a entrega de produtos.

- A empresa Mercado Live tem o tópico ”Resolução de Conflitos” como problema principal no CGOV e no RA, o que indica que independente do grupo de cliente que utilizam as plataformas a principal reclamação é a mesma, indicando que o problema é recorrente e afeta muitos clientes.

- A empresa Netshoes tem o ”Produto” como problema principal no CGOV e ”Entrega” como principal reclamação no RA. Estes tipos de reclamação são diferentes e devem ser tratadas individualmente, mas devido ao modelo de negócio da empresa estão muito relacionadas entre si.

- A empresa Submarino tem o "Produto" como problema principal no CGOV e "Entrega" como principal tipo de reclamação no RA. Porém, devido ao modelo de negócio da empresa, esses tipos de reclamações não necessariamente estão relacionados entre si.

Os achados aqui apresentados se mostram valiosos e com grande possibilidade de orientar a atuação de empresa nas duas plataformas. Através dos mesmos, é possível obter uma ordem sequencial de problemas recorrentes que desgastam a relação entre empresa e consumidor. Além disto, devido a presença de grupos distintos de usuários nas plataformas, é possível estruturar estratégias de soluções que melhor se adéquem a cada grupo.

5.5 CONSIDERAÇÕES FINAIS

Neste artigo, foram analisadas reclamações de cinco empresas do setor de *ecommerce* brasileiro em duas plataformas de reclamações *online*. O objetivo destas análises é prover uma base para a compreensão das oportunidades que elas oferecem para orientar a tomada de decisões em negócios. Foram aplicadas técnicas de mineração de dados textuais a fim de explorar as plataformas e extrair suas características. Os resultados mostram que apesar de as plataformas terem a mesma finalidade, elas divergem em grupos de usuários, confiabilidade de solução, conteúdo e principais tópicos das reclamações. Ademais, as análises se mostraram viáveis mesmo com grandes quantidades de dados, o que torna possível a antecipação e correção de problemas de forma rápida e eficaz, e abre a possibilidade de fazer amplas análises de mercado e concorrência de acordo com a perspectiva de cada plataforma.

Diante disto, a principal contribuição deste estudo está na provisão de uma abordagem de análise de reclamações para as duas plataformas que permite identificar diferentes formas de atuação que uma empresa deve seguir para melhorar o relacionamento com seus clientes. No entanto, não foram realizadas análises de significância estatística nas comparações apresentadas neste estudo. E as análises e resultados obtidos não incluem aplicações práticas nas empresas. Como trabalhos futuros, pretende-se expandir as análises de modo a incluir análise de significância estatística nas comparações, conduzir um estudo prático em empresas que explore tanto as reclamações quanto a respostas.

5.6 AGRADECIMENTOS

Os autores agradecem ao Serviço de Intercâmbio Alemão (DAAD), à Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA) e a Universidade Federal do Oeste do Pará (UFOPA) por meio da Pró-Reitoria da Cultura, Comunidade e Extensão (PROCCE), pelos fomentos destinados à execução desta pesquisa.

6 ARTIGO - GAINING INSIGHTS ON STUDENT SATISFACTION BY APPLYING SOCIAL CRM TECHNIQUES FOR HIGHER EDUCATION INSTITUTIONS

ABSTRACT

Social Media and Customer Relationship Management (CRM) are already widely used in business settings, but other non-commercial sectors started only recently to adopt them. Among them are Higher Education Institutions (HEIs). Even though research shows positive effects on the quality of services, student satisfaction, and attractiveness towards international students, the adoption is very low. This research in progress reviews the state of research about Social CRM in HEIs and gives an example of the potential of social media for CRM approaches of HEIs by applying Social CRM concepts and techniques for better understanding the negative service experiences of students. By applying analytical Social CRM techniques on large amounts of User-Generated-Content (UGC) in complaint platforms the paper gives insights into problem chains inaccessible with manual methods. Based on the scarce research about Social CRM as well as the demonstrated potential of social media for CRM strategies of HEIs, this paper concludes with a call for further research on Social CRM in HEIs.

Keywords: CRM, Social Media, Student Satisfaction, Complaint Management, Text Mining, Topic Modeling

6.1 INTRODUCTION

Social media gain importance in higher education not only since the COVID-19 pandemic prevents personal contact between students and professors. Researchers investigate the use and potential of social media in specific areas for several years. For example, as support in lectures (DYSON et al., 2015), support in hybrid learning environments (LI, 2014), for marketing purposes (KARNA; SUPRIANA; MAULIDEVI, 2015), or examined how students use them for study purposes (HRASTINSKI; AGHAEE, 2012). However, their potential for building relationships with students and managing the student life cycle was only sparsely examined. A reason may be the fact that many universities do not compete over the increasing number of students or that the relationship between students and professors is more focused on teaching and students are expected to actively manage their study program. But as an increasing amount of young people, and so students, use social media more frequently, they also use them during their student lifecycle and expect higher education institutions (HEIs) to do the same.

Applying concepts and techniques developed for the management of customer relationship management (CRM) with the help of social media could help to manage this transformation. As in the industry, wherein many cases until 2005 the enterprise still owned the customer experience (GREENBERG, 2009), the student experience is still often owned by HEIs. Universities already pay attention to social media to maintain communication (BONSóN et al., 2012) but use them only for dedicated tasks. Greenberg (2010) pointed out that enterprises need first to figure out the business models, applications, processes, and social characteristics that are required to actually implement the social CRM before social media customer service begins to happen. This counts now for HEIs as they need to figure out the application areas of social media within the student life cycle and service offerings of the university.

This paper aims to initially explore the relevance of social media by analyzing the use of such channels in critical steps within the student life cycle, namely the handling of complaints as part of the service phase. As visible in a famous example of Social CRM (JARVIS, 2005), students will use social media if they are not satisfied with the service experience, regardless of whether HEIs are active on social media or not. Following the concept of Social CRM (GREENBERG, 2010; ALT; REINHOLD, 2020), higher education could build up its presence on social media platforms, provide services by using social media as a channel in workflows, learn from the content in social media and use these channels to perform collaborative tasks with students. Actively using social media and providing a satisfying service experience may decrease the number of complaints. As research about the management of complaints by HEIs via social media is scarce, this paper aims to show that students use social media for complaints, that a link between the number of complaints in social media with the active provision and management of

social media by universities exists and to identify the core topics of student complaints. The research questions are:

- (RQ1) Are students using external and public platforms to complain about education-related services along their student life cycle?
- (RQ2) How can we derive information about major service quality issues affecting student satisfaction?
- (RQ3) What types of insights on customer satisfaction can HEIs managers expect from an analysis of external complaints?

The remainder of the paper is structured as follows. First, research about the role of customer satisfaction and Social CRM is reviewed and key elements for assessing customer satisfaction with the help of Social CRM are identified. Second, an experiment demonstrates an approach for analyzing customer satisfaction in higher education. Third, the results from the experiment are discussed, answering the research questions.

6.2 CRM AND SOCIAL CRM IN HIGHER EDUCATION

In this section, we discuss the CRM and Social CRM applied in high education institutions to improves the services e the students' satisfaction. We first outline the effects of the CRM on the service quality and students' satisfaction in HEIs, and before we discuss the potential of Social CRM to understanding customer satisfaction and to managing the service quality.

6.2.1 CRM AFFECTS SERVICE QUALITY AND STUDENT SATISFACTION IN HEIS

The application of CRM concepts in higher education was examined from different perspectives already. Rigo et al. (2016) show that the main principles of CRM can also be applied to the context of HEIs and that HEIs must consider more stakeholders than just students in their CRM approach, calling for also using social media for linking and interacting with numerous stakeholders. Nair, Chan e Fang (2007) point out that HEIs first need to understand the student lifecycle ($Suspect \rightarrow Prospect \rightarrow Applicant \rightarrow Admitted \rightarrow Enrollee \rightarrow Alumni$) before successfully making use of a CRM approach.

By analyzing social media content, HEIs can improve their understanding of the student life cycle and optimize their CRM. Hrnjic (2016) shows that student satisfaction is a good indicator for the successful adoption of CRM for the creation of a student-oriented environment and constantly adapting its processes. Critical elements are the university organization and management of teaching processes, academic staff skills and competencies, management board activities and institutional development, and quality of study materials

used in the classes, and application of learning methods. Seeman e O'Hara (2006) points out that universities that aim to achieve a leadership position in the higher education sector need to show an additional focus on reproducing highly skilled faculty staff that will have a capacity to improve teaching with regard to technology changes and market requirements. It is also important that the university board and leading people such as deans of HEIs understand the strategic dimension of CRM orientation at universities. Both call for HEIs to adopt social media early as new technology and to develop an integrated management approach for social media and CRM. A study from Seeman e O'Hara (2006) shows the handling of students as customers provides a competitive advantage for higher education and enhances a college's ability to attract, retain and serve its customers. The benefits of implementing CRM in a college setting include a student-centric focus, improved customer data and process management, and increased student loyalty, retention, and satisfaction with the college's programs and services. As colleges increasingly embrace distance learning and e-business, CRM will become more pervasive. The COVID pandemic further increased this need. Badwan et al. (2017) confirm this observation by showing that implementing electronic CRM can cause customer satisfaction, loyalty, retention, and high service quality as students pointed to be a customer.

CRM supports the understanding of customer expectations and thus provides a basis for service customization, which in turn can positively affect service quality. Wali e Wright (2016) show that an effective CRM program to improve service quality affects customer satisfaction and even has the ability to induce positive advocacy behavior from its international students. A key element is gaining and understanding customer's experience feedback.

6.2.2 NEW POTENTIALS FOR UNDERSTANDING CUSTOMER SATISFACTION AND MANAGING THE SERVICE QUALITY ARISE FROM SOCIAL CRM

A key element of Social CRM is the interaction with stakeholders that influence the service system of a business and the knowledge derived from this interaction. As Greenberg (2009) points out, the customer becomes the focal point of the ecosystem, and service providers can make use of social media to understand their role in the customer ecosystem. Social CRM provides the means for that, but as Sablan et al. (2017) a Social CRM model for HEI's is virtually non-existent. However, the first research points out, that applications, data, and information, adapted business processes, social media presences are among the critical success factors for social CRM in HEIs.

Following Meyliana, Hidayanto e Budiardjo (2015) many universities started with web 2.0 and social media adoption but focus mainly on real-time events webcast, widgets, and social networks for the users of the university website. The study of Oliveira (2015) shows that Social CRM propels HEI to engage in dialogical conversations and collaborative

relationships. The use of social media platforms, allowing to reshape the HEI-student formal relationship, strengthening educational bonds through the development of dialogs, provides mutually beneficial value and, ultimately, allows for the growth of social and educational communities.

However, only a few examples have further investigated the potential of analytical Social CRM for HEIs, especially for assessing student satisfaction with experience service quality. Budiardjo et al. (2017) show in general by mapping the student lifecycle with CRM processes and features of Social CRM that the latter can support core CRM processes of HEIs, and the application of Social CRM software can support operational, analytical, and collaboration tasks. Karna, Supriana e Maulidevi (2015) show how data analysis can provide further insights on candidates and help to individualize marketing activities. But unlike Social CRM in a business context, the potential of monitoring and mining for HEIs has not been understanding.

6.3 IMPROVING THE UNDERSTANDING OF NEGATIVE SERVICE EXPERIENCES IN HEIS WITH ANALYTICAL SOCIAL CRM TECHNIQUES

This section discusses the use of analytical Social CRM to understand the students' negative experiences in HEIs on Social Media platforms. In the following subsections, we present an overview of the process to analyze from social media platforms, the methodology to performs the analysis, the potential data source, and the potential analysis methods.

6.3.1 COMPLAINT AND SATISFACTION ANALYSIS IN EXTERNAL SOCIAL MEDIA

Following the concept of analytical Social CRM (REINHOLD; ALT, 2011), building up a customer feedback and satisfaction analysis requires accessing the data of social media platforms where students share feedback and opinions. In a second step, this data needs to be turned into knowledge about customer satisfaction by either manual evaluation, observation of keywords, or applying methods for understanding at the semantic level. While the manual evaluation can provide rich insights, the potential amount of data makes it challenging for HEIs. Applying basic analyses such as looking for trending topics and sentiments of students might support HEIs to understand the behavior of their students. Although these analyses are supported by many social media monitoring tools, they have limits in uncovering larger patterns, such as the relationship between raised issues. For understanding previously unknown factors that affect service quality methods of data mining need to be applied.

6.3.2 PROCESS DESIGN

In this scenario, it is critical to map relevant data sources and evaluate all pertinent analyses that improve customer satisfaction. One well-known data analysis methodology for conducting real-world projects is the Cross-Industry Standard Process for Data Mining (CRISP-DM). This methodology proposes a comprehensive process model for carrying out data mining projects. The process is divided into phases of *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, and *Deployment* and is independent of the industry sector and the technology used (WIRTH, 2000).

In this research, CRISP-DM will be applied for the analysis of student complaints in Brazil. HEIs in Brazil provides a good example, because of the availability of well-used independent social media platforms for publicly raising and discussing complaints independently from specific industry sectors. Thus, insights on service quality and customer satisfaction outside of a HEIs direct control are accessible.

6.3.3 POTENTIAL DATA SOURCES

The second phase of CRISP-DM is called "*Data Understanding*", in which it is possible to identify issues on data and provide initial insights into available data. However, there is plenty of platforms with a large volume of User-Generated Content (UGC) available with some challenges such as data diversity, unstructured data, missing data, *etc* (LOBATO et al., 2017). Despite these challenges, such platforms represent an interesting data source for providing business *insights* with reduced costs when compared with customer surveys and other market research strategies (BAHTAR; MUDA, 2016; VERMEER et al., 2019). In the context of Brazilian HEIs relevant data sources are for example:

- ***Consumidor.gov*** (CONSUMIDOR.GOV, 2021) - Contains data referring to complaints reports about universities. The data types include *strings* (raw text), *timestamp* (Date and Time), and numerical features. This platform does not provide an *Application Programming Interface* (API) for data acquisition. However, considering that it is a public platform with open data (considering Brazilian Legislation), the data can be requested using the national accountability platform;
- **Ministry of Justice and Public Security** (MJSP) (MJSP, 2021) - The MJSP portal provides details about the complaints published on the Consumidor.gov platform. Large parts of this data are in textual format. This platform provides an interface to get the available data;
- **Data from Higher Education Census** (INEP, 2021) - It is data related to the 2019 higher education census, and contains information about Brazilian students, courses, and high educational institutions. Most of the data are numerical features, requiring a mapping procedure with a data dictionary;

- **Data from Social Media platforms** - Contains information related to followers, publications, and Comments. However, it is important to verify the privacy rules and API restrictions of each platform.

6.3.4 POTENTIAL METHODS FOR ANALYSIS

"Data preparation" and "Modeling" phases of the CRISP-DM methodology represent the core for Computer Scientists (WIRTH, 2000). The pre-processing methods aim to improve data quality, consequently, improving the reliability of the results (LOBATO et al., 2017). Considering that most UGCs are encoded in text, there is a lot of effort on the development of *text mining* methods and pipelines (FERNANDES et al., 2020). The canonical pipeline includes the application of specific pre-processing methods for text data (ALLAHYARI et al., 2017; GARCÍA et al., 2015; HAN et al., 2016). It can be followed by a data fusion/enrichment step, aiming to combine different data sources, which can reduce bias and uncertainties, increase reliability, and improve accuracy (ZHANG, 2010; QI et al., 2020).

The data preparation is following by the data analysis itself. There are common tasks on Text Mining. For instance, **topic modeling** allows the identification of the most frequent topics and their terms, which would be difficult to discover through a manual process that (ROBERTS et al., 2014; PARK; CHAE; KWON, 2018). The topics discovered that can be modeled as a *graph*, indicating the relationship between the topics, and allowing the identification of terms chains (EASLEY; KLEINBERG, 2010). In addition to that, **sentiment analysis** can be used to extract the feeling through automatic polarity detection (JO; FERREIRA, 2017; RAVI; RAVI, 2015), and the text quality can be measured by the evaluation of the legibility and by the determination of the quality according to the expected in each phase of regular education (HIRSCH et al., 2017; FLESCH, 1948; OTHMAN et al., 2012).

6.4 DEMONSTRATION

This section presents an analysis of the textual content of students' complaint about two higher education institutions in Brazil, named *University A* and *University B*. These two universities have multiple internal channels of customer services available to students, which include the website online chats, email, telephone and social networks (see Table 17). On social networks, these universities have in general a high number of followers. However, *University A* has a low number of interactions on their content, opposite to *University B* with a very number of interactions. Furthermore, their students are using external platforms for making complaints. As it is known that on average two-thirds of consumers check product, service, and brand evaluations before deciding on the purchase, (HE et al., 2020; CONSTANTINIDES; HOLLESCHOVSKY, 2016) these external complaints can

have an impact on reputation and in consequence the willingness of future students to join the university.

	University A	*University B*
Total of Students (INEP, 2021)	302,841	393,578
Customer Service Channels	- Website - Phone - WhatApp - Social Networks: —— Twitter with 8,804 followers —— Instagram with 62,300 followers —— Facebook with 617,702 followers - Reclame Aqui - Consumidor.gov	- Website - Phone - WhatApp - Social Networks: —— Instagram with 283,000 followers —— Facebook with 1,534,463 followers - Reclame Aqui - Consumidor.gov
Complaints on Consumidor.gov [1]	- Total of Complaints: 2,743 - Solution Rate: 63.7%	- Total of Complaints: 3,965 - Solution Rate: 57.2%
Complaints on ReclameAqui [2]	- Total of Complaints: 37,764 - Solution Rate: 58.6%	- Total of Complaints: 52229 - Solution Rate: 52%

Tabela 17 – Summary information about the universities

Following the methodology described before, the analysis comprised a pre-processing step and the analysis phase, using topic modeling and topic correlation methods. The data were collected on *Consumidor.gov* [3] (CONSUMIDOR.GOV, 2021) from January to March of 2021, with a total of 652 complaints about *University A* and 803 about *University B*. Topic modeling was carried out using Negative Matrix Factorization (NMF) (CHEN et al., 2019) and revealed the ten main topics in complaints about each university (see Figure 14). Some topics are not exclusive, indicating that the two institutions facing the same type of problem, related to the "**Payments**", "**Attestation of degree**", and "**Classroom**".

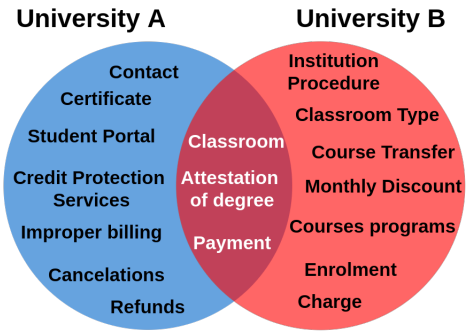


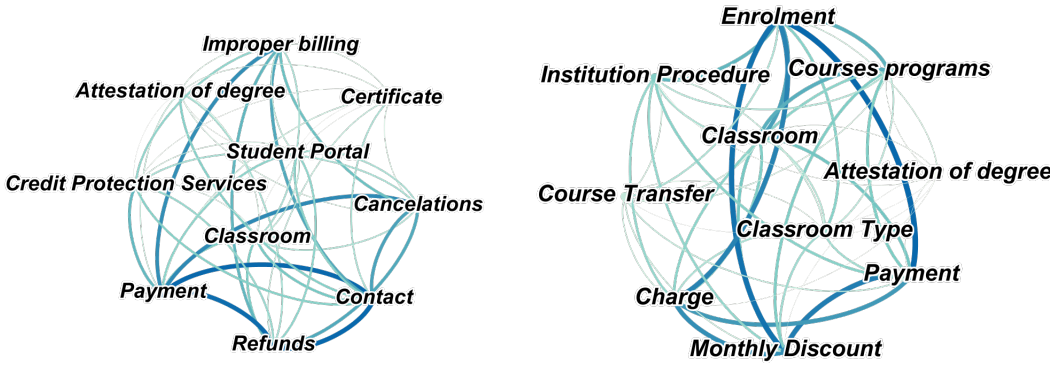
Figura 14 – Main Topics in complaints of the *University A* and *University B*

The correlation of the topics allows the identification of relationship degrees between the topics. Figure 15 was produced using open-source software called *Gephi*[4], and it shows the relationships between most frequent topics/aspects, and the strongest blue tone of the

[3] *Disclaimer: It is important to emphasize that all data collected on Consumidor.gov will be used exclusively for proof-of-concept purposes. The authors have no interest in the commercial use of this data.*

[4] https://gephi.org/

line indicates a stronger degree of correlation. In the topics of *University A* (see Figure 15a) the formation of a chain of main problems refer to **Payments**, **Refunds**, and the **Contact**. Figure 15b shows the chain of main problems for *University B*, which are related to **Payments**, **Charges**, **Monthly Discount**, and **Enrollment**.



(a) *University A*

(b) *University B*

Figura 15 – Topic correlations of complaints

Modeling and correlation of topics give HEIs managers insights on potential service quality issues. But, they are also a starting point for planning actions to solve the problem chains that make up principal causes of dissatisfaction among students. For example, among the problems of *University A* are several types of issues related to finance and customer service. They indicate students are dissatisfied because of financial issues together with the university customer Service. For solving these problems, more detailed analysis can be made on complaints related to these areas, to understand and identify products and services that are the source of the problems.

The demonstrated analysis provides insights based on a large number of complaints related to two universities. Such aggregated insights on major students' problems and their relations can support decision-making and as a basis for executing efficient actions for solving the chain of issues at the root of student complaints. Thus they can contribute to CRM efforts by helping and assisting in developing better products and services centered on the expectations and needs of students.

6.5 CONCLUSION AND IMPLICATIONS

This paper investigates the potential of Social Media for improving service quality and student satisfaction as part of the ever-increasing adoption of CRM in HEIs. Building upon recent research about critical success factors of CRM for HEIs, the paper demonstrates how universities can improve their service system with insights from Social CRM analytics. Based on an examination of selected universities in Brazil this research shows that students use external platforms to complain about their service experience, regardless if

the university provides its own platforms (RQ1). With the application of data analysis and mining techniques employed in social CRM (FERNANDES et al., 2020; REINHOLD; ALT, 2011) in the enterprise context, major issues can be efficiently summarized (RQ2) from large data volumes. HEIs can learn from this data (RQ3) about the service quality experience of students and factors that negatively affect student satisfaction. It is also a basis for comparison and evaluation with other HEIs and an opportunity to identify more successful approaches. In addition, from the results produced by the text mining tasks used in the demonstration, it was possible to observe that complaints are a rich source to extract knowledge about students and services offered by universities.

Besides these direct insights, this research provides, on a more general level, a further example of the benefits of Social CRM for HEIs and the importance of further research about the application fields and implementation approaches for them. Only an integrated Social CRM allows to get data from internal and external sources and to use obtained knowledge for improving services. Many insights on the successful integration of Social Media with CRM in enterprises are available and HEIs can build on this extensive knowledge while adapting it to their specific environment.

This paper is an example of the potential of advanced semantic analysis in a limited case. Therefore the paper has limitations. The obtained insight is not generalizable and demonstrates only the insights that can be acquired by using the proposed approach. Also, the insights have not been evaluated with managers in HEIs in terms of novelty and meaningfulness compared to other sources. Therefore, this research is planned to be expanded in the next step, by using the approach for a larger set of universities, expanding the insights by comparing the results in a larger set of universities and evaluating results of managers of HEIs regarding their contribution towards a better understanding of student satisfaction and potential improvements in service quality.

7 CONSIDERAÇÕES FINAIS

Esta dissertação foi composta por um compêndio de estudos relacionados à análise inteligente de mídias sociais para potencializar a gestão de relacionamento com clientes. Diferentes abordagens de melhoria foram avaliadas e testadas neste trabalho com o objetivo de melhorar os sistemas de Social CRM. Os quatro artigos apresentados demonstraram que o objetivo precípuo do projeto foi alcançado.

Além dos estudos apresentados nos capítulos 3, 4, 5 e 6, outros trabalhos foram publicados, os quais representam contribuições marginais a este trabalho e foram incluídas nos Apêndices A e B. A seguir, cada um desses artigos é brevemente descrito:

1. Artigo *"Ferramentas para Análise de Mídias Sociais: Um levantamento sistemático"* - Este estudo teve como objetivo principal mapear o estado da arte e o estado da prática em relação à análise de mídias sociais, para identificar as bases, métodos e ferramentas mais utilizadas pelos pesquisadores em suas análises. Os resultados deste estudo têm potencial de orientar futuras pesquisas, pois poderão ser conduzidas de maneira mais objetiva, de modo a economizar tempo e esforços nas fases de escolha de fontes de dados, além de apontar para os pesquisadores as tecnologias de análise de dados mais utilizadas. O manuscrito completo deste artigo está disponível no Apêndice A desta dissertação.

2. Artigo *"Social CRM as a business strategy: developing dynamic capabilities of Micro and Small Enterprises"* - Este estudo teve por objetivo principal analisar o uso de Social CRM por Micro e pequenas empresas, a fim de identificar oportunidades de intervenção. A execução deste estudo ocorreu por meio da aplicação de um questionário em 31 empresas e através de estudos de casos conduzidos em quatro empresas. Os resultados permitiram a identificação de pontos comuns de falha na utilização de Social CRM por micro e pequenas empresas, e que a utilização do Social CRM ainda é pouco explorada por essas empresas. O manuscrito completo deste artigo está disponível no Apêndice B desta dissertação.

De acordo com os resultados deste estudo, acredita-se que as abordagens propostas e avaliadas nos artigos contribuem para a melhoria de sistemas de Social CRM por meio:

- Da automatização da avaliação da efetividade de comunicações empresariais. Pois eliminam o viés da avaliação manual e permite a análise de grandes quantidades de dados de forma eficiente e barata;

- A partir da análise do potencial, da identificação das oportunidades e desafios e da demonstração da aplicabilidade em um cenário real, foi atestada e validada a viabilidade do uso de reclamações online para o aprimoramento de sistemas de Social CRM;
- Do auxílio na escolha da base de dados e métodos para as análises. O que otimiza o início do processo de extração de conhecimento das mídias sociais;
- Do entendimento dos principais problemas enfrentados por micro e pequenas empresas na utilização do Social CRM. O que permite a implementação personalizada dos sistemas de Social CRM por esse setor da economia.

Além do impacto científico demonstrado pela publicação dos seis artigos supracitados, o presente trabalho também contribui para com o estado da técnica, fornecendo materiais e métodos para a condução de análises de dados de mídias sociais por outros pesquisadores, por meio da disponibilização de uma base de dados de treinamento para auxiliar na classificação de conteúdos empresariais e de um *pipeline* de análise que inclui diversos métodos e otimiza o processo de extração de conhecimentos das mídias sociais. A base de dados foi tornada pública[1] (SOUSA; JUNIOR; LOBATO, 2021) e o *pipeline* será submetido ao registro de *software* junto ao Instituto Nacional da Propriedade Industrial (INPI).

7.1 TRABALHOS FUTUROS

Como trabalhos futuros, planeja-se a adição e aprimoramento dos métodos presentes no pipeline de análise. Para isso, serão estudadas e avaliadas as possibilidades de melhorar o processamento da linguagem natural através do uso de *reinforcement learning*, *word-embeddings*, *deep-learning* e algoritmos como *Bidirectional Encoder Representations from Transformers* (BERT), *Convolutional Neural Network* (CNN), *Hierarchical Text Classification* (HTC), dentre outros.

Planeja-se também a construção da documentação do pipeline de análises de dados. Isso permitirá que a manutenção e o uso da ferramenta sejam facilitados, uma vez que todos os métodos e formas de uso foram descritos e exemplificados. Além disso, essa documentação permitirá que novos membros sejam integrados à equipe de desenvolvimento de forma simples e rápida, facilitando a ampliação do número de métodos e análises disponíveis no *pipeline*.

Por fim, planejamos realizar uma análise de viabilidade de mercado para monetização do pipeline. Este processo permitirá o mapeamento e avaliação de potenciais consumidores, bem como a delimitação de oportunidades e desafios para a criação de um

[1] https://doi.org/10.5281/zenodo.5113266

produto a partir do pipeline construído neste trabalho. Desta forma, pode-se elaborar um plano de negócios e implementar um protótipo de produto considerando as reais necessidades e desafios impostos pelo mercado.

7.2 DIFICULDADES ENCONTRADAS

Durante o desenvolvimento deste projeto, alguns obstáculos e dificuldades foram encontrados. No início dos estudos foram utilizadas ferramentas implementadas com *JavaScript*[2] e com a biblioteca *Selenium*[3] do Python, porém observou-se que essas ferramentas não eram eficientes e demandavam muito tempo para coletar e organizar os dados, com o avanço dos estudos, as dificuldades técnicas foram superadas e novas ferramentas de coleta foram construídas a partir da biblioteca *Requests* [4] do Python, o que tornou o processo mais rápido e otimizado. Além disso, a conexão com a internet foi um problema constante na coleta inicial de dados, diferentes soluções foram avaliadas e testadas, como a utilização da plataforma *Kaggle*[5], que permitia a execução por 6 horas, e normalmente, o processo era interrompido antes do fim. Assim, foi necessária a contratação de um servidor virtual privado com acesso ininterrupto à internet para a execução das ferramentas de coleta de dados necessárias às análises.

Após a coleta de dados, a execução das análises foi dificultada pelos recursos computacionais disponíveis. Essa limitação teve um grande impacto nas tarefas de pré-processamento, devido ao tamanho do conjunto de dados, esses processos exigiram muitos recursos computacionais e demoravam horas para serem concluídos, às vezes, era necessário dividir os dados em conjuntos menores para que pudessem ser executados dentro das 6 horas fornecidas pela plataforma *Kaggle*. Da mesma forma, as tarefas de avaliar o melhor número de tópicos para modelagem foram impactadas, mas neste processo, não foi possível dividir o banco de dados em conjuntos menores, pois a avaliação deveria ocorrer considerando todos os dados, portanto esse processo ficou em execução por vários dias até a sua conclusão.

Além destas, encontramos dificuldades na validação dos resultados obtidos nas análises de dados. Este estudo envolveu a análise de dados de várias empresas de diferentes setores, mas não foi possível estabelecer contato com elas para que os resultados fossem validados e utilizados internamente. Além disso, não foi possível estabelecer contato direto com especialistas em Social CRM que atuam no Brasil, e os resultados tiveram que ser avaliados e validados por especialistas que atuam no exterior, e devido a não familiarização com o domínio de aplicação e com o contexto brasileiro, este processo foi realizado em diversas fases, que incluíram a contextualização e compreensão dos resultados.

[2] https://developer.mozilla.org/pt-BR/docs/Web/JavaScript
[3] https://selenium-python.readthedocs.io/
[4] https://docs.python-requests.org/en/master/
[5] https://www.kaggle.com/

REFERÊNCIAS

ABCOMM. Crescimento do e-commerce no Brasil. Disponível em: https://abcomm.org/noticias/crescimento-do-e-commerce-no-brasil. Acesso em: 28/03/2020. 2019. Disponível em: <https://abcomm.org/noticias/crescimento-do-e-commerce-no-brasil/>. Citado na página 69.

AGHASIAN, E.; GARG, S.; MONTGOMERY, J. An automated model to score the privacy of unstructured information—Social media case. *Computers and Security*, Elsevier Ltd, v. 92, p. 101778, 2020. ISSN 01674048. Disponível em: <https://doi.org/10.1016/j.cose.2020.101778>. Citado na página 26.

Agência Nacional de Telecomunicações (ANATEL). *Anatel - Panorama*. 2020. Disponível em: <https://www.anatel.gov.br/paineis/acessos/panorama>. Citado 2 vezes nas páginas 24 e 52.

AHMAD, S. N.; LAROCHE, M. Analyzing electronic word of mouth: A social commerce construct. *International Journal of Information Management*, Elsevier Ltd, v. 37, n. 3, p. 202–213, 2017. ISSN 02684012. Disponível em: <http://dx-.doi.org/10.1016/j.ijinfomgt.2016.08.004>. Citado 4 vezes nas páginas 25, 32, 39 e 51.

ALDOUS, K. K.; AN, J.; JANSEN, B. J. View , Like , Comment , Post : Analyzing User Engagement by Topic at 4 Levels across 5 Social Media Platforms for 53 News Organizations. In: *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media*. [S.l.: s.n.], 2019. p. 47–57. Citado 4 vezes nas páginas 53, 67, 68 e 70.

ALEXA. *Alexa - Top Sites in Brazil - Alexa*. 2019. Citado 2 vezes nas páginas 51 e 55.

ALI, M.; JOORABCHI, M. E.; MESBAH, A. Same App, Different App Stores: A Comparative Study. *Proceedings - 2017 IEEE/ACM 4th International Conference on Mobile Software Engineering and Systems, MOBILESoft 2017*, p. 79–90, 2017. Citado na página 51.

ALLAHYARI, M. et al. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. 2017. Citado 3 vezes nas páginas 28, 29 e 83.

ALMEIDA, G. R. de; CIRQUEIRA, D. R.; LOBATO, F. M. Improving social crm through eletronic word-of-mouth: a case study of reclameaqui. In: SBC. *Anais Estendidos do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web*. [S.l.], 2017. p. 107–110. Citado 2 vezes nas páginas 51 e 53.

ALMEIDA, G. R. T. de; LOBATO, F.; CIRQUEIRA, D. Improving Social CRM through eletronic word-of-mouth: a case study of ReclameAqui. *XIVWorkshop de Trabalhos de Iniciação Científic*, 2017. Citado 2 vezes nas páginas 24 e 39.

ALMEIDA, J. S. D. N. G. R. T. de; LOBATO, F. M. F.; JUNIOR, A. F. L. J. Melhorando Sistemas de Social CRM por meio de Eletronic Word-of-Mouth. *Revista Eletrônica de Iniciação Científica em Computação*, v. 17, n. 4, 2019. ISSN 1519-8219. Citado na página 67.

ALT, R.; REINHOLD, O. Social Customer Relationship Management (Social CRM). *Business & Information Systems Engineering*, v. 4, n. 5, p. 287–291, 2012. ISSN 1867-0202. Disponível em: <http://link.springer.com/10.1007/s12599-012-0225-5>. Citado 2 vezes nas páginas 23 e 51.

ALT, R.; REINHOLD, O. Social crm: Challenges and perspectives. In: *Social Customer Relationship Management*. [S.l.]: Springer, 2020. p. 81–102. Citado na página 78.

ANATEL. *Movel_Pessoal - Files - ownCloud*. 2020. Disponível em: <https://cloud.anatel-.gov.br/index.php>. Citado 2 vezes nas páginas 13 e 62.

BADWAN, J. J. et al. Adopting technology for customer relationship management in higher educational institutions. IJARW, 2017. Citado na página 80.

BAHTAR, A. Z.; MUDA, M. The Impact of User – Generated Content (UGC) on Product Reviews towards Online Purchasing – A Conceptual Framework. *Procedia Economics and Finance*, v. 37, p. 337–342, 2016. ISSN 2212-5671. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2212567116301344>. Citado 3 vezes nas páginas 51, 67 e 82.

BANKOLE, F. O.; OSEI-BRYSON, K. M.; BROWN, I. The Impacts of Telecommunications Infrastructure and Institutional Quality on Trade Efficiency in Africa. *Information Technology for Development*, v. 21, n. 1, p. 29–43, 2015. ISSN 15540170. Citado na página 52.

BARATA, G. M. et al. Social CRM in Digital Marketing Agencies: An Extensive Classification of Services. *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, IEEE, p. 750–753, 2018. Citado na página 47.

BARRETO, A. M. The word-of-mouth phenomenon in the social media era. *International Journal of Market Research*, SAGE Publications Sage UK: London, England, v. 56, n. 5, p. 631–654, 2014. Citado na página 39.

BELLO-ORGAZ, G.; JUNG, J. J.; CAMACHO, D. Social big data: Recent achievements and new challenges. *Information Fusion*, Elsevier B.V., v. 28, p. 45–59, 2016. ISSN 15662535. Disponível em: <http://dx.doi.org/10.1016/j.inffus.2015.08.005>. Citado 2 vezes nas páginas 21 e 39.

BERTHON, P. R. et al. Marketing meets web 2.0, social media, and creative consumers: Implications for international marketing strategy. *Business Horizons*, v. 55, n. 3, p. 261 – 271, 2012. ISSN 0007-6813. SPECIAL ISSUE: STRATEGIC MARKETING IN A CHANGING WORLD. Disponível em: <http://www.sciencedirect.com/science/article-/pii/S0007681312000080>. Citado na página 21.

BONSóN, E. et al. Local e-government 2.0: Social media and corporate transparency in municipalities. *Government Information Quarterly*, v. 29, n. 2, p. 123–132, 2012. ISSN 0740-624X. Disponível em: <https://www.sciencedirect.com/science/article/pii-/S0740624X1200010X>. Citado na página 78.

BUDIARDJO, E. K. et al. Social crm features identification for higher education. *Journal of Engineering and Applied Sciences*, Medwell Journals, v. 12, n. 9, p. 2327–2333, 2017. Citado na página 81.

CARR, C. T.; HAYES, R. A. Social Media: Defining, Developing, and Divining. *Atlantic Journal of Communication*, v. 23, n. 1, p. 46–65, 2015. ISSN 15456889. Citado 2 vezes nas páginas 21 e 51.

CARRASCAL, A. I. O. et al. Descubrimiento de conocimiento en historias clínicas mediante minería de texto. *RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação*, Associação Ibérica de Sistemas e Tecnologias de Informação (AISTI), n. 34, p. 29–43, 2019. Citado na página 53.

CHAKRABORTY, A. et al. Who Makes Trends? Understanding Demographic Biases in Crowdsourced Recommendations. n. Icwsm, p. 22–31, 2017. Citado na página 51.

CHANEY, A. J.; BLEI, D. M. Visualizing topic models. *ICWSM 2012 - Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, p. 419–422, 2012. Citado 2 vezes nas páginas 28 e 29.

CHARLES-SMITH, L. E. et al. Using social media for actionable disease surveillance and outbreak management: A systematic literature review. *PLoS ONE*, v. 10, n. 10, p. 1–20, 2015. ISSN 19326203. Citado na página 71.

CHEN, H.; H.L.CHIANG, R.; C. Storey, V. Business Intelligence and Analytics: From Big Data To Big Impact. *MIS Quarterly*, v. 36, n. 4, p. 1165–1188, 2018. ISSN 01406736. Disponível em: <http://www.jstor.org/stable/41703503>. Citado na página 26.

CHEN, Y. et al. Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowledge-Based Systems*, Elsevier B.V., v. 163, p. 1–13, 2019. ISSN 09507051. Citado 4 vezes nas páginas 28, 29, 56 e 84.

CHINCHILLA, L. D. C. C.; FERREIRA, K. A. R. Analysis of the behavior of customers in the social networks using data mining techniques. *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, IEEE, p. 623–625, 2016. Citado 4 vezes nas páginas 27, 28, 53 e 54.

CHIRUMALLA, K.; OGHAZI, P.; PARIDA, V. Social media engagement strategy: Investigation of marketing and R&D interfaces in manufacturing industry. *Industrial Marketing Management*, v. 74, n. February 2017, p. 138–149, 2018. ISSN 00198501. Citado na página 53.

CIRQUEIRA, D. et al. A Literature Review in Preprocessing for Sentiment Analysis for Brazilian Portuguese Social Media. In: *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. [S.l.]: IEEE, 2018. p. 746–749. ISBN 978-1-5386-7325-6. Citado 4 vezes nas páginas 26, 53, 56 e 69.

CIRQUEIRA, D. et al. Performance evaluation of sentiment analysis methods for Brazilian Portuguese. *Lecture Notes in Business Information Processing*, v. 263, p. 245–251, 2017. ISSN 18651348. Citado 2 vezes nas páginas 43 e 53.

CIRQUEIRA, D. et al. Opinion Label : A Gamified Crowdsourcing System for Sentiment Analysis Annotation. *XVI Workshop de Ferramentas e Aplicações*, p. 209–213, 2017. Citado na página 43.

COLOMO-PALACIOS, R. et al. Towards a social and context-aware mobile recommendation system for tourism. *Pervasive and Mobile Computing*, Elsevier B.V., v. 38, p. 505–515, 2017. ISSN 15741192. Citado na página 39.

CONSTANTINIDES, E.; HOLLESCHOVSKY, N. I. Impact of Online Product Reviews on Purchasing Decisions. *Proceedings of the 12th International Conference on Web Information Systems and Technologies*, p. 271–278, 2016. Disponível em: <http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0005861002710278>. Citado 7 vezes nas páginas 22, 25, 32, 34, 51, 67 e 83.

CONSUMIDOR.GOV. *Brazilian Consumer complaints*. 2021. Disponível em: <https://www.consumidor.gov.br>. Citado 2 vezes nas páginas 82 e 84.

COOPER, T.; STAVROS, C.; DOBELE, A. R. Domains of influence: exploring negative sentiment in social media. *Journal of Product Brand Management*, 2019. ISSN 1061-0421. Disponível em: <https://doi.org/10.1108/JPBM-03-2018-1820>. Citado na página 24.

DAMANI, O. P. Improving Pointwise Mutual Information (PMI) by incorporating significant co-occurrence. In: *CoNLL 2013 - 17th Conference on Computational Natural Language Learning, Proceedings*. [S.l.: s.n.], 2013. p. 20–28. ISBN 9781937284701. Citado na página 70.

DOMO. *Data Never Sleeps 7.0*. 2019. 1 p. Disponível em: <https://www.domo.com/learn/data-never-sleeps-7>. Citado na página 26.

DUNN, K.; HARNESS, D. Whose voice is heard? The influence of user-generated versus company-generated content on consumer scepticism towards CSR. *Journal of Marketing Management*, Routledge, v. 35, n. 9-10, p. 886–915, 2019. ISSN 14721376. Disponível em: <https://doi.org/10.1080/0267257X.2019.1605401>. Citado na página 22.

DYSON, B. et al. Evaluating the use of Facebook to increase student engagement and understanding in lecture-based classes. *Higher Education*, Kluwer Academic Publishers, v. 69, n. 2, p. 303–313, feb 2015. ISSN 00181560. Disponível em: <https://link.springer.com/article/10.1007/s10734-014-9776-3>. Citado na página 78.

EASLEY, D.; KLEINBERG, J. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. [s.n.], 2010. 23–46 p. Disponível em: <http://www.cs.cornell.edu/home/kleinber/networks-book/>. Citado na página 83.

EBIT. 43ª Webshoppers. p. 1–40, 2021. Citado na página 25.

EINWILLER, S. A.; STEILEN, S. Handling complaints on social network sites - An analysis of complaints and complaint responses on Facebook and Twitter pages of large US companies. *Public Relations Review*, Elsevier Inc., v. 41, n. 2, p. 195–204, 2015. ISSN 03638111. Citado 2 vezes nas páginas 51 e 58.

ERNALA, S. K. et al. Characterizing audience engagement and assessing its impact on social media disclosures of mental illnesses. In: *Twelfth international AAAI conference on web and social media*. [S.l.: s.n.], 2018. Citado na página 53.

FAASE, R.; HELMS, R.; SPRUIT, M. Web 2.0 in the CRM domain: defining social CRM. *International Journal of Electronic Customer Relationship Management*, v. 5, n. 1, p. 1, 2011. ISSN 1750-0664. Disponível em: <http://www.inderscience.com/link.php?id=39797>. Citado 2 vezes nas páginas 22 e 23.

FANG, A. et al. Examining the coherence of the top ranked tweet topics. *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 825–828, 2016. Citado 3 vezes nas páginas 68, 69 e 71.

FANG, A. et al. Examining the coherence of the top ranked tweet topics. *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 825–828, 2016. Citado na página 70.

FANG, B. et al. Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. *Tourism Management*, Elsevier Ltd, v. 52, p. 498–506, 2016. ISSN 02615177. Disponível em: <http://dx.doi.org/10.1016/j.tourman-.2015.07.018>. Citado na página 68.

FARSEEV, A.; CHUA, T. S. TweetFit: Fusing multiple social media and sensor data for wellness profile learning. *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, p. 95–101, 2017. Citado na página 25.

FERNANDES, L. C. et al. An extensive analysis of online restaurant reviews: A case study of the Amazonian Culinary Tourism. *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020*, v. 21, p. 81–84, 2020. Citado 2 vezes nas páginas 83 e 86.

FLESCH, R. F. A new readability yardstick. *The Journal of applied psychology*, v. 32 3, p. 221–33, 1948. Citado 2 vezes nas páginas 69 e 83.

GARCÍA, S. et al. *Data Preprocessing in Data Mining*. [S.l.: s.n.], 2015. 195–243 p. ISSN 18684408. ISBN 9783319102467. Citado na página 83.

GAVILANES, J. M.; FLATTEN, T. C.; BRETTEL, M. Content Strategies for Digital Consumer Engagement in Social Networks: Why Advertising Is an Antecedent of Engagement. *Journal of Advertising*, v. 47, n. 1, p. 4–23, 2018. ISSN 00913367. Citado 12 vezes nas páginas 13, 22, 26, 31, 32, 40, 41, 42, 46, 47, 51 e 58.

GENC-NAYEBI, N.; ABRAN, A. A systematic literature review: Opinion mining studies from mobile app store user reviews. *Journal of Systems and Software*, Elsevier Inc., v. 125, p. 207–219, 2017. ISSN 01641212. Citado na página 71.

GLOBALWEBINDEX. *Brazilian Consumers: Social Media Habits and Coronavirus Views - GWI*. [S.l.], 2020. Disponível em: <https://www.globalwebindex.com/reports-/coronavirus-brazil-consumers>. Citado na página 25.

GÓMEZ, M. et al. A Recommender System of Buggy App Checkers for App Store Moderators. *Proceedings - 2nd ACM International Conference on Mobile Software Engineering and Systems, MOBILESoft 2015*, p. 1–11, 2015. Citado na página 71.

GREENBERG, P. Social crm comes of age. *Sponsored by Oracle*, 2009. Citado 2 vezes nas páginas 78 e 80.

GREENBERG, P. *CRM at the Speed of Light: Social CRM Strategies, Tools, and Techniques*. [S.l.]: McGraw-Hill New York, 2010. Citado na página 78.

GRETZEL, U. et al. Conceptual foundations for understanding smart tourism ecosystems. *Computers in Human Behavior*, Elsevier Ltd, v. 50, p. 558–563, 2015. ISSN 07475632. Citado na página 39.

HAN, H. J. S. et al. What guests really think of your hotel: Text analytics of online customer reviews. *Cornell Hospitality Report*, v. 16, n. 2, p. 3–17, 2016. Disponível em: <http://scholarship.sha.cornell.edu/chrreports>. Citado na página 83.

HARRIGAN, P. et al. Customer engagement with tourism social media brands. *Tourism Management*, v. 59, p. 597–609, 2017. ISSN 02615177. Citado na página 39.

HE, L. et al. The voice of drug consumers: Online textual review analysis using structural topic model. *International Journal of Environmental Research and Public Health*, v. 17, n. 10, 2020. ISSN 16604601. Citado na página 83.

HIRSCH, M. et al. Googling endometriosis: a systematic review of information available on the Internet. *American Journal of Obstetrics and Gynecology*, Elsevier, v. 216, n. 5, p. 451–458.e1, 2017. ISSN 10976868. Citado 3 vezes nas páginas 69, 70 e 83.

HRASTINSKI, S.; AGHAEE, N. M. How are campus students using social media to support their studies? an explorative interview study. *Education and Information Technologies*, Springer, v. 17, n. 4, p. 451–464, 2012. Citado na página 78.

HRNJIC, A. The transformation of higher education: evaluation of crm concept application and its impact on student satisfaction. *Eurasian Business Review*, Springer, v. 6, n. 1, p. 53–77, 2016. Citado na página 79.

HUSSAIN, S. et al. Consumers' online information adoption behavior: Motives and antecedents of electronic word of mouth communications. *Computers in Human Behavior*, Elsevier Ltd, v. 80, p. 22–32, 2018. ISSN 07475632. Citado na página 39.

IBGE. *Estimativas da população com referência a 1° de julho de 2019 (xls)*. 2019. Disponível em: <https://agenciadenoticias.ibge.gov.br/agencia-detalhe-de-midia-.html?view=mediaibgecatid=2103id=3098>. Citado 4 vezes nas páginas 13, 24, 57 e 62.

IBGE. *PNAD contínua - Pesquisa Nacional por Amostra de Domicílios Contínua*. [S.l.], 2019. 8 p. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros-/liv101794_informativo.pdf>. Citado na página 24.

IBGE - Instituto Brasileiro de Geografia e Estatistica. *Produto Interno Bruto - PIB | IBGE*. 2020. Disponível em: <https://www.ibge.gov.br/explica/pib.php>. Citado 3 vezes nas páginas 13, 58 e 62.

INEP. *Brazilian Higher Education Census*. 2021. Disponível em: <https://www.gov.br-/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior/resultados>. Citado 2 vezes nas páginas 82 e 84.

JARVIS, J. My Dell hell. *The Guardian*, 2005. Disponível em: <https://www.theguardian-.com/technology/2005/aug/29/mondaymediasection.blogging>. Citado na página 78.

JO, J. M.; FERREIRA, M. An evaluation of sentiment analysis for mobile devices. n. October, 2017. Citado na página 83.

KAPLAN, A. M.; HAENLEIN, M. Users of the world , unite ! The challenges and opportunities of Social Media. 2010. Citado na página 21.

KARNA, N.; SUPRIANA, I.; MAULIDEVI, N. Social crm using web mining for indonesian academic institution. In: IEEE. *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*. [S.l.], 2015. p. 1–6. Citado 2 vezes nas páginas 78 e 81.

KEMP, S. *Digital 2021 - We Are Social*. 2021. 8 p. Disponível em: <https://wearesocial-.com/digital-2021>. Citado na página 25.

KIM, A. J.; JOHNSON, K. K. Power of consumers using social media: Examining the influences of brand-related user-generated content on Facebook. *Computers in Human Behavior*, Elsevier Ltd, v. 58, p. 98–108, 2016. ISSN 07475632. Disponível em: <http://dx.doi.org/10.1016/j.chb.2015.12.047>. Citado 4 vezes nas páginas 22, 32, 51 e 67.

KUBINA, M.; LENDEL, V. Successful Application of Social CRM in The Company. *Procedia Economics and Finance*, Elsevier, v. 23, p. 1190–1194, jan 2015. ISSN 2212-5671. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2212567115004876>. Citado na página 22.

KUMAR, A.; DABAS, V.; HOODA, P. Text classification algorithms for mining unstructured data: a SWOT analysis. *International Journal of Information Technology*, Springer Singapore, 2018. ISSN 2511-2104. Disponível em: <https://doi.org/10.1007-/s41870-017-0072-1>. Citado na página 26.

LI, D. et al. Integration of Knowledge Graph Embedding Into Topic Modeling with Hierarchical. n. Ccl, p. 940–950, 2019. Citado 4 vezes nas páginas 28, 29, 53 e 71.

LI, R. Traditional to hybrid: Social media's role in reshaping instruction in higher education. In: *Digital Arts and Entertainment: Concepts, Methodologies, Tools, and Applications*. [S.l.]: IGI Global, 2014. p. 387–411. Citado na página 78.

LI, Y. et al. The concept of smart tourism in the context of tourism information services. *Tourism Management*, Elsevier Ltd, v. 58, p. 293–300, 2017. ISSN 02615177. Citado na página 39.

LIU, B. et al. Growing story forest online from massive breaking news. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. [S.l.: s.n.], 2017. p. 777–785. Citado na página 40.

LIU, J. et al. Like it or not: The Fortune 500's Facebook strategies to generate users' electronic word-of-mouth. *Computers in Human Behavior*, Elsevier Ltd, v. 73, p. 605–613, 2017. ISSN 07475632. Citado na página 51.

LOBATO, F. et al. Social CRM: Biggest challenges to make it work in the real world. *Lecture Notes in Business Information Processing*, v. 263, p. 221–232, 2017. ISSN 18651348. Citado 7 vezes nas páginas 22, 23, 39, 51, 67, 82 e 83.

MAIZ, A.; ARRANZ, N.; JUAN, J. C. Factors affecting social interaction on social network sites: the Facebook case. *Journal of Enterprise Information Management*, v. 29, n. 5, p. 630–649, 2016. ISSN 17410398. Citado 2 vezes nas páginas 39 e 41.

MAROLT, M.; ZIMMERMANN, H. D.; PUCIHAR, A. Exploratory study of social CRM use in SMEs. *Engineering Economics*, v. 29, n. 4, p. 468–477, 2018. ISSN 13922785. Citado na página 23.

MARSHALL, A.; MUECK, S.; SHOCKLEY, R. How leading organizations use big data and analytics to innovate. *Strategy and Leadership*, v. 43, n. 5, p. 32–39, 2015. ISSN 10878572. Citado na página 23.

MCILROY, S. et al. Analyzing and automatically labelling the types of user issues that are raised in mobile app reviews. *Empirical Software Engineering*, Empirical Software Engineering, v. 21, n. 3, p. 1067–1106, 2016. ISSN 15737616. Citado 2 vezes nas páginas 32 e 51.

MEYLIANA, P.; HIDAYANTO, A. N.; BUDIARDJO, E. K. Social media adoption for social crm in higher education: An insight from indonesian universities. *International Journal of Synergy and Research*, Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, v. 4, n. 2, 2015. Citado na página 80.

MJSP, M. da justiça e segurança publica. *Data from Consumidor.gov.br - Datasets - Brazilian Open Data Portal*. 2021. Disponível em: <https://dados.gov.br/dataset-/reclamacoes-do-consumidor-gov-br1>. Citado na página 82.

MUJAHID, S. et al. An empirical study of Android Wear user complaints. *Empirical Software Engineering*, Empirical Software Engineering, v. 23, n. 6, p. 3476–3502, 2018. ISSN 15737616. Citado na página 52.

NAIR, C.; CHAN, S.; FANG, X. A case study of crm adoption in higher education. In: CITESEER. *Proceedings of the 2007 Information Resources Management Association International Conference*. [S.l.], 2007. Citado na página 79.

NETRICA. TOP ECOMMERCE RANKING REPORTS. Disponível em: https://ecommerce-brasil.rankings.netquest.digital. Acesso em: 27/01/2020 . 2020. Disponível em: <https://ecommerce-brasil.rankings.netquest.digital/>. Citado na página 69.

NETTO, J. S. D. et al. Melhorando sistemas de social crm por meio de eletronic word-of-mouth. *Revista Eletrônica de Iniciação Científica em Computação*, v. 17, n. 4, 2019. Citado na página 53.

NGUYEN, D. T. et al. Rapid Classification of Crisis-Related Data on Social Networks using Convolutional Neural Networks. n. Icwsm, p. 632–635, 2016. Disponível em: <http://arxiv.org/abs/1608.03902>. Citado 2 vezes nas páginas 28 e 29.

NOGUEIRA DE SOUSA, G. et al. Adoption of Social Crm in Micro and Small Enterprises: an Analysis of Santarém's Market. *Proceedings of the 15th CONTECSI International Conference on Information Systems and Technology Management*, v. 15, p. 0–2, 2018. Citado 2 vezes nas páginas 39 e 41.

NUSAIR, K. et al. A theoretical framework of electronic word-of-mouth against the backdrop of social networking websites. *Journal of Travel and Tourism Marketing*, Routledge, v. 34, n. 5, p. 653–665, 2017. ISSN 10548408. Citado na página 51.

OLIVEIRA, B.; CASAIS, B. The importance of user-generated photos in restaurant selection. *Journal of Hospitality and Tourism Technology*, 2018. ISSN 17579899. Citado 2 vezes nas páginas 24 e 39.

OLIVEIRA, L. Social student relationship management in higher education: extending educational and organisational communication into social media. In: *9th Annual International Technology, Education and Development Conference, IATED*. [S.l.: s.n.], 2015. Citado na página 80.

OLMEDILLA, M.; MARTÍNEZ-TORRES, M. R.; TORAL, S. Harvesting big data in social science: A methodological approach for collecting online user-generated content. *Computer Standards & Interfaces*, Elsevier, v. 46, p. 79–87, 2016. Citado na página 52.

ORENGA-ROGLÁ, S.; CHALMETA, R. Social customer relationship management: taking advantage of Web 2.0 and Big Data technologies. *SpringerPlus*, Springer International Publishing, v. 5, n. 1, 2016. ISSN 21931801. Citado 3 vezes nas páginas 23, 24 e 67.

OTHMAN, I. W. et al. Text readability and fraud detection. *ISBEIA 2012 - IEEE Symposium on Business, Engineering and Industrial Applications*, IEEE, n. 99, p. 296–301, 2012. Citado 2 vezes nas páginas 70 e 83.

PARK, E. O.; CHAE, B. K.; KWON, J. The structural topic model for online review analysis: Comparison between green and non-green restaurants. *Journal of Hospitality and Tourism Technology*, v. 11, n. 1, p. 1–17, 2018. ISSN 17579899. Citado na página 83.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, JMLR. org, v. 12, p. 2825–2830, 2011. Citado na página 43.

PORTO, J. S. et al. Legibilidade de artigos de um periódico nacional na área de melhoramento vegetal. *Cultivando o Saber*, v. 7, n. 2, p. 205–211, 2014. Citado na página 70.

PRADIPTARINI, C. Social Media Marketing: Measuring Its Effectiveness and Identifying the Target Market. *Journal of Undergraduate Research*, v. 14, p. 1–11, 2011. Citado na página 39.

PRESS, G. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. *Forbes Tech*, p. 4–5, 2016. Disponível em: <https://www.forbes.com-/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=661ce5206f63 https://www.forbes.com/sites/gilpress/2016-/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/>. Citado na página 26.

QI, J. et al. An overview of data fusion techniques for Internet of Things enabled physical activity recognition and measure. *Information Fusion*, Elsevier B.V., v. 55, n. September 2019, p. 269–280, 2020. ISSN 15662535. Citado na página 83.

QUATTRONE, G. et al. Is the sharing economy about sharing at all? A linguistic analysis of Airbnb reviews. In: *12th International AAAI Conference on Web and Social Media, ICWSM 2018*. [S.l.: s.n.], 2018. p. 668–671. ISBN 9781577357988. Citado na página 68.

RAMAN, R.; MENON, P. Using social media for innovation – market segmentation of family firms. *International Journal of Innovation Science*, 2018. ISSN 17572231. Citado na página 24.

RAVI, K.; RAVI, V. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, Elsevier, v. 89, p. 14–46, nov 2015. ISSN 0950-7051. Disponível em: <https://www.sciencedirect.com/science/article-/pii/S0950705115002336?via%3Dihub>. Citado na página 83.

REINHOLD, O.; ALT, R. Analytical social crm: Concept and tool support. In: *Bled eConference*. [S.l.: s.n.], 2011. p. 50. Citado 2 vezes nas páginas 81 e 86.

REINHOLD, O.; ALT, R. How companies are implementing social customer relationship management: Insights from two case studies. *26th Bled eConference - eInnovations: Challenges and Impacts for Individuals, Organizations and Society, Proceedings*, p. 206–221, 2013. Citado 4 vezes nas páginas 12, 13, 23 e 24.

RHEE, H. T.; YANG, S.-B. How does hotel attribute importance vary among different travelers? an exploratory case study based on a conjoint analysis. *Electronic markets*, Springer, v. 25, n. 3, p. 211–226, 2015. Citado na página 53.

RIGO, G.-E. et al. Crm adoption in a higher education institution. *JISTEM-Journal of Information Systems and Technology Management*, SciELO Brasil, v. 13, n. 1, p. 45–60, 2016. Citado na página 79.

ROBERTS, M. E. et al. Structural topic models for open-ended survey responses. *American Journal of Political Science*, v. 58, n. 4, p. 1064–1082, 2014. ISSN 15405907. Citado na página 83.

RODRIGUES, L.; JR, A. B.; LOBATO, F. Disability-related news: An analysis of user-generated content on social media posts. In: *Proceedings of the 16th National Meeting on Artificial and Computational Intelligence*. [S.l.: s.n.], 2019. Citado na página 53.

ROLLINS, J. B. Foundational Methodology for Data Science A 10-stage data science methodology that spans technologies and approaches. *IBM Analytics*, 2015. Citado 4 vezes nas páginas 27, 29, 53 e 54.

ROSENBERGER, M. Social customer relationship management: An architectural exploration of the components. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 9373, p. 372–385, 2015. ISSN 16113349. Citado na página 23.

ROSSOW, A. B. O uso do big data no estudo do mercado consumidor, potencialidades para aplicação no brasil. *Revista Sodebras [on line]*, v. 14, n. 159, p. 198–202, 2019. Citado na página 39.

ROY, G.; DATTA, B.; MUKHERJEE, S. Role of electronic word-of-mouth content and valence in influencing online purchase behavior. *Journal of Marketing Communications*, Routledge, v. 25, n. 6, p. 661–684, 2019. ISSN 14664445. Disponível em: <https://doi.org/10.1080/13527266.2018.1497681>. Citado 2 vezes nas páginas 34 e 67.

SABLAN, B. et al. The critical success factors (csfs) of social crm implementation in higher education. In: IEEE. *2017 International Conference on Research and Innovation in Information Systems (ICRIIS)*. [S.l.], 2017. p. 1–6. Citado na página 80.

SALMINEN, J. et al. Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, n. Icwsm, p. 330–339, 2018. Citado na página 56.

SASHI, C.; BRYNILDSEN, G.; BILGIHAN, A. Social media, customer engagement and advocacy. *International Journal of Contemporary Hospitality Management*, p. IJCHM–02–2018–0108, 2019. ISSN 0959-6119. Disponível em: <https://www-.emeraldinsight.com/doi/10.1108/IJCHM-02-2018-0108>. Citado na página 24.

SCHAFER, F. et al. Synthesizing CRISP-DM and Quality Management: A Data Mining Approach for Production Processes. *2018 IEEE International Conference on Technology Management, Operations and Decisions, ICTMOD 2018*, p. 190–195, 2019. Citado 4 vezes nas páginas 27, 28, 53 e 54.

SCHMÄH, M.; WILKE, T.; ROSSMANN, A. Electronic Word-of-Mouth: A Systematic Literature Analysis. *Lecture Notes in Informatics (LNI)*, p. 147, 2017. Citado 4 vezes nas páginas 22, 39, 51 e 67.

SEEMAN, E. D.; O'HARA, M. Customer relationship management in higher education: Using information systems to improve the student-school relationship. *Campus-wide information systems*, Emerald Group Publishing Limited, 2006. Citado na página 80.

SHARMA, R. et al. Digital literacy and knowledge societies: A grounded theory investigation of sustainable development. *Telecommunications Policy*, v. 40, n. 7, p. 628–643, 2014. ISSN 03085961. Citado na página 52.

SHIAU, W. L.; DWIVEDI, Y. K.; LAI, H. H. Examining the core knowledge on facebook. *International Journal of Information Management*, Elsevier, v. 43, n. December 2017, p. 52–63, 2018. ISSN 02684012. Disponível em: <https://doi.org/10.1016/j.ijinfomgt.2018.06-.006>. Citado 2 vezes nas páginas 21 e 39.

SILVA, J. R. d. et al. Redes sociais e promoção da saúde: utilização do facebook no contexto da doação de sangue. *RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação*, Associação Ibérica de Sistemas e Tecnologias de Informação (AISTI), n. 30, p. 107–122, 2018. Citado na página 51.

SILVA, W. et al. A methodology for community detection in Twitter. *Proceedings of the International Conference on Web Intelligence - WI '17*, p. 1006–1009, 2017. Citado 2 vezes nas páginas 39 e 67.

Simon Kemp. *Digital 2021: Brazil*. 2021. Disponível em: <https://datareportal.com-/reports/digital-2021-brazil>. Citado na página 25.

SOUSA, G. N. de; JUNIOR, A. F. L. J.; LOBATO, F. M. F. *Facebook posts for analyzing Content Strategies for Digital Consumer Engagement: a curated dataset*. Zenodo, 2021. Disponível em: <https://doi.org/10.5281/zenodo.5113266>. Citado 2 vezes nas páginas 32 e 88.

STATISTA. • *Global social media ranking 2019 | Statistic*. 2019. Citado na página 39.

STONE, M. The evolution of the telecommunications industry-What can we learn from it? *Journal of Direct, Data and Digital Marketing Practice*, v. 16, n. 3, p. 157–165, 2015. ISSN 17460174. Citado na página 57.

TANG, C.; GUO, L. Digging for gold with a simple tool: Validating text mining in studying electronic word-of-mouth (ewom) communication. *Marketing Letters*, Springer, v. 26, n. 1, p. 67–80, 2015. Citado na página 53.

TIRUNILLAI, S.; TELLIS, G. J. Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science*, v. 31, n. 2, p. 198–215, 2012. ISSN 07322399. Citado na página 25.

TRSTENJAK, B.; MIKAC, S.; DONKO, D. KNN with TF-IDF based framework for text categorization. *Procedia Engineering*, Elsevier B.V., v. 69, p. 1356–1364, 2014. ISSN 18777058. Citado na página 56.

VECCHIO, P. D. et al. Creating value from Social Big Data: Implications for Smart Tourism Destinations. *Information Processing and Management*, Elsevier, v. 54, n. 5, p. 847–860, 2018. ISSN 03064573. Citado na página 39.

VERMEER, S. A. et al. Seeing the wood for the trees: How machine learning can help firms in identifying relevant electronic word-of-mouth in social media. *International Journal of Research in Marketing*, Elsevier B.V., n. xxxx, p. 1–17, 2019. ISSN 01678116. Citado 4 vezes nas páginas 32, 51, 67 e 82.

VU, P. M. et al. Mining user opinions in mobile app reviews: A keyword-based approach. *Proceedings - 2015 30th IEEE/ACM International Conference on Automated Software Engineering, ASE 2015*, IEEE, p. 749–759, 2016. Citado 2 vezes nas páginas 32 e 51.

VYDISWARAN, V. G. et al. "Bacon bacon bacon": Food-related tweets and sentiment in metro detroit. In: *12th International AAAI Conference on Web and Social Media, ICWSM 2018*. [S.l.: s.n.], 2018. p. 692–695. ISBN 9781577357988. Citado 2 vezes nas páginas 68 e 69.

WALI, A. F.; WRIGHT, L. T. Customer relationship management and service quality: Influences in higher education. *Journal of Customer Behaviour*, Westburn Publishers Ltd, v. 15, n. 1, p. 67–79, 2016. Citado na página 80.

WANG, R. et al. Review on mining data from multiple data sources. *Pattern Recognition Letters*, v. 109, p. 120–128, 2018. ISSN 01678655. Citado na página 25.

WANG, Y.; YU, C. Social interaction-based consumer decision-making model in social commerce: The role of word of mouth and observational learning. *International Journal of Information Management*, v. 37, p. 179–189, 2015. ISSN 02684012. Citado na página 22.

WHITING, A.; WILLIAMS, D. Why people use social media: a uses and gratifications approach. *Qualitative Market Research: An International Journal*, v. 16, n. 4, p. 362–369, 2013. ISSN 13522752. Citado na página 21.

WIRTH, R. CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, n. 24959, p. 29–39, 2000. Citado 8 vezes nas páginas 12, 27, 28, 29, 53, 54, 82 e 83.

WITTWER, M.; REINHOLD, O.; ALT, R. Customer Context and Social CRM : A Literature Review and Research Agenda. *30th Bled eConference Digital Transformation – From Connecting Things to Transforming Our Lives*, p. 1–14, 2017. Citado na página 23.

XIANG, Z. et al. A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, Elsevier Ltd, v. 58, p. 51–65, 2017. ISSN 02615177. Citado 4 vezes nas páginas 67, 68, 69 e 71.

ZHANG, J. Multi-source remote sensing data fusion: Status and trends. *International Journal of Image and Data Fusion*, v. 1, n. 1, p. 5–24, 2010. ISSN 19479824. Citado na página 83.

ZHANG, L.; YANG, W. Consumers' responses to invitations to write online reviews. *International Journal of Contemporary Hospitality Management*, p. IJCHM–01–2018–0022, 2019. ISSN 0959-6119. Disponível em: <https://www.emeraldinsight.com/doi/10.1108-/IJCHM-01-2018-0022>. Citado na página 24.

APÊNDICES

APÊNDICE A – ARTIGO PUBLICADO NO *COMPUTER ON THE BEACH* (2020) - *FERRAMENTAS PARA ANÁLISE DE MÍDIAS SOCIAIS: UM LEVANTAMENTO SISTEMÁTICO*

Ferramentas para Análise de Mídias Sociais: Um levantamento sistemático

Emanuel Gilvan Souza Lima Júnior
juniorlima.e@gmail.com
Universidade Estadual do Maranhão

Gustavo Nogueira de Sousa
sougusta@gmail.com
Universidade Estadual do Maranhão

Antonio Fernando Lavareda Jacob Junior
antonio.jacob@gmail.com
Universidade Estadual do Maranhão

Fábio Manoel França Lobato
lobato.fabiof@gmail.com
Universidade Federal do Oeste Pará

RESUMO

Social media are increasingly present in the daily life of human beings. As a consequence, the volume of content generated by users grows considerably. These contents are published on digital platforms, such as blogs, communities and online social networks. The analysis of these data requires different approaches and methods to obtain a satisfactory result. Seeking to know the current scenario of social media analysis, this work performs a systematic mapping to identify the most used databases, algorithms, and tools in research in this area. The results presented provides the identification of the main research topics and how they are related to each other and can serve as a basis to guide new researchers, both in the choice of data sources and in the definition of tools and algorithms in the solution of the identified problems.

KEYWORDS

Mídias sociais, Análise de Dados, Mineração de Texto

1 INTRODUÇÃO

O advento da Web 2.0 possibilitou o surgimento de plataformas interativas como as mídias sociais [1, 2]. Com isso, percebe-se que os usuários vêm expressando opiniões sobre produtos, serviços ou até mesmo sobre acontecimentos em geral neste tipo de mídia [3–6]. Neste sentido, as mídias sociais estão cada vez mais presentes no cotidiano do ser humano [7]. De acordo com um levantamento realizado por [8], mais de 4 bilhões de pessoas acessam a rede mundial de computadores, e mais de 3 bilhões utilizam mensalmente redes sociais. Consequentemente, o volume de dados gerados pelos usuários cresce continuamente [9, 10].

Estes dados estão presentes em mídias sociais como blogs, redes sociais *online*, comunidades de jogos virtuais, ou mundos sociais virtuais [2, 6] e trazem consigo informações relevantes [11]. Por esta razão, tais plataformas têm sido usadas como fonte de dados para estudos de interesse dos mais variados campos de atuação científica no mercado, governo, academia, e até movimentos sociais, como os protestos de junho de 2013 no Brasil [12–15]. Comumente, os dados gerados pelos usuários encontram-se na forma não-estruturada, consistindo sobretudo de *corpus* textuais [16].

Em meio aos novos desafios analíticos impostos pela migração digital, a ciência de dados tem apoiado a descoberta de conhecimento por meio do desenvolvimento de recursos de aquisição, armazenamento, análise e visualização de dados, a fim de trazer vantagens competitivas às corporações, bem como possibilitar uma gama de estudos de fenômenos sociais [17]. Em vista disso, há na literatura uma ampla variedade de bases de dados, ferramentas e algoritmos utilizados para a implementação e desenvolvimento de análises em mídias sociais [16]. No entanto, para que o processo de descoberta de conhecimentos tenha êxito, é necessário que a seleção das bases de dados, técnicas e algoritmos seja feita cuidadosamente, considerando-se que problemas diferentes exigem diferentes soluções [18, 19].

Nesse sentido, visando mapear o estado da arte e o estado da prática no que tange a análise de mídias sociais, o presente trabalho descreve um mapeamento sistemático da literatura relacionada a análises de dados advindos de mídias sociais, com o objetivo de identificar as bases de dados mais prevalentes, bem como algoritmos, métodos e ferramentas mais utilizadas. Ademais, é conduzida ainda uma análise de tópicos baseada nas palavras-chave dos artigos, com o intuito de identificar as correlações de temas entre as pesquisas.

Para tanto, mapeou-se sistematicamente trabalhos publicados nas conferências de maior impacto na área, a saber: *International AAAI Conference on Web and Social Media* (ICWSM), *Workshop on Computational Approaches to Subjectivity* (WASSA), *ACM Conference on Hypertext and Hypermedia* (ACM HT) e *International Conference on Social Media & Society* (ICSMS), identificando trabalhos destinados a análise de mídias sociais e, neles, quais bases de dados foram utilizadas, quais ferramentas foram usadas, e quais algoritmos foram aplicados. Além disso, realizou-se também uma análise de redes de palavras-chave. A partir deste presente estudo, futuros pesquisadores podem ter uma base para orientar os seus estudos na área, pois neste trabalho é apresentado a relação entre os principais termos, além da listagem das principais bases de dados, ferramentas e algoritmos mais utilizados por estudos que são referência em análises de mídias sociais.

O restante do artigo encontra-se organizado como segue. Na Seção 2 são apresentados conceitos de redes sociais e mídias sociais. Além disso, alguns trabalhos relacionados são discutidos. Na Seção 3 a metodologia utilizada neste trabalho é descrita. Os resultados são discutidos na Seção 4. Por fim, as conclusões do estudo e sugestões de trabalhos futuros são apresentadas na Seção 5.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção, são apresentados alguns conceitos pertinentes ao foco do presente estudo, bem como alguns trabalhos relacionados que serviram de base para a construção do mesmo.

Formato de Referência: JACOB JUNIOR, Antonio F L. et al. Ferramentas para Análise de Mídias Sociais: Um levantamento sistemático. In COMPUTER ON THE BEACH (COTB'20), 11., 2020, Balneário Camboriú. **Anais...** Balneário Camboriú: Universidade do Vale do Itajaí, 2020, p. 1-8.

2.1 Redes Sociais e Mídias Sociais

Redes Socais *online* podem ser definidas como serviços baseados na *Web* que permitem a indivíduos: construir um perfil (público ou semipúblico) dentro de um sistema limitado; articular uma lista de outros usuários com os quais compartilham uma conexão; e visualizar e percorrer sua própria lista de conexões e de outras pessoas no sistema. A peculiaridade das redes sociais não está em permitir aos usuários novas conexões, mas em permiti-los articular e tornar visível sua rede (relacionamentos já existentes *offline*) [20].

Mídias sociais são canais baseados na Internet para interação entre usuários em tempo real e/ou de forma assíncrona, produzindo valor da interação e do conteúdo gerado pelos usuários [2, 21]. As mídias sociais podem ser definidas ainda como atividades, práticas e comportamentos entre comunidades de pessoas que se reúnem *online*, via aplicativos da *Web*, para compartilhar informações, conhecimentos e opiniões usando meios de conversação e empregando palavras, imagens, vídeos ou áudios [22]. As definições de sites de redes socais por vezes são aplicadas de forma inconsistente a mídias sociais, prejudicando o avanço de estudos na área. Frente ao exposto, faz-se necessário explicitar a diferença entre redes sociais *online* e mídias sociais.

Grosso modo, redes sociais *online* têm como característica precípua a conexão de nós (usuários) e incluem em si aspectos de mídias sociais, uma vez que permitem a geração de conteúdo por parte dos usuários. Neste sentido, as mídias sociais podem ser consideradas mais amplas, uma vez que tais mídias estão presentes em todas as esferas da sociedade, não restringindo-se às redes sociais *online*, e se expandiram a partir dos avanços da tecnologia da informação e da possibilidade de acesso por diferentes tipos de aparelhos eletrônicos, como computadores pessoais, *smartphones*, *tablets* e TVs [23].

2.2 Análise de Redes e Mídias Sociais

Considerando as diferenças entre redes sociais e mídias sociais apresentadas anteriormente, o entendimento das análises possíveis fica facilitado. No que tange ao estudo de redes, as análises destinam-se a entender as dependências (conexões) entre as entidades sociais (nós) dispostos nos dados, caracterizando seu comportamento e seus efeitos na rede como um todo e também no aspecto temporal, quando se considera redes dinâmicas [24].

Uma lista não exaustiva de aplicações de análise de redes sociais comuns inclui: i) detecção de comunidades [19]; ii) identificação de *Hubs* ou autoridades/influenciadores; iii) análise da evolução temporal de atores; iv) estudo de fluxo de informação [19]. É importante reforçar que estes aspectos são analisados com base nos nós e suas conexões.

Já os estudos de mídias sociais baseiam-se, sobretudo, na análise do conteúdo gerado pelos usuários (do inglês UGC – *User Generated Content*) [9, 10]. O UGC assume diferentes formas, como *tweets*, atualizações de *status* do *Facebook* e vídeos no *Youtube*, bem como comentários em sites de notícias ou *reviews* de produtos [25, 26]. A análise de mídias sociais pode ser realizada de diversas formas, como análise de sentimentos [12, 27], modelagem de tópicos [12], classificação/categorização de postagens [28], identificação de tendências [29], dentre outros.

Em resumo, a análise de mídias sociais foca no conteúdo gerado pelos usuários enquanto a análise de redes sociais tem como objeto o estudo das conexões entre os usuários.

2.3 Trabalhos Relacionados

Nos últimos anos, as mídias sociais têm sido estudadas em diversos contextos de aplicação. A partir disto, [11] avaliaram artigos relevantes na área com o intuito de encontrar os principais temas relacionados a mídias sociais, incluindo seus benefícios e efeitos colaterais. Para isso, foi realizada uma revisão na literatura partindo de palavras-chave e de uma pesquisa manual em periódicos, tendo em vista a identificação da evolução geral e os direcionamentos dos trabalhos analisados. Foram consideradas as publicações mais relevantes nos principais periódicos no período de 1997 a 2016. Assim, foi possível verificar que as mídias sociais estão sendo conhecidas por sua capacidade de agregação, e não somente pela socialização e agrupamento. Verificou-se também que o *Facebook*, *Twitter* e comunidades *online* são fontes de dados presentes na maioria dos trabalhos, e que ao longo dos anos houve mudanças no objeto de análise dos trabalhos publicados. Por exemplo, em 2011 as publicações relatavam o conteúdo gerado pelo usuário como um novo tipo de conteúdo *online*, já em 2013 as publicações avaliavam os aspectos comuns em avaliações e recomendações populares, e por fim, nos anos de 2015 e 2016 os estudos focaram no comércio em mídias sociais.

Com o enfoque na aplicação prática, mas ainda no contexto da evolução e aprimoramento do uso das mídias sociais, [30] realizaram uma pesquisa com o objetivo de determinar as implicações de estudos sobre mídias sociais no processo de tomada de decisão nos negócios ou em *marketing*. Para isso, o estudo utiliza mapeamento sistemático da literatura para criar classificações e conduzir análises. Foram considerados apenas estudos que utilizaram dados advindos do *Twitter*, pois esta plataforma é uma poderosa ferramenta para negócios aumentarem sua performance e ganhos. Com isso, obteve-se da base de dados *Web of Science* um total de 41 artigos do período de 2014 e 2015. Desta forma, concluiu-se que a análise descritiva é a principal ferramenta utilizada para o gerenciamento de *marketing* em mídias sociais. Por outro lado, foi verificado que além da análise descritiva, alguns autores propõem a mídia social como ferramenta para solucionar dilemas de *marketing* como a criação de segmentos e a percepção da marca.

Em [31] há um estudo das implicações diretas das mídias sociais em negócios. Foi realizado um mapeamento sistemático de artigos relacionados a métricas e análises de mídia social em *marketing*. Foram coletados 60 artigos de cinco grandes periódicos da área, do período de 2011 a 2016. Os artigos foram analisados de acordo como a sua metodologia de pesquisa, tipos de análises, campo de estudo, objetivos de *marketing* e de acordo com os tipos de plataformas de mídias utilizadas. As pesquisas analisadas mostram que em estratégias de *marketing* em mídia social as plataformas dominantes são o *Facebook* e o *Twitter*, e conceitos dominantes e recorrentes estão relacionados à indústria do turismo e ao *marketing* centrado no consumidor. Ademais, as pesquisas mostram que análise de conteúdo e atividade de mídia social, processamento de linguagem natural, análise de texto e análise de sentimentos estão presentes na maioria dos trabalhos.

As aplicações práticas e estudos de mídias sociais necessitam de tecnologias, algoritmos e bases de dados específicas para atingir os seus objetivos. Isto é observado ao comparar os trabalhos de [19, 32–34], nos quais o *Twitter* foi utilizado como base de dados principal, no entanto com algoritmos e propósitos diferentes. Além destes, outros estudos utilizam os dados extraídos do *Facebook*, tal como [35], que compara diversos algoritmos de extração de sentimento em conteúdo. Ainda no contexto de análise de sentimentos, em [36] é apresentada uma plataforma de *crowdsourcing* para a anotação de sentimentos. Por fim, uma modelagem de tópicos para avaliar reclamações presentes em uma plataforma *online* destinadas ao recebimento de reclamações de empresa é utilizada em [37].

3 METODOLOGIA

Este trabalho faz um mapeamento sistemático da literatura relacionada a mídias sociais com o objetivo de identificar o estado da arte e da prática sobre análise de mídias sociais, a partir da identificação de ferramentas, bases de dados, algoritmos e da construção de uma rede de palavras-chave [38, 39].

Para a condução deste mapeamento sistemático foram seguidas as recomendações providas por [38] e [40]. Com base nisto, este estudo foi conduzido em três etapas: Planejamento, Condução, e Relatório. As definições e atividades referentes ao planejamento e condução podem ser observadas nas subseções abaixo. O Relatório é descrito na Seção 4.

3.1 Planejamento

Na etapa de planejamento ocorre a delimitação do escopo do trabalho por meio da definição das perguntas de pesquisa. Definem-se também as bases de dados a serem consideradas e critérios de seleção dos estudos [38]. A instanciação destes elementos para o presente trabalho encontra-se descrita a seguir.

3.1.1 Perguntas de pesquisa. As perguntas de pesquisa definidas têm o objetivo de esclarecer questões relevantes sobre a área de análise de mídias sociais. Isso inclui métodos quantitativos acerca dos recursos mais frequentes e métodos analíticos acerca dos tópicos mais relevantes e como estes se relacionam em publicações de análises de mídias sociais. As perguntas de pesquisa deste mapeamento sistemático são:

- PP1: Quais são as principais bases de dados para análise de mídias sociais?
- PP2: Quais são as principais técnicas e algoritmos nas análises de mídias sociais?
- PP3: Como os tópicos abordados pela literatura sobre análise de mídias sociais estão relacionados?

Tais perguntas visam contemplar as dimensões de interesse, a saber: Bases de dados mais utilizadas (PP1), Técnicas e Algoritmos mais prevalentes (PP2) e quais as aplicações e temas relevantes (PP3).

3.1.2 Processo de Pesquisa. Devido à dificuldade de mapear os termos-chave utilizados e considerando a vasta aplicabilidade de análise de mídias sociais e da necessidade de se considerar apenas trabalhos atuais, optou-se por considerar apenas trabalhos das conferências mais relevantes para a área de análise de mídias sociais. As conferências foram selecionadas observando a relevância medida através da métrica *Índice h5* da ferramenta *Google Scholar Metrics*[1]. Assim, foram escolhidas as seguintes conferências, com maior relevância em análise de dados em mídias sociais: ICWSM, WASSA, ICSMS e ACM HT.

Tabela 1: Critérios de inclusão e exclusão de trabalhos.

Critérios de inclusão	Trabalhos completos
	Trabalhos publicados nos últimos quatro anos
	Trabalhos sobre análise de mídias sociais
Critérios de exclusão	Trabalhos não escritos em língua inglesa
	Trabalhos sem base de dados coletada em mídias sociais
	Revisões Sistemáticas
	Trabalhos não classificados como *Full Paper* e *Short Paper*

3.1.3 Seleção dos estudos. A fim de selecionar somente trabalhos relevantes na área de análise de mídias sociais, a seleção dos trabalhos obedeceu aos critérios de inclusão e exclusão apresentados na Tabela 1. Como critérios de inclusão foram considerados todos os trabalhos completos, publicados nos últimos 4 anos e que fizessem análise de mídias sociais. Por outro lado, resumos simples e expandidos, revisões sistemáticas, trabalhos que não fizessem análise de mídias sociais e trabalhos que não utilizavam nenhuma base de dados foram excluídos do presente estudo. Faz-se necessário ressaltar que os anais da edição de 2019 da conferência *ACM HT* foram publicados durante a realização deste trabalho, portanto ficaram de fora da análise aqui apresentada.

3.2 Condução

A etapa de condução consiste na execução dos passos definidos na fase de planejamento. A seguir, alguns aspectos técnicos são descritos.

3.2.1 Coleta dos trabalhos. As conferências consultadas na presente análise publicam seus trabalhos em diferentes repositórios. Assim, fez-se necessário o desenvolvimento de diferentes ferramentas de *web scraping* para a coleta de alguns dados disponibilizados em cada plataforma, como título, link, resumo e, em alguns casos, palavras-chave. Estes dados coletados foram então armazenados em um *dataset*, onde foram filtrados aplicando-se os critérios de inclusão e exclusão e, por fim, somente as publicações que atendem aos critérios selecionados foram utilizadas neste mapeamento sistemático.

3.2.2 Análises dos trabalhos. Os trabalhos filtrados foram então escrutinados através de uma leitura direcionada, a fim de identificar ferramentas, bases de dados e algoritmos utilizados. Além disto, os trabalhos foram avaliados considerando suas palavras-chave. Para isso, foi construída uma rede de palavras-chave com o objetivo de avaliar a correlação das mesmas, e consecutivamente avaliar a relação dos trabalhos avaliados neste estudo.

4 RESULTADOS & DISCUSSÕES

O processo de coleta dos trabalhos resultou em 964 trabalhos publicados nos mais variados tipos aceitos pelas conferências. Destes, os critérios de inclusão e exclusão foram responsáveis pela eliminação

[1] https://scholar.google.com/citations?view_op=top_venues

Tabela 2: Trabalhos resultantes.

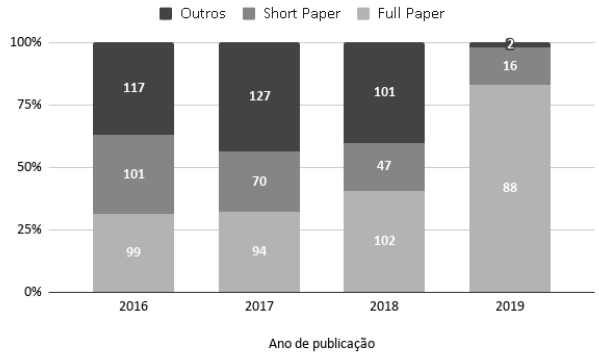
Conferência	Pesquisa inicial	Trabalhos excluídos	Trabalhos analisados
ICSMS	320	151	169
ICWSM	403	311	92
ACM HT	125	48	77
WASSA	116	14	102
Total	964	524	440

de 524 trabalhos da análise. O restante, um total de 440 trabalhos, foi analisado e usado na construção do mapeamento sistemático. Na Tabela 2 são apresentadas as quantidades de trabalhos em cada conferência.

A conferência ICWSM apresentou uma grande quantidade de trabalhos na pesquisa inicial. No entanto, aproximadamente 77% destes trabalhos foram excluídos do presente estudo, sendo a maior taxa de trabalhos excluídos dentre as conferências analisadas. Isto se deve à presença de trabalhos publicados como *Tutorials* e *Datasets*, além da grande aceitação de trabalhos do tipo *Poster*, que, por sua natureza resumida, não são objetos de análise neste trabalho.

Da conferência WASSA foram coletados um total de 116 trabalhos na pesquisa inicial. Uma das justificativas dos organizadores da conferência para a pequena quantidade de trabalhos submetidos (e, por consequência, de trabalhos aceitos) é o curto intervalo de tempo da edição de 2018 para a edição de 2019 [41]. No entanto, a taxa de exclusão destes trabalhos é de apenas 12% sendo, portanto, a menor taxa de exclusão entre as conferências analisadas. Novamente, isto deve-se à natureza dos trabalhos aceitos: em sua maioria, *Full Papers*.

Figura 1: Distribuição dos trabalhos ao longo dos anos.



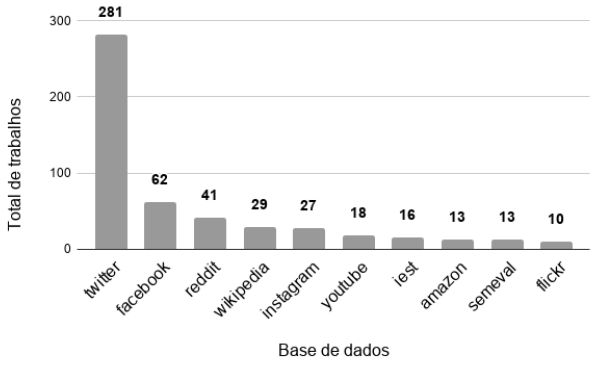
Na Figura 1 é possível observar a distribuição dos trabalhos ao longo dos anos analisados. Nos anos de 2016 a 2018 os trabalhos apresentam uma distribuição razoavelmente homogênea, mas no ano de 2019 essa distribuição foi totalmente alterada. Essa discrepância pode ser explicada devido à ausência dos trabalhos da edição de 2019 da *ACM HT*, que possui um grande percentual de *Full* e *Short Papers*.

A seguir, encontram-se os resultados de cada dimensão de pesquisa deste mapeamento sistemático.

4.1 Bases de dados

Para identificar as bases de dados mais frequentes em análises de mídias sociais, foram contabilizadas as fontes de dados utilizadas nas análises dos trabalhos selecionados. O gráfico da Figura 2 ilustra as 10 bases de dados mais frequentes.

Figura 2: As dez fontes de dados mais utilizadas nos trabalhos analisados.



As fontes de dados identificadas como *Iest* e *SemEval* são bases de dados relacionadas a *shared tasks*.

Por meio da análise da Figura 2 é possível responder à PP1: "*Quais são as principais bases de dados para análise de mídias sociais?*". Percebe-se que o *Twitter* é, de longe, a base de dados mais utilizada, seguido de *Facebook*, *Reddit* e *Wikipedia*. O número de trabalhos que realizaram análises em bases de dados extraídas do *Twitter* é aproximadamente quatro vezes maior que o número de trabalhos que utilizam o *Facebook*, e aproximadamente 6 vezes maior que o número de trabalhos que utilizam o *Reddit*. A prevalência do *Twitter* pode ser explicada tanto pela quantidade de usuários ativos (330 milhões mensalmente [42]) e de publicações (511.200 *tweets* por minuto [43]), quanto pela disponibilização de sua API com acesso completo aos dados, enquanto outras plataformas de redes sociais *online* têm políticas restritiva de extração de dados, e impõem limites às requisições que tornam algumas pesquisas inviáveis. A segunda fonte de dados mais utilizada é o *Facebook*, plataforma de rede social *online* com maior quantidade de usuários no mundo [44]. Em terceiro lugar está o *Reddit*, que já ultrapassou o *Twitter* em quantidade de usuários ativos mensalmente [44, 45]. A diversidade de tópicos encontrados na rede social permite afirmar que existem mais de 130 mil comunidades ativas na plataforma [46].

4.2 Ferramentas e algoritmos

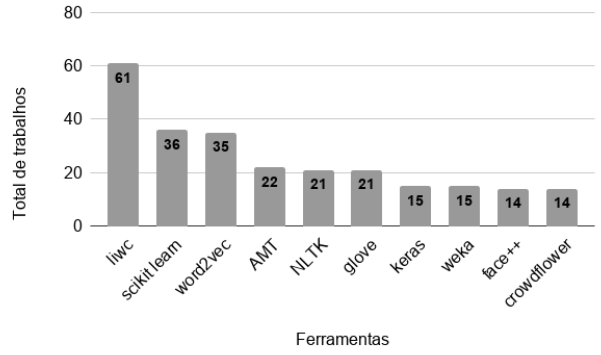
A análise quantitativa das ferramentas e algoritmos presentes nas publicações selecionadas ocorreu em dois momentos: primeiro, foi realizado um levantamento das principais ferramentas utilizadas, conforme ilustrado na Figura 3. Em seguida, os algoritmos mais utilizados foram contabilizados, o resultado pode ser visualizado na Figura 4. Para estes resultados, foi considerada a quantidade de trabalhos que utilizam cada um destes recursos.

A Tabela 3 apresenta a descrição de cada uma das ferramentas mais frequentes.

Tabela 3: Ferramentas mais relevantes e suas descrições.

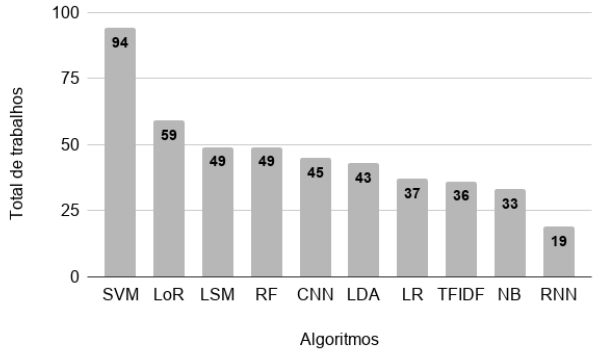
Ferramenta	Descrição
LIWC	Em constante desenvolvimento desde a década de 1990, *Linguistic Inquiry and Word Count* (LIWC) é uma ferramenta de análise de texto com mais de 6.400 palavras da língua inglesa cadastradas em seus dicionários. Desenvolvida com o objetivo de mapear estados sociais e psicológicos do autor do texto, a ferramenta é muito utilizada para tarefas de processamento de linguagem natural [47].
Scikit-learn	Ferramenta de aprendizagem de máquina desenvolvida em Python para mineração de dados e análise de texto. Seu uso está associado a tarefas como pré-processamento, redução de dimensionalidade, classificação, regressão e *clustering*, ou mesmo modelos computacionais, como redes neurais [48].
Word2vec	Grupo de modelos associados à tarefa de *word embedding*, responsáveis por associar cada palavra presente em um *corpus* a uma representação vetorial n-dimensional para aplicações de processamento de linguagem natural [49].
MTurk	*Amazon Mechanical Turk* (*MTurk*) é uma plataforma de *crowdsourcing* que permite o uso de uma força de trabalho distribuída para a execução de tarefas de processos, como validação de dados e respostas a uma pesquisa. A cooperação massiva dos indivíduos envolvidos no processo diminui o tempo de trabalhos de coleção e análise de dados [50].
NLTK	*Natural Language Toolkit* (NLTK) é uma plataforma para processamento e análise de linguagem natural, oferecendo suporte através de uma vasta biblioteca com *tokenizers*, *stemmers*, *POS-taggers*, *corpora*, *lexicons*, entre outros [51].
GloVe	Criado pelo Grupo de Processamento de Linguagem Natural da Universidade de Stanford, *Global Vectors for Word Representation* (GloVe) é um modelo de regressão log-bilinear não supervisionado para aprendizagem de representação vetorial de palavras [52, 53].
Keras	Projeto de código aberto disponibilizado no *GitHub* [54], *Keras* é uma plataforma para *Deep Learning* escrita em Python [55]. Com uma API de fácil compreensão utilizada por mais de 250 mil usuários e empresas como *Netflix* e *Uber*, *Keras* é a segunda ferramenta mais utilizada para *Deep learning* no mundo [56].
Weka	*Waikato Environment for Knowledge Analysis* (Weka) é uma ferramenta computacional para preparação, classificação, regressão, *clustering*, associação e visualização de dados [57].
Face++	Ferramenta que permite a busca, comparação e detecção de faces e partes do corpo humano [58].
CrowdFlower	Plataforma de *crowdsourcing* com mais de 5 milhões de trabalhadores cadastrados [59]. Nela, os trabalhadores podem ser recrutados por nível de confiança/acurácia, país e/ou idioma principal [60], já os trabalhadores podem escolher as tarefas com base na recompensa, número de tarefas restantes, nível de satisfação dos trabalhadores etc [61].

Figura 3: As dez ferramentas mais utilizadas nos trabalhos analisados.



AMT = *"Amazon Mechanical Turk"*, NLTK = *"Natural Language Toolkit"*

Figura 4: Os dez algoritmos mais utilizadas nos trabalhos analisados.



SVM = *"Support Vector Machine"*, LoR = *"Logistic Regression"*, LSM = *"Long Short-term Memory"*, RF = *"Random Forest"*, CNN = *"Convolutional Neural Network"*, LDA = *"Latent Dirichlet Allocation"*, LR = *"Linear Regression"*, TFIDF = *"Frequency–Inverse Document Frequency"*, NB = *"Naive Bayes"*, RNN = *"Recurrent Neural Network"*

Reunindo-se estes resultados, pode-se responder à PP2: "*Quais são as principais técnicas e algoritmos nas análises de mídias sociais?*". Os resultados obtidos permitem verificar o predomínio de ferramentas que auxiliam na tarefa de processamento de texto, como observado nas três primeiras colocações do gráfico ilustrado na Figura 3. Sobre estas ferramentas, algumas conclusões acerca do seu uso podem ser obtidas. A ferramenta mais utilizada é a *Linguistic Inquiry and Word Count* (LIWC), seu uso pode ser justificado pelo fato dela possuir um grande número de palavras em língua

inglesa cadastradas e categorizadas para identificação de padrões emocionais, sociais, cognitivos e estruturais de textos [62]. A segunda ferramenta mais utilizada, *Scikit-learn*, disponibiliza uma vasta gama de algoritmos para auxiliar em tarefas de *Machine Learning*, popularizando o seu uso na área. Já a terceira ferramenta, *Word2Vec*, fornece um espaço vetorial para trabalhos que realizam análise de textos. Sua principal colaboração é, portanto, a transformação de dados não-estruturado (ou semi-estruturados) em dados estruturados, mais adequados a tarefas computacionais. *Softwares* que agrupam e oferecem diversas soluções em análise de dados continuam presentes entre as 10 ferramentas mais frequentes, como o *Weka*, *Keras* e *GloVe*. *MTurk* e *CrowdFlower* auxiliam pesquisadores em tarefas com necessidade de emprego massivo de material humano, poupando tempo e recursos.

4.3 Palavras-chave

Para compreender como se relacionam os tópicos de pesquisa dos trabalhos analisados, foi realizado um estudo baseado nas palavras-chave presentes nos trabalhos analisados. Inicialmente, cada trabalho com palavras-chave foi transformado em registros contendo pares destas, por meio de uma combinação simples denotada por

$$r = \frac{n!}{p!(n-p)!}$$

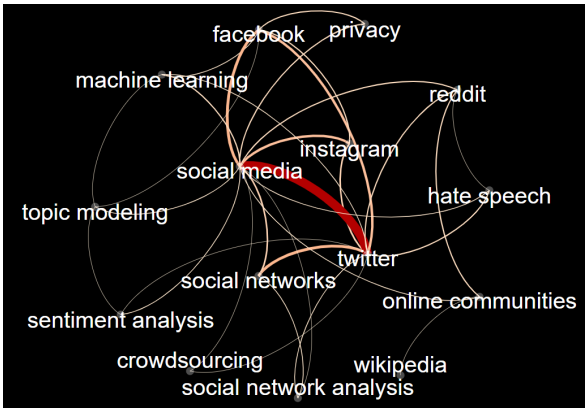
em que r é a quantidade de registros resultantes, n é o número de palavras-chave do trabalho e $p = 2$, pois estas são tomadas aos pares. Todos os registros são inseridos em um arquivo CSV. O resultado é um documento similar a uma lista de adjacência de uma rede de palavras-chave, que pode ser visualizada com o suporte de uma ferramenta de análise de redes. Neste trabalho, foi feita a escolha pelo *software Gephi*[2] para este fim.

Foram identificados 827 nós e 2560 arestas pertencentes ao grafo gerado. Para uma melhor visualização, a Figura 5 ilustra um subgrafo da rede contendo apenas nós com grau igual ou superior a 20 (em um grafo não-dirigido, o grau de um nó indica o número de arestas incidentes a ele). Esta visualização ajuda a responder à PP3: "*Como os tópicos abordados pela literatura sobre análise de mídias sociais estão relacionados?*". Nela, é possível ter uma noção de como as palavras-chave mais frequentes em análise de mídias sociais estão relacionadas, através das arestas adjacentes aos nós que representam as palavras-chave. É notório, por exemplo, que os termos "*social media*" e "*Twitter*" têm uma relação forte, pois a coloração e espessura da aresta que conecta os dois termos indicam seu elevado peso na visualização da rede ilustrada.

A aparição do termo inglês para privacidade entre os termos mais utilizados denota a preocupação com o acesso e a exposição dos dados publicados nas redes sociais como o *Facebook*, que é um nó vizinho na rede. O termo "*social media*" naturalmente encontra-se no centro do subgrafo. Já o nó com o termo "*Social network analysis*" no subgrafo indica a presença significante de trabalhos relacionados a esse termo, fato que pode ser comprovado não somente na prevalência dos termos *Facebook*, *Instagram*, *Reddit* e *Twitter*, mas também na visualização de arestas com peso alto incidentes a ambas as redes sociais, apontando uma grande coocorrência das fontes de dados em trabalhos de análise de redes sociais. A palavra-chave "*Crowdsourcing*" corrobora com os resultados encontrados na resposta à PP2, onde foi identificada a popularidade das ferramentas "*MTurk*" e "*CrowdFlower*". Por fim, "*topic modeling*", "*sentiment analysis*" e "*hate speech*" são evidências fortes do grande interesse das pesquisas na identificação dos comportamentos e dos padrões emocionais e cognitivos do conteúdo gerado por usuários.

[2] https://gephi.org/

Figura 5: Grafo que apresenta a relação entre as palavras-chave.



Na Tabela 4 são descritas algumas estatísticas básicas extraídas no *Gephi*. Elas são interpretadas de acordo com [63]. A rede de palavras-chave completa (com todos os nós) apresenta um grau médio de 6,191, o que indica que as palavras-chave se relacionam em média com seis outras palavras. Outro ponto que chama a atenção na Tabela 4 é o número de componentes conectados, que são 46, e o coeficiente *clustering* médio que tem valor de 0,933, estas duas métricas indicam a formação de *clusters* em torno de 46 palavras-chave.

Tabela 4: Estatísticas da rede de palavras-chave.

Métrica	Valor
Número de nós	827
Número de arestas	2560
Grau médio	6,191
Grau médio ponderado	6,568
Diâmetro da rede	7
Densidade do grafo	0,007
Componentes conectados	46
Coeficiente *clustering* médio	0,933
Comprimento médio do caminho	2,914

5 CONSIDERAÇÕES FINAIS

Neste artigo, foi realizado um mapeamento sistemático da literatura relacionada a análise de mídias sociais. O foco desse mapeamento sistemático é a identificação de bases de dados, ferramentas e algoritmos utilizados. Além disto, foi identificado como os principais tópicos de pesquisa dos trabalhos estavam relacionados entre si.

Os resultados obtidos durante a análise dos estudos mostram que as pesquisas existentes na área de análise de mídias sociais usam majoritariamente dados do *Twitter* em suas análises, sendo esta fonte de dados aproximadamente quatro vezes mais utilizada que a segunda, *Facebook*. Outro ponto que chama atenção é a forma na qual os estudos estão relacionados. Sendo que há 46 termos que agrupam outros termos à sua volta, os quais podem ser identificados como temas centrais para estes trabalhos. Ademais, os principais termos foram apresentados na forma de um grafo, no qual as arestas indicaram as relações entre os termos.

No entanto, este estudo apresenta algumas limitações que precisam ser tratadas em trabalhos futuros. A primeira está relacionada à acurácia, pois o estudo não apresenta validação cruzada. A segunda, é relacionada às análises realizadas, uma vez que os resultados não incluem a finalidade na qual cada algoritmo foi utilizado nos estudos. Devido a isto, nos trabalhos futuros pretendemos incluir a validação cruzada em todas as etapas da metodologia, e também expandir as análises e incluir o uso prático de cada base de dados, ferramenta e algoritmo utilizados nos estudos sobre mídias sociais.

REFERÊNCIAS

[1] Pierre R. Berthon, Leyland F. Pitt, Kirk Plangger, and Daniel Shapiro. Marketing meets web 2.0, social media, and creative consumers: Implications for international marketing strategy. *Business Horizons*, 55(3):261 – 271, 2012. ISSN 0007-6813. doi: https://doi.org/10.1016/j.bushor.2012.01.007. URL http://www.sciencedirect.com/science/article/pii/S0007681312000080. SPECIAL ISSUE: STRATEGIC MARKETING IN A CHANGING WORLD.

[2] Andreas M. Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59 – 68, 2010. ISSN 0007-6813. doi: https://doi.org/10.1016/j.bushor.2009.09.003. URL http://www.sciencedirect.com/science/article/pii/S0007681309001232.

[3] Rui Fan, Jichang Zhao, Yan Chen, and Ke Xu. Anger is more influential than joy: Sentiment correlation in weibo. *CoRR*, abs/1309.2402, 2013. URL http://arxiv.org/abs/1309.2402.

[4] Yany Grégoire, Audrey Salle, and Thomas M Tripp. Managing social media crises with your customers: The good, the bad, and the ugly. *Business Horizons*, 58(2): 173 – 182, 2015. ISSN 0007-6813. doi: https://doi.org/10.1016/j.bushor.2014.11.001. URL http://www.sciencedirect.com/science/article/pii/S0007681314001566. EMERGING ISSUES IN CRISIS MANAGEMENT.

[5] Thomas M Tripp and Yany Grégoire. When unhappy customers strike back on the internet. *MIT Sloan Management Review*, 52(3):37–44, 2011.

[6] Helen Donelan, Karen Kear, and Magnus Ramage. *Online communication and collaboration: A reader*. Routledge, 2012.

[7] Nick Couldry. *Media, society, world: Social theory and digital media practice*. Polity, 2012.

[8] We Are Social. Global digital report 2018. *Erişim: https://wearesocial.com/blog/2018/01/global-digital-report-2018*, 2018.

[9] Andrew Mcafee and Erik Brynjolfsson. Spotlight on Big Data Big Data: The Management Revolution, 2012. Acedido em 15-03-2017. *Harvard Business Review*, (October):1–9, 2012. URL http://tarjomefa.com/wp-content/uploads/2017/04/6539-English-TarjomeFa-1.pdf.

[10] Theodore Lynn, Philip Healy, Steven Kilroy, Graham Hunt, Lisa Van Der Werff, Shankar Venkatagiri, and John Morrison. Towards a general research framework for social media research using big data. *IEEE International Professional Communication Conference*, 2015-September:1–8, 2015. ISSN 21581002. doi: 10.1109/IPCC.2015.7235843.

[11] Kawaljeet Kaur Kapoor, Kuttimani Tamilmani, Nripendra P. Rana, Pushp Patil, Yogesh K. Dwivedi, and Sridhar Nerur. Advances in Social Media Research: Past, Present and Future. *Information Systems Frontiers*, 20(3):531–558, jun 2018. ISSN 15729419. doi: 10.1007/s10796-017-9810-y.

[12] Weiguo Fan and Michael D Gordon. The Power of Social Media Analytics. *Commun. ACM*, 57(6):74–81, 2014. ISSN 0001-0782. doi: 10.1145/2602574.

[13] Raj Agnihotri, Rebecca Dingus, Michael Y. Hu, and Michael T. Krush. Social media: Influencing customer satisfaction in b2b sales. *Industrial Marketing Management*, 53:172 – 180, 2016. ISSN 0019-8501. doi: https://doi.org/10.1016/j.indmarman.2015.09.003. URL http://www.sciencedirect.com/science/article/pii/S0019850115002631.

[14] Michael J Magro. A review of social media use in e-government. *Administrative Sciences*, 2(2):148–161, 2012.

[15] Luiz Antonio Joia and Carla Danielle Soares. Social media and the trajectory of the "20cents movement" in brazil: An actor-network theory-based investigation. *Telematics and Informatics*, 35(8):2201 – 2218, 2018. ISSN 0736-5853. doi: https://doi.org/10.1016/j.tele.2018.08.007. URL http://www.sciencedirect.com/science/article/pii/S0736585318303277.

[16] Stefan Stieglitz, Milad Mirbabaie, Björn Ross, and Christoph Neuberger. Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, 39(October 2017): 156–168, 2018. ISSN 02684012. doi: 10.1016/j.ijinfomgt.2017.12.002. URL https://doi.org/10.1016/j.ijinfomgt.2017.12.002.

[17] Tobias Brandt, Johannes Bendler, and Dirk Neumann. Social media analytics and value creation in urban smart tourism ecosystems. *Information and Management*, 54(6):703–713, 2017. ISSN 03787206. doi: 10.1016/j.im.2017.01.004. URL http://dx.doi.org/10.1016/j.im.2017.01.004.

[18] Li Cai and Yangyong Zhu. The challenges of data quality and data quality assessment in the big data era. *Data science journal*, 14, 2015.

[19] Wendel Silva, Ádamo Santana, Fábio Lobato, and Márcia Pinheiro. A methodology for community detection in Twitter. *Proceedings of the International Conference on Web Intelligence - WI '17*, pages 1006–1009, 2017. doi: 10.1145/3106426.3117760. URL http://dl.acm.org/citation.cfm?doid=3106426.3117760.

[20] Danah M Boyd and Nicole B Ellison. Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication*, 13(1):210–230, 2007.

[21] Caleb T Carr and Rebecca A Hayes. Social media: Defining, developing, and divining. *Atlantic Journal of Communication*, 23(1):46–65, 2015.

[22] Lon Safko and David K Brake. A bíblia da mídia social: táticas, ferramentas e estratégias para construir e transformar negócios. *São Paulo: Blucher*, 2010.

[23] Per Andersen. *What is Web 2.0?: ideas, technologies and implications for education*, volume 1. JISC Bristol, 2007.

[24] Tabassum Shazia, Pereira Fabiola S F., Fernandes Sofia, Gama João, Shazia Tabassum, Fabiola S. F. Pereira, Sofia Fernandes, and João Gama. Social network analysis: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 0(0):e1256, 2018. ISSN 19424787. doi: 10.1002/widm.1256. URL http://doi.wiley.com/10.1002/widm.1256https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1256.

[25] Lucas Rodrigues, Jorge Junior, and Fábio Lobato. A culpa é dela! É isso o que dizem nos comentários das notıcias sobre a tentativa de feminicıdio de elaine caparroz. In *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining*, pages 47–58, Porto Alegre, RS, Brasil, 2019. SBC. URL https://sol.sbc.org.br/index.php/brasnam/article/view/6547.

[26] Fábio Lobato, Márcia Pinheiro, Antonio Jacob, Olaf Reinhold, and Ádamo Santana. Social CRM: Biggest Challenges to Make it Work in the Real World. In Witold Abramowicz, Rainer Alt, and Bogdan Franczyk, editors, *Business Information Systems Workshops: BIS 2016 International Workshops, Leipzig, Germany, July 6-8, 2016, Revised Papers*, volume 263, pages 221–232. Springer International Publishing, Cham, 2017. ISBN 978-3-319-52464-1. doi: 10.1007/978-3-319-52464-1_20. URL http://dx.doi.org/10.1007/978-3-319-52464-1{_}20http://link.springer.com/10.1007/978-3-319-52464-1{_}20.

[27] Jorge A. Balazs and Juan D. Velásquez. Opinion Mining and Information Fusion: A survey. *Information Fusion*, 27:95–110, 2016. ISSN 15662535. doi: 10.1016/j.inffus.2015.06.002.

[28] GUSTAVO NOGUEIRA DE SOUSA, ISABELLE DA SILVA GUIMARÃES, ANTONIO FERNANDO LAVAREDA JACOB JR, and FÁBIO MANOEL FRANÇA LOBATO. GERENCIAMENTO DE PUBLICIDADES NA PLATAFORMA DAS REDES SOCIAIS DE ACORDOCOMCATEGORIAS DE CONTEÚDO. *Revista SODEBRAS*, 14(166):18–23, oct 2019. ISSN 18093957. doi: 10.29367/issn.1809-3957.14.2019.166.18. URL http://sodebras.com.br/edicoes/N166.pdf.

[29] Diego Saez-Trumper, Giovanni Comarela, Virgílio Almeida, Ricardo Baeza-Yates, and Fabrício Benevenuto. Finding trendsetters in information networks. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, page 1014, 2012. doi: 10.1145/2339530.2339691. URL http://dl.acm.org/citation.cfm?doid=2339530.2339691.

[30] Carolina Nicolas Alarcón, Angélica Urrutia Sepúlveda, Leslier Valenzuela-Fernández, and Jaime Gil-Lafuente. Systematic mapping on social media and its relation to business. *European Research on Management and Business Economics*, 24(2):104–113, may 2018. ISSN 24448834. doi: 10.1016/j.iedeen.2018.01.002.

[31] Nikolaos Misirlis and Maro Vlachopoulou. Social media metrics and analytics in marketing – S3M: A mapping literature review, feb 2018. ISSN 02684012.

[32] Marjori N M Klinczak and Celso A Kaestner. Identificação de Temas em Redes Sociais por meio de técnicas de agrupamento. 2017.

[33] Augusto Zangrandi and Luis A Rivera. Sensores Sociais em Detecção de Eventos Sociais. 2019.

[34] Jader Fabiano, Batista Marques, Fernanda Dos, Santos Cunha, Anita Mª, and Rocha Fernandes. Sistema de Monitoração e Análise de Comentários nas Mídias Sociais. *Anais do Computer on the Beach*, 2017.

[35] Douglas Cirqueira, Márcia Pinheiro, Thaís Braga, Antonio Jacob, Olaf Reinhold, Rainer Alt, and Ádamo Santana. Improving relationship management in universities with sentiment analysis and topic modeling of social media channels. *Proceedings of the International Conference on Web Intelligence - WI '17*, pages 998–1005, 2017. doi: 10.1145/3106426.3117761. URL http://dl.acm.org/citation.cfm?doid=3106426.3117761.

[36] Douglas Cirqueira, Lucas Vinícius, Márcia Pinheiro, Jacob Jr. Antônio F. L., Fábio Lobato, and Ádamo Santana. Opinion Label : A Gamified Crowdsourcing System for Sentiment Analysis Annotation. *XVI Workshop de Ferramentas e Aplicações*, pages 209–213, 2017.

[37] Gustavo Rangel Torres de Almeida, Fabio Lobato, and Douglas Cirqueira. Improving Social CRM through eletronic word-of-mouth: a case study of ReclameAqui. *XIVWorkshop de Trabalhos de Iniciação Científic*, 2017.

[38] Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64:1–18, 2015. ISSN 09505849. doi: 10.1016/j.infsof.2015.03.007.

[39] Beatriz Vendrame, Emely Albernaz, Rodrigo Sacchi, and Valguima Odakura. Mapeamento sistemático sobre ferramentas digitais online para o ensino-aprendizagem de algoritmos e programação no ensino superior. *Anais do Computer on the Beach*, 2019.

[40] Barbara A. Kitchenham, David Budgen, and O. Pearl Brereton. Using mapping studies as the basis for further research - A participant-observer case study. *Information and Software Technology*, 53(6):638–651, 2011. ISSN 09505849. doi: 10.1016/j.infsof.2010.12.011. URL http://dx.doi.org/10.1016/j.infsof.2010.12.011.

[41] *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Minneapolis, USA, June 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W19-1300.

[42] J CLEMENT. Twitter: number of active users 2010-2019 | statista, 2019. URL https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.

[43] DOMO. Data never sleeps 7.0 inforgraphic | domo, 2019. URL https://www.domo.com/learn/data-never-sleeps-7.

[44] We Are Social. Global digital report 2019 - we are social, 2007. URL https://wearesocial.com/blog/2018/01/global-digital-report-2018.

[45] Andrew Hutchinson. Reddit now has as many users as twitter, and far higher engagement rates | social media today, 2019. URL https://www.socialmediatoday.com/news/reddit-now-has-as-many-users-as-twitter-and-far-higher-engagement-rates/521789/.

[46] Inc Reddit. Press - reddit, 2019. URL https://www.redditinc.com/press.

[47] LIWC. Liwc 2015: How it works | liwc, 2019. URL http://liwc.wpengine.com/how-it-works/.

[48] scikit-learn developers. Documentation scikit-learn: machine learning in python 8212; scikit-learn 0.21.3 documentation, 2019. URL https://scikit-learn.org/stable/documentation.html.

[49] Google Code Archive. Google code archive - long-term storage for google code project hosting., 2019. URL https://code.google.com/archive/p/word2vec/.

[50] Amazon Mechanical Turk. Amazon mechanical turk, 2019. URL https://www.mturk.com/.

[51] NLTK Project. Natural language toolkit 8212; nltk 3.4.5 documentation, 2019. URL https://www.nltk.org/.

[52] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://www.aclweb.org/anthology/D14-1162.

[53] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[54] Keras Team. Github - keras-team/keras: Deep learning for humans, 2019. URL https://github.com/keras-team/keras.

[55] Keras Team. Home - keras documentation, 2019. URL https://keras.io/.

[56] Keras Team. Why use keras - keras documentation, 2019. URL https://keras.io/why-use-keras/.

[57] LIWC. Weka 3 - data mining with open source machine learning software in java, 2019. URL https://www.cs.waikato.ac.nz/ml/weka/.

[58] Face++. Face++ - face++ cognitive services, 2019. URL https://www.faceplusplus.com/.

[59] J.C.F. de Winter, M. Kyriakidis, D. Dodou, and R. Happee. Using crowdflower to study the relationship between self-reported violations and traffic accidents. *Procedia Manufacturing*, 3:2518 – 2525, 2015. ISSN 2351-9789. doi: https://doi.org/10.1016/j.promfg.2015.07.514. URL http://www.sciencedirect.com/science/article/pii/S2351978915005156. 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015.

[60] Nicholas B King, Sam Harper, Meredith Young, Sarah C Berry, and Kristin Voigt. The impact of social and psychological consequences of disease on judgments of disease severity: An experimental study. *PloS one*, 13(4):e0195338, 2018.

[61] Guoliang Li, Jiannan Wang, Yudian Zheng, Ju Fan, and Michael J. Franklin. *Crowdsourcing Background*, pages 11–20. Springer Singapore, Singapore, 2018. ISBN 978-981-10-7847-7. doi: 10.1007/978-981-10-7847-7_2. URL https://doi.org/10.1007/978-981-10-7847-7_2.

[62] James W Pennebaker, Cindy K Chung, Molly Ireland, Amy Gonzales, and Roger J Booth. The development and psychometric properties of liwc2007, 2007. URL http://www.liwc.net/LIWC2007LanguageManual.pdf.

[63] Shazia Tabassum, Fabiola S.F. Pereira, Sofia Fernandes, and João Gama. Social network analysis: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(5):1–21, 2018. ISSN 19424795. doi: 10.1002/widm.1256.

APÊNDICE B – ARTIGO PUBLICADO NO *ICRM* (2021) - *SOCIAL CRM AS A BUSINESS STRATEGY: DEVELOPING DYNAMIC CAPABILITIES OF MICRO AND SMALL ENTERPRISES*

Social CRM as a business strategy: developing dynamic capabilities of Micro and Small Enterprises

Isabelle da Silva Guimarães[1], Gustavo Nogueira de Sousa[2], Antônio Jacob Junior[2] and Fábio Manoel França Lobato[1,2]

[1] Federal University of Western Pará
[2] State University of Maranhão
fabio.lobato@ufopa.edu.br

Abstract. The global pandemic, caused by the spread of COVID-19, has transformed the way people go through shopping. In this sense, Social Media is an important channel for sharing information about products, services, and brands. Thus, considering the market relevance of these channels in this new scenario, this paper describes the results of a survey on how to use Social Media to improve customer relationship management processes on 31 companies. The focus was on digital marketing for micro and small enterprises. Moreover, we examined four companies in-depth, identifying challenges and perspectives on Social CRM adoption by micro and small enterprises. The results show that this is still an unexplored environment for micro and small companies, but with great potential to boost sales, develop customer loyalty and increase brand visibility. The lessons learned have the potential of guiding policymakers to create more adequate measures for boosting this market sector.

Keywords: Digital Marketing, Micro and Small Enterprises, Social CRM.

1 Introduction

The global pandemic, caused by the spread of COVID-19, has transformed the way people shop [1]. Consumers migrated their purchases to e-commerce platforms, a fact that positions the modality as one of the main pillars on sustaining the global economy [2]. According to [3], the post-coronavirus trade will no longer be the same. Therefore, it is necessary that companies invest time and resources to understand their client in the best possible way [4]. In this scenario, Social Media (SM) platforms stand out. In January 2021, SM platforms reported over 4.2 billion active users [5]. It is noteworthy that this was the biggest growth in three years. This growing phenomenon of content generation may become one of the most fundamental ones when it comes to product development and brands by small, medium, and large companies [6]. From User-Generated Content on social media, companies are able to make predictions about new trends and remain competitive in their market niche [7]. However, it is not feasible to remain on traditional customer management strategies [8]. Hence, it was necessary to direct traditional Customer Relationship Management (CRM) processes that correspond in marketing, innovation, sales, user experience, and after-sales, to the SM environment [9].

Social CRM as the adaptation of interactions between companies and customers through social media. It can also be defined as a set of processes that make use of technologies to capture and analyze customer data in order to know their individual needs [10]. Moreover, with the knowledge generated from these data, companies are able to understand the behavior of their consumers and objectively identify their needs [11]. In addition, marketing campaigns can be more efficient, in order to directly reach their target audience and convert leads into buyers [12]. This scenario is valid for both large companies and Micro and Small Companies (MSCs) [13]. It is important to highlight the representativeness of MSCs in the Brazilian economy. Currently, small businesses account for 99% of the 6.4 million Brazilian companies and are responsible for about 52% of formal jobs in the private sector [14].

Despite this economic relevance, this sector reports a shortage of qualified professionals to fulfill their demands and their low expertise to implement Social CRM strategies in their businesses. It was identified in the existing literature that few studies are directed to study the use of Social CRM strategies by MSCs. Therefore, based on its economical relevance for the country, it is necessary to analyze the implementation of Social CRM strategies by Micro and Small Companies in order to identify opportunities for intervention. In these gaps, the following research questions defined: RQ1: Which aspects of Social CRM are important for MSCs? RQ2: How do MSCs adopt Social CRM strategies in their businesses?

To answer these questions we conducted a 2-step investigation on how to use Social Media to improve customer relationship management, identifying challenges and perspectives on Social CRM adoption by micro and small enterprises. The first step consisted of applying a survey on 31 companies, aiming to identify some quantitative aspects. In the second step, we collected and analyzed four case studies companies from different areas in the market. We also highlighted some lessons learned and perspectives that have the potential of guiding policymakers to create more adequate measures for boosting the MSCs sector. The remainder of this paper is structured as follows: Section 2 presents a brief survey of related works followed by Section 3, which deals with the description of the methodology. Right after, in Section 4, the results of the research are presented and, finally, the conclusions.

2 Related Works

[15] examine how firms can improve CRM capabilities such as marketing and sales through Social Media usage, and his results demonstrate that social media can amplify the positive results of customer engagement and consequently, firm performance. [16]l investigated the usage of Facebook Commerce (F-commerce) to verify how/how much it is influencing the organizational performance of micro and small companies.

Similarly, [17] sought to verify whether the adoption of Facebook has an influence on the company's performance. As a result, the authors agree that the use of Facebook platforms for business processes helps companies performance incomes and outcomes. In addition, external pressures such as competition and consumer influence are characterized as reasons to boost relationships with the customers. [18] manage the role of Social CRM in Social Information Systems (SIS), based on the results of four

case studies. The authors' conclusions show that the systems have applications that are not restricted to strengthening the company-customer relationship but can also be used to promote business in order to increase its expansion capabilities.

In the same niche, [19] surveyed the technologies in the marketing area that are adopted in small businesses. The results show that managers see them as an opportunity to develop their relationship with customers. The authors also present a model built from the research's insights that consists of the process of collecting information, building content, and measuring results.

Then [16] makes a study based on quantitative variables, obtained through an online survey, and qualitative, raised from an exploratory search in the state of the art. [18] points out results in relation to Social CRM based on case studies built from interviews, providing quantitative and qualitative data, however, the approach was made in Large Companies, fleeing the reality in which this work is inserted.

3 Methodology

This study was conducted in a quantitative way, through the application of a survey; And qualitative, through case studies. After recognizing the literature and defining the relevant concepts, the scope of the work and the construction of the research questions were defined. The scope was limited to MSCs and Individual Micro-Entrepreneur (IME) who use Social Media in their businesses. The research questions, presented in the Introduction, were defined from the gaps in the existing literature. The quantitative data was collected through a self-administered online survey divided into three subsections: i) general information about the company (e.g. size, the activity segment, social media management, *etc*.); ii) social media management (e.g. media used, published content, difficulties encountered, *etc*.), and, iii) customer relationship management (e.g. perception of the CRM concept, software used, *etc*.). The response method was divided between essays and Likert-scale (from 1 to 5). The response time could vary between 5 and 10 minutes.

The second phase occurred with the collection of qualitative data through the case study methodology adapted from [20]. Four MSCs were selected, who actively participated in the events promoted by our research group. To guarantee unbiased results, all four companies belong to different market niches, described in the subsection Case Studies. The data collection was conducted through semi-structured interviews lasting between 50 and 60 minutes. Table 1 presents the script defined in conjunction with practitioners.

Table 1. Script for conducting semi-structured interviews.

Section	Question
Manager data	General information of the interviewee
General information about social media	Who manages Social Media? Do you have external consultancy? If so, how does it happen? How did the evolution take place when using the networks?
Motivations for using social media in business	What are the perceived benefits of adopting social media in business? Were external factors perceived as influencing the motivation for using social media?

Investments and internal evaluations	How do you define the success of a publication? Do you invest money in Social Media? Does investing in Social Media / Networks make a difference to the business?
Use of Social CRM	Do you adopt CRM and/or Social CRM strategies in the company? How do you conceptualize Social CRM? What are the reasons for adoption? Who is responsible for the strategies? How is media management done? Which social networks and metrics are most used? What are the difficulties experienced?
Pre-in-pos COVID-19	What are the strategies/actions to face the crisis? What are the post-pandemic perspectives?

It is important to highlight that all interviews were conducted through telephone calls, with the purpose of avoiding social contact, since this phase of the research coincided with the period of dissemination of the new coronavirus. The results were first organized individually in a feedback format for the interviewees, highlighting the positive and negative aspects of the business in the areas of customer relations. After that, the information collected was cross-checked in order to assess the equality and differences between the companies' processes. The results were validated in partnership with specialists in Social CRM and Digital Marketing. The research results were organized and culminated in this work.

4 Results

In this section, we present the results from the survey and interviews, contextualizing and discussing some relevant findings.

4.1 Companies Characterization

This survey was a prerequisite to attend some training courses provided by our research group. We used this approach instead of broadcasting the survey because it was possible to co-validate the answers. Moreover, it provided a good engagement with the open questions. We obtained 31 responses from MSCs, IME, and informal businesses that actively use Social Media in their business. Table 2 shows the main information of the companies.

Table 2. General information of the companies.

Size	Total in numbers	Percentage
MSCs	5	16,1%
IME	14	45,2%
Informal	12	38,7%
Operating segment	**Total in numbers**	**Percentage**
Foodservice	8	25,8%
Advisory and communication	6	19,3%
Entertainment	5	16,1%
Engineers and Information Systems	5	16,1%

Clothing and handcrafts	4	12,9%
Beauty and well-being	2	6,4%
Veterinary service	1	3,3%

Table 3 shows that entrepreneurs consider the internet as a very important channel for the development of their businesses. For this question, the entrepreneurs gave their answers on a Likert scale ranging from 1 to 5. For a better understanding of the result, groupings were made between 1 and 2 were grouped in “low relevance”; 3 and 4 in “medium relevance”; and 5 was called “high relevance”.

Table 3. Internet and social media management.

Internet relevance for revenues	Total in numbers	Percentage
Low relevance	1	3,2%
Medium relevance	3	9,7%
High relevance	27	87,1%
Who manages social media	Total in numbers	Percentage
Owner/Manager	30	96,8%
Employee	1	3,2%
Outsourced	0	0%

Also in Table 3, it is possible to verify that, in almost all cases, the owner or manager is responsible for managing the company's social networks. Regarding the use of social media, it is clear that the most widespread are WhatsApp with 27 positive responses, followed by Instagram with 24 and Facebook with 22 of the 31 companies. This result agrees with [21], which says that these platforms are adopted because they are free of charge and easy to access. This distribution can be explained by the fact that these three social media are among the most used by Brazilians, only behind YouTube [22]. From this information, it can be inferred that the use of YouTube for business is still unexplored by companies, as in this research it represents only 6.5% of the sample, which, in turn, corresponds to only 2 positive responses, thus identifying gaps in the adoption of social media by small businesses.

Table 4. Types of content published on social media

Types of content	Total in numbers	Percentage
Sharing products	26	83,9%
Promotion	18	58,1%
Advertising	17	54,8%
Relevant information for costumers	15	48,4%
Feedback search	11	35,5%
Institutional material of the company	10	32,3%
Drawing and conquest	7	22,6%

In Table 4, it is possible to see that the main purpose of using social media is for the dissemination of products and services already consolidated in companies, followed

by the sharing promotions and construction of advertisements for new products. However, one of the least used topics is the search for customer feedback, which corresponds to less than 36% of the companies. It is known that one of the purposes of using social media for business is to create proximity between consumer and company, therefore, instigating sharing of shopping and relationship experiences with the company is of considerable importance for the strengthening of relations of customer loyalty.

Regarding the difficulties encountered by managers in the use of online social networks, Figure 1 shows that the biggest complication is related to the lack of knowledge of tools for the management of these platforms. In addition, a recurring topic among entrepreneurs is also the lack of time devoted to creating content and evaluating results. This can be explained by the information present in Table 3, wherein 96.8% of the cases it is the owner or manager who is responsible for these tasks. It is known that entrepreneurs of small businesses can perform several functions within his company, which makes him, most of the time, overloaded with tasks. In this way, they do not have time to learn new tools, nor learn to deal with the reports produced by them, a fact that directly impacts the management of social media in business.

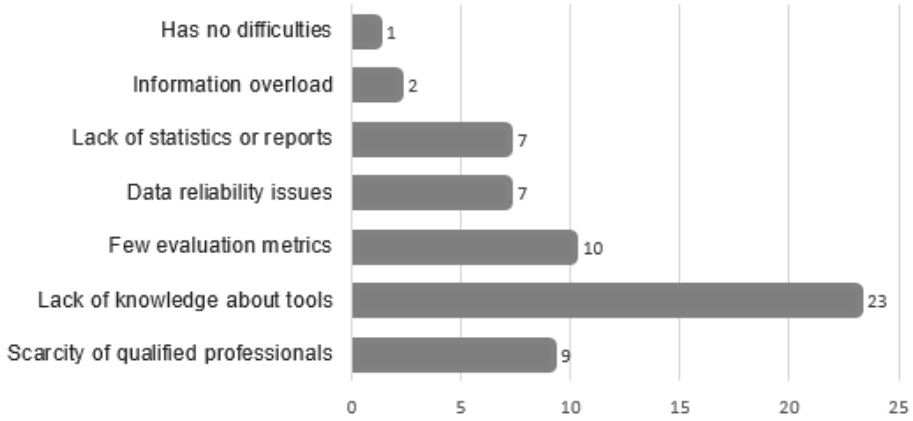


Fig. 1. Difficulties in managing results in Social Media.

Referring to Social CRM, Figure 2 says that 25 companies answered that they do not know how to conceptualize it. From the 6 that gave their answers, only 2 said that it is present in the entire process of relationship with customers. The other 4 responses were related only to the sales aspect. Some examples can be read in Figure 2.

Fig. 2. Conceptualization of Social CRM by small businesses.

4.2 Case studies

To conduct the case studies, four companies from different sectors were selected from the 31 companies that attended our courses. It is important to justify the omission of the names of the companies, as this was one of the criteria agreed upon between the researchers and the interviewees, namely: **Retail Company** - belongs to the retail footwear and clothing industry. In operation since 2009; **Food Company** - operates in the sector of supplying prepared foods, selling beverages, and the like. It has been on the market since 2018; **PET Company** - offering services for small animals, such as veterinary care and beautification. It has been in operation since 2018; **Photography Company** - its activities are described as a filming service organization. It has been offering its services since 2018. This information was collected on the Federal Government Redesim's free access platform, where it is possible to search in the National Register of Legal Entities using the companies' fancy name. All participants (interviewees) were the owners of the companies. More information about the profile of each interviewee can be seen in Table 5.

Table 5. General information of the interviewee.

	Retail	Food	PET	Photography
Degree	Administration	Psychology	Veterinary medicine	Veterinary medicine
Professional experience	Worked in a shoe store, in a real estate company, and seller of home appliances	No previous work experience	No previous work experience	Worked in his training area for two years
Favorite hobby	Non-informed	Enjoys reading and studying Gestalt Therapy	Loves going to the beach, but lately doesn't have much time	Go to the beach, ride a bike and go out to bars.

In the case of the Food, Photography, and PET companies, the interviewees reported that they themselves are responsible for managing the business's digital platforms. The company Retail faced two phases, in the beginning, the owner was the one responsible for taking care of social networks, while currently, a thirty-party company manages their social media channels. Only the Food and Retail companies have participated in

external consultancies offered by the Brazilian Micro and Small Business Support Service, known by its acronym SEBRAE.

Then, it was moved to the questioning phase about the motivations that the owners have in relation to the use of social media. The responses were very similar, focusing on “brand presence”, “product promotion” and “getting the customer's attention”. A relevant fact was that none of the three companies cited social media as a channel of communication with customers. Therefore, a gap is identified here in the use of these platforms for business. With regard to the external factors that lead to the adoption of these platforms, the companies Retail and Food cited the competitor’s pressure as a direct influence on the concern with the digital presence. PET talked about the collection it receives from customers in relation to the disclosure of procedures that are done at the clinic, how animals are treated, etc. Photography reported that it uses social media to evaluate competitors.

In sequence, the most used metric by companies was the number of likes that a publication receives. Only Retail cited the number of people interested in a particular product that was posted on social media, demonstrating that social media is being used for trend identification, which guides the portfolio selection process. After that, it was entered in the section regarding Social CRM and its processes. First, respondents were asked whether they knew how to conceptualize Social CRM. None of the four could answer. This fact was evident when the companies Retail and PET said that they have software responsible for cash flow, inventory control, customer database, and sales control. The Food company also said it uses software to control orders. The company Photography reported that it started measuring customer satisfaction after participating in a course related to the topic. In other words, companies have internal CRM and Social CRM processes, but they still do not understand what it is and where it is located.

Regarding the difficulties experienced, the Food portfolio selection process spoke about the limitations of the physical space, which is still small to serve the public it receives. Another point was the administration of multiple virtual service channels. Orders arrive via WhatsApp, phone, Instagram/Facebook messages, and ordering application - in addition to on-site assistance. One of the solutions was to concentrate orders, excluding calls and messages via Instagram and Facebook. Despite disliking some customers, there was an improvement in the speed of service and a reduction in any delays and errors in orders, increasing the level of satisfaction of service users.

In the PET company, the biggest problems are related to the time management of the owner, who sometimes needs to perform many functions and ends up being overloaded with work. An interesting dynamic in the relationship with the client was the use of WhatsApp as a tool for bringing together tutors and hosted pets - daily, the manager sends photos and videos of the pets (hosted or in daycare) to each tutor. Some of the photos were posted on Instagram, but this action was canceled, given that tutors who did not have their pets contemplated complained to the manager. It should be noted that the manager stated that she does not feel comfortable using Facebook, which is why the company abandoned this platform. Photography cited seasonality as a problem faced by the company. In periods of low demand, one solution seen as product diversification - such as photography courses, an increase in promotions, and the use of sponsored posts to increase the visibility of these actions.

Finally, questions were asked to the owners regarding the COVID-19 pandemic period. The PET company implemented the “Taxi Dog” service, in order to reduce the

impact caused by the decrease in the circulation of people and restrictions imposed by social isolation. The company Retail is investing in online sales through Facebook, Instagram, and WhatsApp, intensifying the campaigns carried out on the latter platform. The Food company stated that it will continue with the delivery service and that it will also take the time without moving in the physical space to start the renovation and expand the environment to be able to accommodate customers with social distancing. Photography said it was necessary to postpone the work and, currently, they are betting on the gastronomic photography market for restaurant delivery services.

After grouping the data and information obtained through the survey and case studies, individual reports were built for the four companies that participated in the interviews. For these reports, aspects such as feasibility and feasibility were taken into account in carrying out the suggested tasks. These tasks were grouped into positive points and points that could be improved with regard to the adoption of social media for business. Table 6 shows the information contained in the reports, which were also presented to the research partners in order to validate these suggestions.

Table 6. Social media evaluation reports.

	Positive points	Points to improve
Retail	Investments in "Boosting publication", contracting an outsourced service for managing social media, using WhatsApp Business as a communication channel, using software for managing internal processes.	Automation of responses in WhatsApp Business, interaction with the public on social media through stories, creation of the marketplace on social media, promotion of in-store promotions.
Food	Use of delivery service platforms; interaction with the public in comments and reposts; and active presence on social media.	Definition of digital presence and branding; study ways to adopt WhatsApp Business as a communication channel; company registration on Google My Business; adoption of software for publishing automation.
PET	The excellent interval between posts, use of WhatsApp Business as a communication channel, search for customer feedback, constant innovation in services.	Automation of responses in WhatsApp Business, use of social media for market research, not feeding all social media of the company, standardization of service.
Photography	Adaptation of the business in line with market trends, investments in "boosting publication", concern in optimizing the feedback process with customers, and market variability.	Studying ways to adopt WhatsApp Business as a communication channel with customers, building a database with customer information, adopting anonymous feedback strategies.

5 Lessons learned and perspectives

From the results, it is possible to see that companies still make use of a few functionalities that are made available within social media, for that of Customer Relationship Management. For example, the metrics used by entrepreneurs are mainly the number of likes, which is just one of the data offered by the Facebook and Instagram platforms to assess the success of a publication. If they also had access to the engagement data of the publications and reach of people, they could more broadly identify what content is being well received by the public. With this, they would be able to build more targeted and assertive campaigns, and, consequently, save efforts in time and money, which are two of the main pains reported by managers.

A point that is also common among the four companies studied is the use of WhatsApp Business as a business tool. This platform can assist them in several ways, such as automating frequent responses, in which the manager saves time in the first contact with potential customers. In addition, WhatsApp Business can serve as a direct communication channel with customers to disseminate news and promotions and act as a catalog of products and services, in order to optimize the work of the social media manager.

Another aspect that was observed has a direct relationship with the branding of companies. According to [23, 24], branding can be defined as the set of strategies that define the brand. It is everything that involves it: from colors and shapes to define the value of products and services and sales practices. Moreover, for small companies, this can make a huge difference, since it serves to position the brand in the market and help to identify it among others. This process can be initiated with the definition of the values perceived in the products, logo creation, and colors and visual elements definition, which will identify the company. After these steps, the visual identity creation process can begin, in which the way in which the company will communicate with its public will be defined. Then, we move on to the digital presence stage, where the company will set up its social media channels.

These channels are defined according to the company's niche. For example, the food company cited iFood as one of its sales channels. In addition, for the Retail company, the report sent to the manager indicated the configuration of the store's marketplace, where he will be able to count on another online sales channel, in addition to the presence. This modality is known for bringing together in a single place, several offers of products and services from different companies, such as an online catalog (e.g. Amazon, Mercado Livre, OLX, etc.). Facebook currently has its own marketplace tool. The correct definition of social media can make a big difference in the sales volume achieved by companies.

To assist managers in these adjustments, it is, therefore, necessary to offer individualized courses and consultancies, with the purpose of guiding them in the adjustments that were suggested in the reports. It is also important to assist them in understanding metrics for evaluating results and how to optimize their internal processes. With this, it is expected to positively impact its business in the growth of brand visibility and increase of sales.

6 Conclusions

This article analyzed quantitative aspects, through a survey, and qualitative aspects, with the help of case studies, of the implementation of Social CRM in Micro and Small Companies, a sector that is highly relevant for Brazilian trade. The purpose of these analyzes was to identify which aspects of Social CRM are relevant to MSCs (RQ1) and how they implement them in their business (RQ2).

The results referring to RQ1 show that entrepreneurs classify the internet as a highly relevant means of communication for their businesses, with Facebook, Instagram, and WhatsApp as the main platforms. They use these media to publicize their products and promotions, in addition to disseminating news and promoting sweepstakes. However, only a portion of managers uses them as a channel for seeking feedback from their customers, which ends up neglecting the loyalty process. The results also show that entrepreneurs are unaware of Social CRM concepts and applications. Additionally, the lack of this knowledge means that they are unable to achieve better results since they themselves are responsible for managing social media. As a solution to this problem, short and asynchronous training on this topic can be offered to managers.

Regarding RQ2, the results obtained indicate that the main reason for using social media is directly linked to the presence of the brand. However, as in the previous phase, it became evident that managers still do not see social networks as a two-way communication channel, which is used only for the broadcasting of products and services. A fact that drew attention was that entrepreneurs implement CRM processes in their companies, such as using software for cash and inventory management, at the same time; they report that they do not know the concept, which also contributes to the neglect of related functionalities. Despite this, the pandemic forced micro and small companies to adapt to the new reality, demonstrating their adaptability with actions that include launching new service packages and promotions, in addition to offering new services through online sales and delivery.

In view of the results, this work contributes to the visualization of the main gaps in the Customer Relationship Management sector in social media, in common among small companies. We concluded that Social CRM is still an unexplored environment for micro and small companies, but with great potential to boost sales, develop customer loyalty and increase brand visibility. The lessons learned have the potential of guiding policymakers to create more adequate measures for boosting this market sector. As future work, it is intended to plan a set of actions with an interventionist view, in order to help entrepreneurs to explore the functionalities of social media, thus contributing to the optimization of their internal and external processes.

7 Acknowledgment

We would like to thank the Federal University of Western Pará and the National Council for Scientific and Technological Development (CNPq) for their financial support for conducting this work, and the research partners from the State University of Maranhão for their direct collaboration in conducting the research. We would also

like to thank the insights given by the reviewers of the article that will help us to improve the work even more.

References

1. Bartik, A.W., Bertrand, M., Cullen, Z., Glaeser, E.L., Luca, M., Stanton, C.: The impact of COVID-19 on small business outcomes and expectations. Proc. Natl. Acad. Sci. U. S. A. 117, 17656–17666 (2020). https://doi.org/10.1073/pnas.2006991117.
2. Elrhim, M.A., Elsayed, A.: The Effect of COVID-19 Spread on the E-Commerce Market: The Case of the 5 Largest E-Commerce Companies in the World. SSRN Electron. J. 1–14 (2020). https://doi.org/10.2139/ssrn.3621166.
3. Kim, R.Y.: The Impact of COVID-19 on Consumers: Preparing for Digital Sales. IEEE Eng. Manag. Rev. 48, 212–218 (2020). https://doi.org/10.1109/EMR.2020.2990115.
4. Nisar, T.M., Prabhakar, G., Strakova, L.: Social media information benefits, knowledge management and smart organizations. J. Bus. Res. 94, 264–272 (2019). https://doi.org/10.1016/j.jbusres.2018.05.005.
5. We Are Social: Digital 2021 - We Are Social, https://wearesocial.com/digital-2021, last accessed 2021/01/27.
6. Alalwan, A.A., Rana, N.P., Dwivedi, Y.K., Algharabat, R.: Social media in marketing: A review and analysis of the existing literature. Telemat. Informatics. 34, 1177–1190 (2017). https://doi.org/10.1016/j.tele.2017.05.008.
7. Kim, A.J., Johnson, K.K.P.: Power of consumers using social media: Examining the influences of brand-related user-generated content on Facebook. Comput. Human Behav. 58, 98–108 (2016). https://doi.org/10.1016/j.chb.2015.12.047.
8. Lobato, F., Pinheiro, M., Jr, A.J.: Social CRM: Biggest Challenges to Make it Work in the Real World. Int. Conf. Bus. Inf. Syst. 303, 221–232 (2017). https://doi.org/10.1007/978-3-319-69023-0.
9. Marolt, M., Zimmermann, H.D., Žnidaršič, A., Pucihar, A.: Exploring social customer relationship management adoption in micro, small and medium-sized enterprises. J. Theor. Appl. Electron. Commer. Res. 15, 38–58 (2020). https://doi.org/10.4067/S0718-18762020000200104.
10. Woodcock, N., Green, A., Starkey, M.: Social CRM as a business strategy. J. Database Mark. Cust. Strateg. Manag. 18, 50–64 (2011). https://doi.org/10.1057/dbm.2011.7.
11. Reinhold, O., Alt, R.: How companies are implementing Social Customer Relationship Management. Insights from two case sudies. 26th Bled eConference. 206–221 (2013).
12. Felix, R., Rauschnabel, P.A., Hinsch, C.: Elements of strategic social media marketing: A holistic framework. J. Bus. Res. 70, 118–126 (2017). https://doi.org/10.1016/j.jbusres.2016.05.001.
13. Rosenberger, M., Lehmkuhl, T., Jung, R.: Conceptualising and Exploring User Activities in Social Media. In: IFIP International Federation for Information Processing. pp. 107–118. Springer Verlag (2015). https://doi.org/10.1007/978-3-319-25013-7_9.
14. de León-Sigg, M., Vázquez-Reyes, S., Villa-Cisneros, J.L.: Factores que Afectan la Adopción de Tecnologías de Información en Micro y Pequeñas empresas: Un Estudio Cualitativo. RISTI - Rev. Iber. Sist. e Tecnol. Inf. 20–36 (2017). https://doi.org/10.17013/risti.22.20-36.
15. Wang, Z., Kim, H.G.: Can Social Media Marketing Improve Customer Relationship Capabilities and Firm Performance? Dynamic Capability Perspective. J. Interact. Mark. 39, 15–26 (2017). https://doi.org/10.1016/j.intmar.2017.02.004.
16. Longaray, A.A., Anselmo, C.R., Maia, C., Lunardi, G., Munhoz, P.: Análise do emprego do F-commerce como impulsionador do desempenho organizacional em micro e pequenas empresas no Brasil. RISTI - Rev. Ibérica Sist. e Tecnol. Informação. 27, 67–

85 (2018). https://doi.org/10.17013/risti.27.67-85.
17. Ainin, S., Parveen, F., Moghavvemi, S., Jaafar, N.I., Shuib, N.L.M.: Factors influencing the use of social media by SMEs and its performance outcomes. Ind. Manag. Data Syst. 115, 570–588 (2015). https://doi.org/10.1108/IMDS-07-2014-0205.
18. Rodrigues Chagas, B.N., Nogueira Viana, J.A., Reinhold, O., Lobato, F., Jacob, A.F.L., Alt, R.: Current Applications of Machine Learning Techniques in CRM: A Literature Review and Practical Implications. Proc. - 2018 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2018. 452–458 (2019). https://doi.org/10.1109/WI.2018.00-53.
19. Alford, P., Page, S.J.: Marketing technology for adoption by small business. Serv. Ind. J. 35, 655–669 (2015). https://doi.org/10.1080/02642069.2015.1062884.
20. Hancock, D.R., Algozzine, B.: Doing Case Study Research. , New York (2006).
21. Baah-Ofori, R., Amoako, G.K.: Electronic Customer Relationship Management (E-CRM) Practices of Micro, Small, and Medium Scale Enterprises in Ghana. In: Strategic Customer Relationship Management in the Age of Social Media. pp. 72–94 (2015). https://doi.org/10.4018/978-1-4666-8586-4.ch005.
22. Imme, A.: As 10 Redes Sociais mais usadas no Brasil em 2020, https://resultadosdigitais.com.br/blog/redes-sociais-mais-usadas-no-brasil/, last accessed 2020/07/31.
23. Swaminathan, V., Sorescu, A., Steenkamp, J.B.E.M., O'Guinn, T.C.G., Schmitt, B.: Branding in a Hyperconnected World: Refocusing Theories and Rethinking Boundaries. J. Mark. 84, 24–46 (2020). https://doi.org/10.1177/0022242919899905.
24. Key, B.A., Tool, M.: Branding: A Key Marketing Tool. (1992). https://doi.org/10.1007/978-1-349-12628-6.