Faculdades Integradas de Taquara - Faccat
Av. Oscar Martins Rangel, 4.500
Taquara, RS, CEP 95612-150

Curso de Sistemas de Informação

MINERAÇÃO RE
MINERAÇÃO DE DADOS COM BASE NO RECLAME AQUI

Felipe Garcia Rocha
Faculdades Integradas de Taquara – Faccat – Taquara – RS – Brasil
feliperocha@sou.faccat.br

Leonardo Augusto Sápiras Professor Orientador
Faculdades Integradas de Taquara – Faccat – Taquara – RS - Brasil
sapiras@faccat.br

Resumo

As opiniões dos usuários em plataformas na internet, como redes sociais e demais canais, permitem expor perspectivas ou experiências em relação a produtos ou serviços utilizados, que são comuns no cotidiano dos consumidores. As empresas que possuem um enorme volume de opiniões com origens diversificadas, tendem a ser mais demoradas na análise e no *feedback* aos usuários em relação a essas opiniões. Este artigo apresenta um projeto de pesquisa que tem como objetivo desenvolver um sistema que encontra as principais reclamações em empresas do mesmo segmento, utilizando o site Reclame Aqui como base de dados. Esse trabalho limita-se a aplicar o sistema desenvolvido no segmento de telecomunicações. Com base nos dados coletados e processados foi possível encontrar os termos de maior incidência nas reclamações dos usuários.

Palavras-chave: Mineração de Dados, Mídia Social, Reclame Aqui.

MINING RE
DATA MINING BASED ON THE COMPLAINT HERE

Abstract

The opinions of users on platforms on the Internet, such as social networks and other channels, allow exposing perspectives or experiences in relation to products or services used, which are common in the daily lives of consumers. Companies that have a huge volume of opinions with diverse origins, tend to take longer to analyze and provide feedback to users regarding these opinions. This article presents a research project that aims to develop a system that finds the main complaints in companies in the same segment, using the Reclame Aqui website as a database. This work is limited to applying the system developed in the telecommunications segment. Based on the data collected and processed, it was possible to find the terms with the highest incidence in user complaints.

Keywords: *Data Mining, Social Media, Reclame Aqui****.***

1 INTRODUÇÃO

A tecnologia tem avançado constantemente e se tornado cada vez mais presente em nosso cotidiano, seja através dos *smartphones*, dos computadores, dos tablets e entre outros meios de acesso à internet. Por meio de fóruns, *blogs*, emails e sites de opiniões, os consumidores têm à disposição maneiras para compartilhar suas experiências com as empresas que tiveram algum contato ou que pretendem ter, logo, a opinião geral dos consumidores na internet em relação às empresas é fator primordial para o marketing na *Web* (PANG; LEE, 2008; SAPIRAS, 2015).

No Brasil, mais precisamente no primeiro semestre de 2019, houve um crescimento consecutivo de mais 12% no *e-commerce*, isso representa uma movimentação em torno de 26,4 bilhões de reais, durante este período, 5.3 milhões de usuários estavam realizando sua primeira compra *online*. As compras *online* são motivadas 25% pelos sites de busca, como o Google por exemplo, o segundo maior motivador são as redes sociais, com 19%, destaque maior para o Facebook, que ocupa 53% deste contexto. Considerando que os clientes que compraram pelas redes sociais são os que acabam elogiando mais a compra que foi realizada (EBIT, 2019).

O usuário está cada vez mais exigente, demandando cada vez mais experiência das empresas *online*, de forma simples, fácil e rápida (EBIT, 2019). Um fator muito importante é o ponto de vista dos usuários em relação a um determinado produto, este ponto de vista é

determinado devido a três fatores: atenção seletiva, distorção seletiva e retenção seletiva (ARANTES, 2016). A voz dos usuários em relação ao consumo online passou a ter uma presença fundamental, valorizam a criatividade, tecnologia e recusam banalidades do consumo (ARANTES, 2016).

De acordo com Bhatia, Chaudhary e Dey (2020, p. 1), o principal componente que pode melhorar a qualidade de um serviço ou produto é a opinião dos usuários. Segundo Bhatia, Chaudhary e Dey (2020, p. 1), fatos são sentenças verdadeiras e podem ser verificadas, enquanto opiniões contém uma certa crença e não podem ser verificadas como uma verdade absoluta, a mineração de dados deve atuar na parte subjetiva da informação, que inclui as opiniões e as emoções, com a crescente popularidade dos sites de compras online, mídias sociais e outros sites de interação, isso gerou milhares de postagens na *World Wide Web* (WWW). Segundo Bhatia, Chaudhary e Dey (2020, p. 2), a necessidade de um sistema automatizado para a análise lógica das informações textuais, permitirá que as empresas investiguem uma grande quantidade de dados a serem processados em um curto espaço de tempo e, entender isto, é um ponto crucial. Neste contexto, como comparar produtos ou serviços entre empresas diferentes? Não foram encontradas soluções que respondam essa pergunta.

O objetivo geral desta pesquisa científica é desenvolver uma ferramenta que seja capaz de sumarizar as principais palavras chaves, com base nas opiniões dos usuários do site Reclame Aqui, referente às empresas de telecomunicações. Como objetivos específicos lista-se: (i) pesquisar ferramentas e metodologias para extrair dados do site Reclame Aqui; (ii) desenvolver a geração de relatórios e (iii) desenvolver informativos ao usuário desta ferramenta que existe reclamação a respeito de sua empresa sem resposta.

Para operar a ferramenta, primeiro deve ser realizada a configuração de seus parâmetros no módulo de configurações, após realizada as configurações conforme o segmento da empresa, o usuário terá que atualizar a base de dados, utilizando as opções disponíveis no módulo de configurações. Após finalizadas as atualizações necessárias, estarão disponíveis as informações referentes às empresas e outras do mesmo segmento na página principal.

A ferramenta opera conforme as empresas predefinidas pelo usuário. Ao realizar a atualização das informações é iniciado o processo de *Web Scraping*, em que é acessado o site do Reclame Aqui, realizando-se a extração dos seguintes elementos: os comentários referentes às empresas cadastradas, dados das empresas e demais estatísticas. Estes dados coletados são armazenados no banco de dados e posteriormente ocorre o processamento de linguagem natural.

Logo em seguida estes dados estão disponíveis na plataforma *Web* para tomadas de decisão do usuário.

A estrutura deste artigo é regida pela seguinte ordem: um referencial teórico que descreve os assuntos abordados na seção 2; a metodologia de desenvolvimento do *software* proposto, que é demonstrada na seção 3; o funcionamento do sistema desenvolvido é relatado na seção 4; por fim, as conclusões sobre o trabalho são demonstradas na seção 5.

2 REFERENCIAL TEÓRICO

Nesta seção será apresentado os conceitos que envolvem este projeto de pesquisa, demonstrando as fontes utilizadas e as referências para criação do *software* Mineração RE.

2.1 Processo de descoberta de informações em bases de dados

Segundo Kantardzic (2020, p. 2) mineração de dados é um processo de descoberta de vários modelos, resumos e derivados, com base em uma coleção de dados, em que o termo "processo" não consiste em apenas selecionar uma ferramenta para solucionar um determinado problema. O motivo para isso é que a mineração de dados não é apenas um conjunto de ferramentas isoladas, cada uma é compatível com determinado problema. O que acontece na prática é um processo interativo em que é realizado um estudo dos dados, onde são examinados por meio de uma técnica analítica e, ao visualizar os resultados por outra perspectiva, talvez seja necessário modificar e retornar para o início do processo, utilizando outra estratégia, para assim atingir melhores resultados.

As etapas envolvidas na descoberta de informações em bases de dados são: seleção, limpeza dos dados, transformação, mineração de dados, avaliação e por último apresentação. São etapas comuns entre as estruturas KDD, CRISP-DM e SEMMA, que serão abordadas nas seções a seguir, quando serão demonstradas as técnicas de mineração de dados e suas metodologias de funcionamento (ARBEX *et al.*, 2019).

2.1.1 KDD

O processo KDD (*Knowledge Discovery in Databases*) realiza a descoberta de conhecimento dentro de um conjunto de dados, como um banco de dados, por exemplo. Em

KDD uma das etapas é a mineração de dados, sendo responsável pelo reconhecimento de padrões conforme foram predefinidos pelo usuário. As demais etapas do processo estão presentes apenas para garantir a extração de conhecimento útil ao usuário, na primeira etapa de compreensão do projeto, é realizada uma análise do projeto que precisa ser desenvolvido, desta forma, a equipe terá um maior entendimento do objetivo que precisar ser atendido. É realizada uma investigação a respeito dos requisitos que precisam ser desenvolvidos e quais os meios serão utilizados, em conjunto com o usuário são definidas as metas conforme suas expectativas (DADERMAN; ROSANDER, 2018).

Na segunda etapa, denominada seleção, é criado um conjunto de dados, estes dados podem já estar disponíveis ou ser necessários integrá-los a partir de um sistema externo, podendo estar concentrado em subconjuntos de variáveis ou em amostras de dados. Na terceira etapa ocorre o pré-processamento, é realizada uma limpeza dos dados, removendo regras gramaticais, preposições e entre outros caracteres conforme a necessidade do projeto. A quarta etapa é a transformação, quando é realizada uma preparação dos dados para a mineração de dados, durante essa fase pode ser adotado recursos de representação de algum dado específico (DADERMAN; ROSANDER, 2018).

Na quinta etapa é realizada a mineração de dados, que consiste em três fases. Na primeira fase os dados são preparados conforme o método de mineração escolhido, podendo ser classificação, regressão ou agrupamento. A segunda fase da mineração de dados é escolher um algoritmo para realização da mineração, este algoritmo será responsável pelo reconhecimento de padrões conforme os parâmetros definidos pelo usuário. Na terceira e última fase, a mineração de dados é implementada e, por meio de um conjunto de dados, são encontrados padrões interessantes ao usuário, essa última fase pode ser repetida diversas vezes até que um resultado satisfatório seja encontrado.

Na sexta etapa ocorre a avaliação ou interpretação, por meio dos resultados obtidos na etapa anterior que é a de mineração de dados, é realizada uma avaliação em relação aos objetivos determinados na primeira etapa do processo KDD, podendo também ser necessário retornar em alguma das etapas anteriores para ser realizada algumas alterações necessárias. Após o KDD com o resultado obtido, a próxima etapa será utilizar este conhecimento novo para implementar em conjunto com algum sistema ou aplicar em alguma documentação ou relatório (DADERMAN; ROSANDER, 2018).

2.1.2 CRISP-DM

A técnica de mineração CRISP-DM(*Cross-Industry Standard Process of Data Mining*), tem como vantagem não ser dependente de uma ferramenta para ser executada. Essa técnica está dividida em seis etapas que são:

- Compreensão do negócio;
- Entendimento dos dados;
- Preparação dos dados;
- Modelagem;
- Avaliação;
- Desenvolvimento.

Na etapa de compreensão do negócio deverá ser identificado o problema que o envolve, devendo ser entregue uma explicação de como o projeto irá sanar esse problema atual, tornando explícito o principal objetivo da solução apresentada e também deverá ser bem claro qual métrica utilizada para análise posteriormente dos resultados. A etapa de entendimento dos dados, é responsável pela coleta dos dados, descrições e estatísticas, sendo uma etapa que demanda bastante tempo de análise por parte do *Data Scientist* (DADERMAN; ROSANDER, 2018).

Em preparação dos dados é dividida em quatro fases que são: seleção dos dados, é o momento em que se deve escolher os dados de maior relevância para utilização no modelo, sendo necessário documentar o motivo da escolha, pois será um facilitador futuramente; Limpeza dos dados, consiste na realização de formatações de dados incorretos e remoção de caracteres indesejáveis; Construção de dados, pode ocorrer de não haver o dado necessário a disposição, sendo necessário criá-lo, isso ocorre com datas em que há feriados por exemplos, havendo a necessidade de sinalizar estes dados; A quarta e última fase é a integração de dados, em que consiste na união de duas fontes de dados distintas (DADERMAN; ROSANDER, 2018).

Durante a etapa de modelagem, com base nas etapas anteriores, será desenvolvido o modelo utilizando o algoritmo mais adequado, pode haver mais de um modelo e assim ser realizadas comparações para ao final obter o melhor dentre eles. A etapa de avaliação é o momento de analisar os resultados com base nos critérios definidos na primeira etapa, caso não tenha correspondido à expectativa, deverá ser refinado o escopo e executados novamente. A última etapa é o desenvolvimento, é o momento de colocar em produção o modelo produzido (DADERMAN; ROSANDER, 2018).

2.1.3 SEMMA

Essa metodologia é semelhante a CRISP-DM, porém durante as etapas do projeto é permitido ir e retornar, ou realizar repetição de alguma etapa, desta forma o desenvolvimento do projeto não fica restrito à finalização de uma etapa para prosseguir para a próxima. SEMMA possui cinco etapas diferentes, que são: a amostra, nesta primeira etapa é realizada a coleta dos dados, que devem ser grandes o suficiente para possuírem informações relevantes, porém também devem ser pequenos o suficiente da perspectiva de processamento dos dados. A segunda etapa é a exploração, onde ocorre a pesquisa por padrões e relacionamentos entre os dados, fazendo uso de visualização ou uma análise estatística. A terceira etapa é modificar, com base na etapa anterior os dados são modificados conforme os requisitos do modelo, podendo ser incluídos novos dados ou variáveis. Na quarta etapa o modelo começa ser criado, aplicando novas modificações com base na etapa anterior, assim se tornando um modelo mais confiável, sendo capaz de prever um resultado ou classificar um dado desconhecido. A última etapa que é a avaliação, com base nas amostras dos resultados e desempenho gerados pelo modelo, será realizado validações e testes para definir se o modelo é útil ou não (DADERMAN; ROSANDER, 2018).

Dentre essas técnicas apresentadas, este projeto de pesquisa foi desenvolvido utilizando a técnica KDD, por conta de possuir etapas bem definidas e também por principalmente ter um foco maior na descoberta de conhecimento e isso tem um alinhamento em relação ao objetivo deste projeto. Sendo adotado como método de mineração de dados a classificação, que segundo Fernando Amaral (2016, p. 88) funcionam como dados históricos, pois são fatos já ocorridos, é por meio destes dados históricos que será desenvolvido um modelo. Neste projeto de pesquisa foi utilizado um dicionário, que já contém as classificações de cada palavra de acordo com sua classe gramatical, o modelo desenvolvido tem como objetivo analisar apenas as palavras classificadas como adjetivos e, para isso, são realizadas as comparações com o dicionário para obter a informação de qual palavra é um adjetivo ou não.

2.2 *Web Scraping*

É um processo de coleta automatizada de dados da internet. Esta coleta ocorre por meio de um programa ou API, cujo programa realizará uma requisição a um servidor *Web* que por sua vez entregará um resultado em HTML e outros arquivos que compõem as páginas *Web*.

Logo em seguida este programa analisará o conteúdo para extrair as informações necessárias (MITCHELL, 2018).

A utilização de *Web Scraping* agiliza o processo de coleta de dados, por serem excelentes em analisar grandes quantidades de dados, pois permitem analisar diversas páginas simultaneamente, ao invés de ser apenas um usuário olhando manualmente em uma única janela estreita cheia de informações. Existem APIs que fornecem grandes volumes de informações, porém geralmente há algumas limitações dos parâmetros ou das requisições, que podem acabar influenciando nos objetivos que se pretende alcançar. Por conta disto o *Web Scraping* é amplamente utilizado, todos os dados que estão públicos e disponíveis na internet. Não possuindo as limitações que ocorrem em APIs (MITCHELL, 2018).

2.3 Plataformas de Opiniões

Os usuários possuem diversos canais na internet para expor suas opiniões referentes a produtos ou serviços. No Brasil existe um canal especializado em reclamações denominado Reclame Aqui, sendo que além das reclamações, este site também acaba servindo de referência para pesquisas dos consumidores. O Reclame Aqui recebe mais de 92% de consulta dos consumidores sobre a reputação de uma determinada empresa antes de realizarem uma compra na internet, este portal de reclamações recebe um grande volume de visualizações contando com em torno de 42 milhões de *views* por mês, possuem 15 milhões de consumidores cadastrados e em torno de 120.000 empresas cadastradas (Reclame Aqui, 2020).

Há também um outro portal *online* denominado Consumidor.gov.br que fundamentou-se no disposto no artigo 4º inciso V da Lei n. 8.078/1990 e no artigo 7º, incisos I, II e III, do decreto n. 7.963/2013, que possui como objetivo resolver conflitos de consumo de forma extrajudicial sendo de maneira rápida e eficiente. Cerca de 80% das reclamações presentes neste portal são resolvidas pelas empresas no prazo médio de sete dias, o portal atualmente conta com 1.388.487 usuários e 519.000 empresas cadastradas, contando com um total de 1.898.217 de reclamações finalizadas (FIGUEIREDO, 2020). Apesar da plataforma “Reclame Aqui” e o portal “Consumidor.gov.br” possuirem semelhanças em seus propósitos, cada uma tem sua respectiva particularidade, no caso do Reclame Aqui é aberto para receber reclamações de todas as empresas disponíveis no mercado, no portal Consumidor apenas é aceito receber reclamações das empresas que se cadastraram voluntariamente na plataforma e com isso gera um impacto diretamente nas quantidades de reclamações referente às empresas cadastradas (SOUSA *et al.*, 2020). Neste projeto de pesquisa foi utilizado apenas a plataforma do Reclame Aqui, por ser

mais difundida entre os usuários.

2.4 Trabalhos Relacionados

2.4.1 API Reclame Aqui

Existe uma seção presente no site do Reclame Aqui que realiza comparações entre empresas diferentes, estando disponível o acesso aos consumidores, este módulo traz algumas informações referente ao tempo de resposta, reputação, quantidade de reclamações e os tópicos mais reclamados, conforme é demonstrado na Figura 1.

Figura 1 – Comparativo entre Empresas

Fonte: Reclame Aqui (2020)

O principal diferencial deste projeto de pesquisa em relação aos comparativos entre as empresas e a API do Reclame Aqui, é a possibilidade do usuário saber os principais termos utilizados pelos consumidores nas reclamações presentes no Reclame Aqui, há também este módulo de comparação que funciona de maneira similar, porém é demonstrado em números as quantidades de reclamações referente a cada tópico.

2.4.2 Análise comparativa das principais plataformas de reclamações online

É realizada uma análise comparativa entre os dois maiores portais de reclamações *online* no Brasil, sendo eles o Reclame Aqui e o Consumidor.gov. Essa análise usou como referência o segmento de *ecommerce*, utilizando como base as maiores empresas deste segmento no período analisado, que foi de 26/01/2019 a 26/01/2020 com uma coleta total de 286.566 reclamações de ambas plataformas, conforme demonstrado na Quadro 1.

Quadro 1 - Demonstrativo das reclamações analisadas no Consudor.gov e no Reclame Aqui

	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%

Fonte: SOUSA *et al.* (2020).

Por meio dos dados coletados também foi aplicado o grau de legibilidade através do *Flesch Readability Ease Score (FRES)*, através do grau de legibilidade é possível realizar uma estimativa do grau de escolaridade dos usuários. Assim, foi obtido como resultados referente ao portal Consumidor.gov, que seu público maior é de pessoas do 6° ao 9° ano e no Reclame Aqui um maior acesso de pessoas com ensino médio ou nível superior. Através da análise das reclamações foi possível constatar as principais reclamações abordadas pelos usuários, agrupadas em tópicos conforme é apresentado na Quadro 2.

Quadro 2 - Tópicos das reclamações mais utilizadas pelos usuários referente as empresas analisadas

	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	

Fonte: SOUSA *et al.* (2020).

As diferenças dos resultados apresentados nesta análise comparativa entre os principais *e-commerce* do Brasil com este projeto de pesquisa desenvolvido, é que no presente trabalho os resultados são apresentados aos usuário por meio de uma interface *Web* e os parâmetros podem ser definidos conforme a necessidade do usuário através desta interface. Os termos apresentados na análise comparativa foram agrupados em tópicos, ou seja, não é a palavra em si que ocorreu nas reclamações, mas sim um conjunto de palavras. Neste projeto de pesquisa os termos apresentados não são agrupados, apenas contabilizados e expostos ao usuário por meio de uma interface *Web*.

3 METODOLOGIA

Esta seção detalha o embasamento literário da arquitetura, da metodologia e das tecnologias utilizadas para a realização do projeto

3.1 Metodologia de desenvolvimento

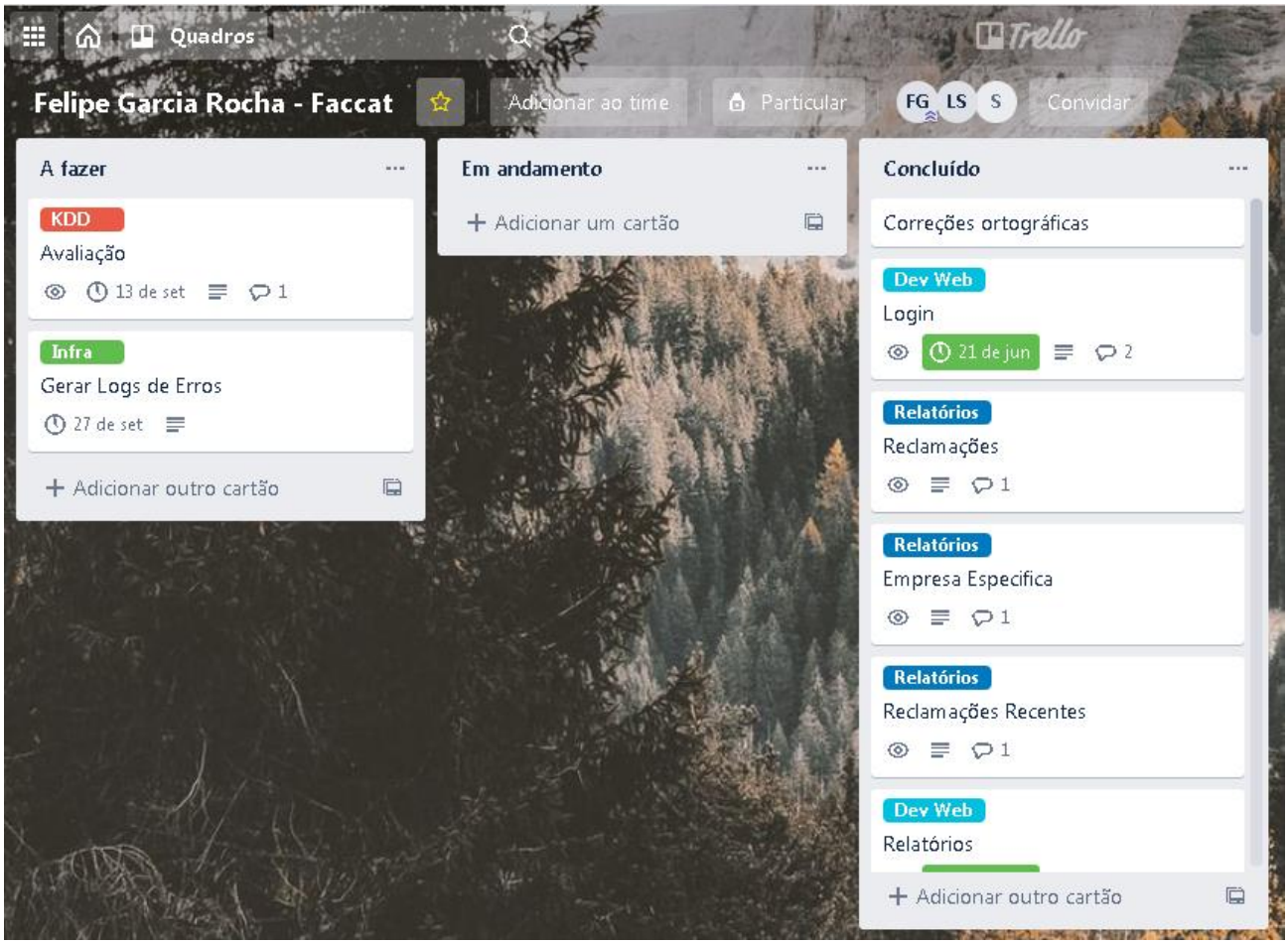
3.1.1 Análise de Requisitos e Modelagem do Sistema

O site Reclame aqui oferece acesso a uma API pelo qual é possível realizar requisições

para obter as reclamações postadas pelos usuários, porém, possibilita acesso somente aos dados referentes à empresa pertencente ao usuário, sendo assim não é possível ter acesso às reclamações de empresas terceiras, impossibilitando comparativos entre as empresas de forma dinâmica, e a utilização da API é um serviço pago.

Primeiro foi realizada a identificação dos requisitos funcionais primordiais para entregar o primeiro MVP (*Minimum Viable Product*), que é uma versão básica do *software* final, gerando o resultado esperado pelo objetivo geral da pesquisa (RIES, 2011). A organização do desenvolvimento destes requisitos foi realizada por meio da ferramenta Kanban, que é um quadro dividido em colunas em que cada coluna representa os estados possíveis de cada requisito ser desenvolvido. Neste trabalho estas colunas são divididas em: a fazer, em andamento e concluído, os requisitos são identificados por meio de "cartões", os quais descrevem o que deve ser feito e o nível de prioridade (HAMMARBERG; SUNDEN, 2014). Por meio da plataforma Trello foi realizado todo o controle da ferramenta Kanban, como o recorte demonstrado na Figura 2.

Figura 2 – Recorte do quadro Kanban virtual do Trello com uma parte das atividades realizadas



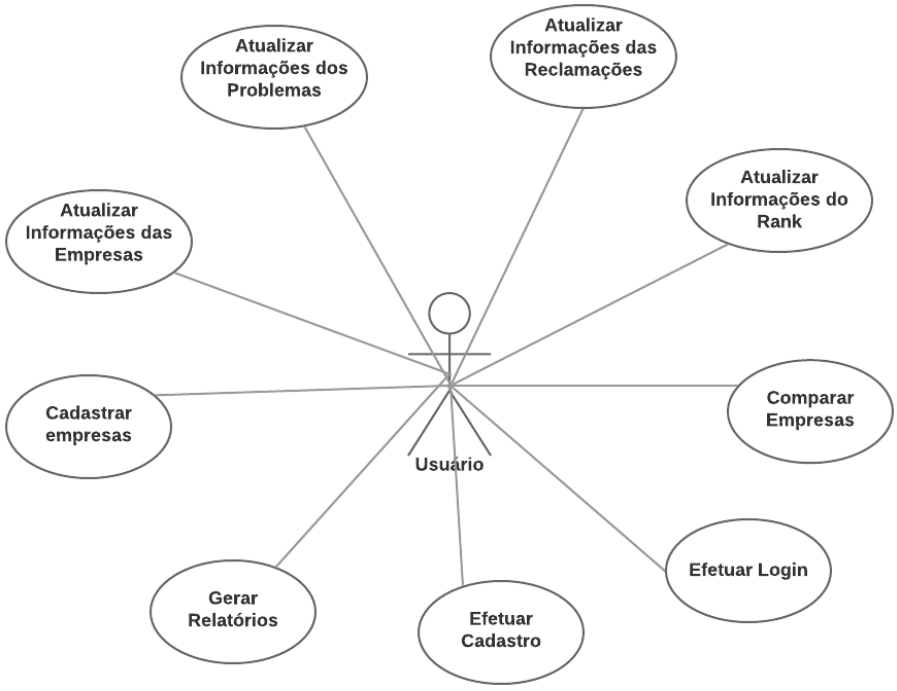
Fonte: Autor (2020).

A metodologia utilizada no desenvolvimento da aplicação foi a iterativa e incremental, na qual a cada requisito desenvolvido há uma iteração, e caso apresente o resultado esperado é homologado e incrementado ao projeto principal. Os requisitos estão representados com a ferramenta sistemática Kanban (PRESSMAN; MAXIM, 2016).

A Figura 3 ilustra os casos de uso mapeados na fase de análise. Tendo em vista o modelo iterativo incremental, alguns casos de uso foram elencados ao longo do desenvolvimento da ferramenta. Os principais casos de usos disponíveis na primeira versão da ferramenta foram: “Cadastrar Empresas”, que possibilita a inclusão de empresas pertencentes ao usuário e seus respectivos concorrentes. As opções de atualizar informações: das empresas, reclamações e problemas, que permitem a adição de novas informações por meio do *Web Scraping* que é a extração de dados do site Reclame Aqui.

A atualização das informações do *ranking* inicia o processo linguagem natural, com base nas reclamações cadastradas no banco de dados. A comparação entre empresas possibilita distinguir o perfil comportamental entre cada empresa em relação aos seus consumidores, e assim encontrar possibilidades de novos mercados ou pontos de melhorias.

Figura 3 – Diagrama de caso de uso



Fonte: Autor (2020).

3.1.2 Tecnologias utilizadas no desenvolvimento

Para o desenvolvimento dos requisitos de *Web Scrapy* e mineração de dados foi utilizada a linguagem de programação Python, por possuir uma alta legibilidade do código produzido, tornando o processo de desenvolvimento mais eficiente, possibilitando também uma prototipagem rápida, com código robusto e sustentável (HILLARD, 2020).

O banco de dados foi escolhido seguindo o modelo relacional, gerenciado por meio do MySQL, que é um sistema de gerenciamento de banco de dados (SGBD). Através deste sistema de gerenciamento são criadas as tabelas e colunas que compõem a base de dados, que ficarão disponíveis para consultas, atualizações, inserções e exclusões.

A biblioteca Selenium é responsável por realizar a extração de dados de sites, opera através do Chromium, que é um navegador de código aberto desenvolvido pela Google. Por meio do Selenium e juntamente com Python, foi possível definir os campos que desejamos extrair os dados, estes dados serão salvos no banco de dados em SQL.

A interface *Web* foi desenvolvida com HTML, CSS e *Javascript*. Para gerenciar a comunicação *Web* com os demais módulos foi utilizado a linguagem de programação PHP que é uma linguagem interpretada, de alto nível e tipagem baixa. Como servidor *Web* foi utilizado o Apache que é responsável pelas requisições dos navegadores (CANALTECH, 2020). O *framework* Laravel é responsável pela organização do projeto, das dependências e a inicialização dos *scripts* Python quando necessários, e também realiza a apresentação da ferramenta no navegador ao usuário. Este *framework* usa o padrão MVC, que define a arquitetura da ferramenta.

3.2.3 Processo de extração e tratamento dos dados

Ao iniciar a execução do *script* Python, o Selenium cria uma instância do navegador Chromium, e assim é realizado o acesso ao site alvo da extração (neste artigo delimitado ao Reclame Aqui) e assim realizada a extração dos dados e posteriormente armazenado no banco de dados. Ao final da execução é encerrada a instância do Chromium.

O processamento de linguagem natural é realizado com auxílio da biblioteca NLTK em Python. Esta biblioteca possui um conjunto de algoritmos que irão auxiliar ao longo dos tratamentos textuais, cujas etapas são ilustradas na Figura 4. A primeira etapa é resgatar os dados do banco de dados, após isto é realizada a remoção dos caracteres especiais. Logo em seguida é iniciada a fase de *tokenize*, que transforma todas as sentenças das frases em *tokens*,

desta forma todas as palavras ficarão separadas e então analisadas individualmente. Esses *tokens* gerados são inseridos em uma lista de *array*, para serem utilizados posteriormente. A partir de agora com os dados da lista que foram criados nas etapas anteriores, é realizada a remoção das *stopwords*, que são artigos, proposições, conectivos e entre outros caracteres gramaticais da língua portuguesa que não representam uma palavra em si, após a remoção é gerada uma nova lista de *array*, contendo os dados sem as *stopwords*. Um exemplo de comentário retirado do Reclame Aqui é ilustrado na Figura 4.

Figura 4 – Etapas do processo de linguagem natural

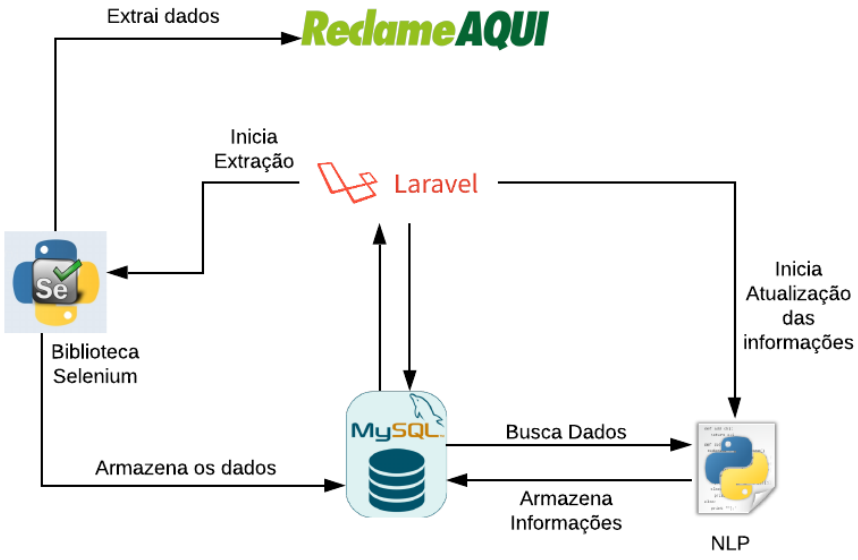
Fonte: Autor (2020)

Na etapa de comparação, são utilizadas palavras que passaram pelos filtros anteriores e

comparadas com as palavras de um dicionário, sendo que através deste dicionário temos as classes gramaticais caracterizadas para cada palavra. Neste projeto de pesquisa analisei apenas os adjetivos, pois são termos que representam qualidade de determinado substantivo. Quando é encontrado um termo igual é contabilizado em um contador, desta forma iremos obter a frequência dos adjetivos mais utilizados, e assim formar um *ranking*.

Por meio do texto extraído deste comentário, é iniciado o processamento de linguagem. O texto então passará pela seguinte sequência: remoção dos caracteres especiais, transformar as sentenças em tokens, remoção das *stopwords* e, por último, é realizada a comparação das palavras extraídas do Reclame Aqui com as palavras de um dicionário. Com a utilização desse léxico é possível identificar quando uma palavra é um adjetivo, pois cada palavra presente no dicionário já possui uma classificação gramatical. Quando uma palavra for igual a do dicionário e seja pertencente à classe gramatical dos adjetivos, ela é contabilizada em um contador, ao final do processo teremos os adjetivos mais utilizados e estes adjetivos serão salvos no banco de dados, que posteriormente irão formar um *ranking* dos adjetivos mais utilizados pelos consumidores. A Figura 5 apresenta um diagrama de funcionamento da aplicação.

Figura 5 – Diagrama de funcionamento da aplicação



Fonte: Autor (2020).

4 DESCRIÇÃO DO SISTEMA DESENVOLVIDO

O projeto de pesquisa e desenvolvimento resultou no *software* Mineração RE. O

usuário, ao acessar a plataforma, terá que realizar o *login* com seu email e senha previamente cadastrados. Em seguida será redirecionado para a tela principal do sistema, que contém as informações gerais das empresas e o *ranking* dos adjetivos mais utilizados, conforme ilustrado na Figura 6, observa-se que o termo "plano" apareceu 316 vezes a cada 775, o contexto de aplicação deste termo pode modificar sua classe gramatical podendo ser um substantivo ou adjetivo, neste trabalho de pesquisa é limitado ao uso apenas de adjetivos, mesmo que tenha sido utilizado como substantivo, pois não é realizada análise de contexto; ainda sim, podemos deduzir que existem muitos usuários insatisfeitos, principalmente em relação ao seus planos contratados com as operadoras de telecomunicação. Ao lado do *ranking* existe o "Perfil das Empresas Relacionadas", que são algumas características do comportamento destas empresas dentro do site Reclame Aqui, com base nestas informações é possível verificar qual empresa possui a melhor relação com os usuários dos quais vendem serviços ou produtos.

Figura 6 – Tela inicial do sistema de Mineração RE

Fonte: Autor (2020).

No menu de navegação presente à esquerda é possível acessar as demais seções da interface, em comparações, há opção de selecionar duas empresas distintas e compará-las, sendo retornado como resultado os dados referentes à diversas categorias de problemas. Em relatórios à esquerda, é possível gerar três formatos diferentes de relatórios. O primeiro irá gerar todas as reclamações dos usuários referente a todas as empresas cadastradas na base de dados. No segundo formato é necessário selecionar a empresa, através do nome da empresa ao qual será gerado o relatório, e por último, é possível gerar um relatório contendo as reclamações dos últimos trinta dias, referentes às empresas pertencentes ao usuário, conforme ilustrado na Figura

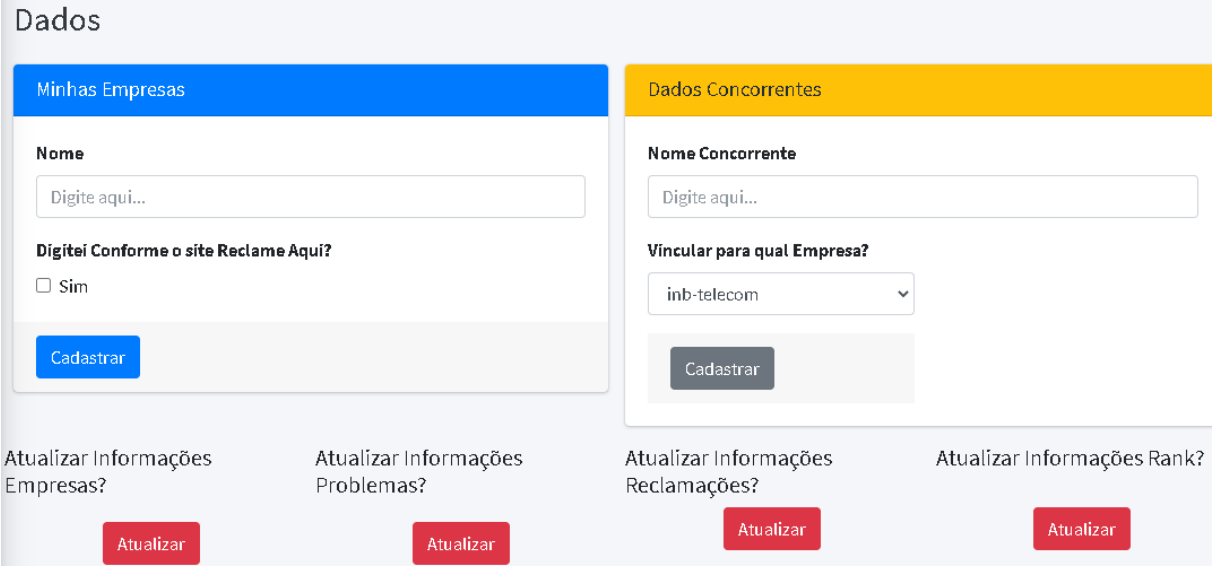
7.

Figura 7 – Relatórios do sistema de Mineração RE

Fonte: Autor (2020).

Em configurações, a última opção à esquerda, é realizada a configuração dos parâmetros da ferramenta. Nesta interface o usuário terá que adicionar as empresas que possui através do nome, que deve ser preenchido igual ao nome presente no site do Reclame Aqui. Em dados dos concorrentes devem ser preenchidas informações referentes às empresas concorrentes. Também deve-se seguir a mesma nomenclatura de empresas do site Reclame Aqui. Logo após é realizada a vinculação da empresa concorrente com a empresa do usuário, conforme é apresentado na Figura 8.

Figura 8 – Configurações do sistema de Mineração RE



Fonte: Autor (2020).

Após ter realizado a configuração dos parâmetros das empresas pertencentes ao usuário e de seus concorrentes, é necessário atualizar a base de dados, para isso existem quatro botões disponíveis na parte inferior da seção de configuração, cada botão é responsável pela atualização de determinada informação referente às empresas.

4 CONCLUSÕES

A mineração de dados pode ser aplicada a diversos conjuntos de dados, existindo muitas técnicas e metodologias disponíveis para se utilizar. Apesar de serem técnicas que existem há bastante tempo, ainda há muitas aplicações possíveis para se implementar, gerando novos conhecimentos com base em padrões e relacionamentos entre os dados. A etapa de *Web Scraping* também é uma técnica com ampla capacidade de obter dados, por meio de acesso às informações públicas na internet é possível extrair e armazenar informações para posterior aplicação de técnicas de mineração de dados.

Durante o desenvolvimento deste projeto foi possível analisar que existem poucos artigos e estudos a respeito de uma ferramenta capaz de encontrar as principais reclamações sobre as empresas de um mesmo segmento. Há alguns projetos científicos que realizavam comparações e buscas pelos principais tópicos citados por usuários, porém não ofereciam ao usuário uma interface da qual poderia se definir os parâmetros e critérios do modelo de mineração de dados, neste trabalho o usuário define a empresa que deseja analisar e acompanhar

os resultados.

Com os resultados obtidos é possível verificar alguns termos que aparecem com mais frequência que outros, também houve um aumento na eficiência do acesso das informações referente às empresas cadastradas. A justificativa de desenvolvimento deste projeto é devido ao aumento constante de canais, portais e sites em que os usuários realizam comentários a respeito dos produtos e serviços oferecidos pelas empresas na internet. Por conta de haver muitas entradas de dados em diferentes canais, o processo de verificação e resposta por parte das empresas é lento e caótico, o que gera frustrações e desgostos aos usuários e consumidores.

Todos os requisitos estipulados neste projeto de pesquisa foram atendidos, sendo que o principal objetivo era criação de uma ferramenta capaz de encontrar as principais reclamações e foi capaz de se observar isto pelos termos encontrados. Essa ferramenta, caso seja implantada em larga escala pelas empresas, acredito que seria uma grande aliada para sanar muitos problemas de produtos ou serviços que as empresas possuem e podem estar passando despercebido de seus controles de qualidade.

Há muitas outras possibilidades de implementações futuras neste projeto, podendo ser modificado seu modelo de mineração para análises de outras classes gramaticais, como substantivos e verbos por exemplo. Existem outras fontes de dados além do Reclame Aqui como as redes sociais que a cada dia ganham novos usuários, sendo muito abundante em dados, há possibilidade também de ser implementada uma análise sentimento em que indicaria o sentido em que determinado termo foi utilizado, podendo ser positivo, negativo ou neutro.

REFERÊNCIAS

AMARAL, Fernando. **Introdução à Ciência de Dados: Mineração de dados e big data**. Alta Books, 1° Edição, 2016.

ARANTES, Viviane D. **E-commerce: A expansão do setor no Brasil e o comportamento do consumidor**. Universidade de São Paulo, 2016

ARBEX, Marcio A.; COSTA, Claudio N.; COUTINHO, Jonatas V.; MAGALHÃES, Lúcia H. **DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS.** FESJ, 2019.

BHATIA, Surbhi.; CHAUDHARY, Poonam.; DEY, Nilanjan.**Opinion Mining in Information Retrieval**. Springer, 2020.

CANALTECH. **O que é servidor Apache?**. Disponível em: https://canaltech.com.br/internet/o-que-e-servidor-apache/. Acesso em: 19 de set. 2020.

DADERMAN, Antonia.; ROSANDER, Sara. **Evaluating Frameworks for Implementing Machine Learning in Signal Processing: A Comparative Study of CRISP-DM, SEMMA and KDD**. STOCKHOLM SVERIGE: EXAMENSARBETE INOM TEKNIK,
GRUNDNIVÅ, 15 HP, 2018.

EBIT. **WebShoppers**. 2019. Disponível em: https://www.ebit.com.br/webshoppers/webshoppersfree. Acesso em: 13 set. 2020.

FIGUEIREDO, Bianca F. **Consumidor.gov.br: a exigência de utilização da plataforma digital de solução adequada de conflitos antes do ajuizamento de ação de consumo como fator de eficiência do Poder Judiciário, à luz da análise econômica do direito**. Revista Eletrônica do CNJ, 2020

HAMMARBERG, Marcus.; SUNDÉN, Joakim. **Kanban in Action**. Manning Publications Co, 2014.

HILLARD, Dane. **Practices of the Python Pro.** MANNING, 2020.

MITCHELL, Ryan. **Web Scraping with Python: Collecting More Data from the Modern Web**. O'Reilly Media; 2° Edição, 2018.

KANTARDZIC, Mehmed. **DATA MINING: Concepts, Models, Methods, and Algorithms**. IEEE PRESS Wiley; 3° Edição, 2020.

ORACLE. **Obtenha Informações Sobre a Tecnologia Java**. 2018. Disponível em: https://www.java.com/pt_BR/about. Acesso em: 26 out. 2019.

PANG, B.; LEE, L. **Opinion mining and sentiment analysis**. Found. Trends Inf Retr. Now Publishers Inc., Hanover, MA, USA, v. 2, n. 1-2, p. 1-135, jan 2008. ISSN 1554-0669.
PRESSMAN, Roger S.; MAXIM, Bruce R. **Engenharia de Software: Uma Abordagem Profissional**. AMGH; 8ª Edição, 2016.

RIES, Eric. **The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses**. Crown Business, 2011.

SAPIRAS, Leonardo Augusto.**Mineração de Opiniões Baseada em Aspectos em Fontes de Opiniões Fracamente Estruturadas**. Porto Alegre: PPGC da UFRGS, 2015.

SOUSA, Gustavo.; GUIMARAES, Isabelle.; JUNIOR, Antonio.; LOBATO, Fabio. **Análise comparativa das principais plataformas de reclamações online: implicações para análise de mídia social em negócios**. Porto Alegre: Sociedade Brasileira de Computação, 2020.