

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
DO RIO GRANDE DO NORTE

THOMAS NOGUEIRA FIGUEREDO

**COLETANDO E ANALISANDO INFORMAÇÕES DO FACEBOOK COM A
LINGUAGEM R**

NATAL
2018

THOMAS NOGUEIRA FIGUEREDO

**COLETANDO E ANALISANDO INFORMAÇÕES DO FACEBOOK COM A
LINGUAGEM R**

Trabalho de Conclusão de Curso apresentado ao Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte, em cumprimento às exigências legais como requisito parcial à obtenção do título de Tecnólogo em Análise e Desenvolvimento de Sistemas.

Orientadora: Dr. Fellipe Araújo Aleixo.

NATAL

2018

Figueredo, Thomas Nogueira.

F475c Coletando e analisando informações do Facebook com a linguagem R /
Thomas Nogueira Figueredo. – Natal, 2018.
52 f : il. color.

Trabalho de Conclusão de Curso (Tecnologia em Análise e
Desenvolvimento de Sistemas) – Instituto Federal de Educação, Ciência e
Tecnologia do Rio Grande do Norte. Natal, 2018.
Orientador: Dr. Fellipe Araújo Aleixo.

1. Mineração de texto. 2. Linguagem R. (Linguagem de programação).
3. Facebook. 4. Análise de sentimento. I. Aleixo, Fellipe Araújo. II.
Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do
Norte. III. Título.

CDU 004.4

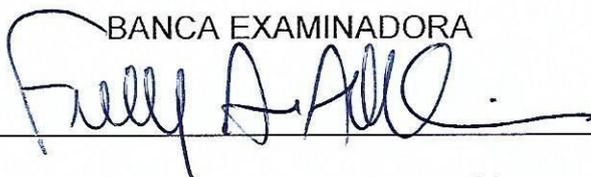
THOMAS NOGUEIRA FIGUEREDO

**COLETANDO E ANALISANDO INFORMAÇÕES DO FACEBOOK COM A
LINGUAGEM R**

Trabalho de Conclusão de Curso apresentado ao Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte, em cumprimento às exigências legais como requisito parcial à obtenção do título de Tecnólogo em Análise e Desenvolvimento de Sistemas.

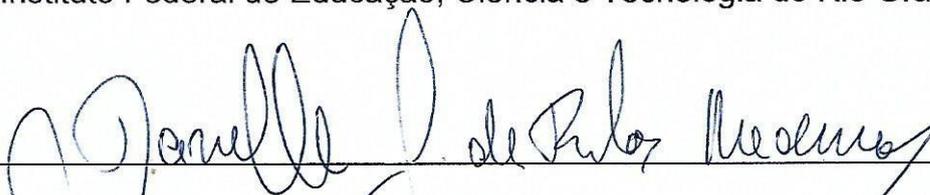
Trabalho de Conclusão de Curso apresentado e aprovado em 13/06/2018, pela seguinte Banca Examinadora:

BANCA EXAMINADORA



Fellipe Araújo Aleixo, Dr. - Presidente

Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte



Danielle Gomes de Freitas Medeiros, M.^a - Examinadora

Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte



Alexandre Gomes de Lima, M.e - Examinador

Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte

AGRADECIMENTOS

Agradeço, em primeiro lugar a Deus, pela força, coragem e sabedoria durante esta longa caminhada e não somente nestes anos como universitário, mas em todos os momentos da minha vida. Por ter me sustentado nas horas de adversidades e me guiado quando não havia esperança para solucionar os problemas enfrentados. Sou grato também pela minha esposa, por ter tido a paciência, como também acreditado em mim por essa grande conquista em minha vida. Não posso esquecer da minha parentela, minha mãe que por muitas vezes tem orado por mim, pedindo a Deus o seu auxílio para me ajudar. Assim como meus irmãos, Deus os recompense.

Conjuntamente, quero agradecer ao apoio dos professores que ao longo desse percurso tive o privilégio de ter sido instruído. Dedicando seu tempo e suas experiências para o nosso aprendizado.

Por fim, o meu orientador Prof. Fellipe Araújo Aleixo, pelo suporte e empenho dedicado no pouco tempo que lhe coube, pelas correções e incentivos para à elaboração deste trabalho. Que Deus te abençoe sempre.

E disse ao homem: Eis que o temor do Senhor é a sabedoria, e apartar-se do mal é a inteligência.

Bíblia Sagrada (Jó 28:28)

RESUMO

Tendo em vista que com surgimento da Internet, as pessoas, empresas e organizações passaram a manter-se relacionados também nesse mundo virtual. As redes sociais surgidas na Internet proporcionam a interatividade dos seus diversos usuários. Os usuários das redes sociais, sejam pessoas ou empresas, podem expressar, através dos comentários, diferentes sentimentos, opiniões a respeito do que foi publicado. Este trabalho se propõe a explorar a análise de sentimentos em comentários postados no Facebook, que foi a rede social escolhida para o aprofundamento desse trabalho. No decorrer do trabalho será apresentada a concepção, implementação e avaliação de uma abordagem para avaliação de tendência em comentários de postagens do Facebook. A abordagem proposta visa a extração de um conjunto de dados não estruturados de uma página do Facebook buscando identificar a tendência, positiva ou negativa do que está sendo comentado. Também faz parte do escopo deste trabalho a aplicação da abordagem em um estudo de caso – Ferramenta DataMy, bem como uma análise dos resultados obtidos na implementação da abordagem e possível definição de melhoramentos na mesma. Através da análise dos resultados obtidos verificou-se que foi possível identificar as polaridades nos comentários processados. Através dos resultados da implementação da abordagem, também foi possível constatar que a técnica de mineração de texto em conjunto com a análise de sentimentos, usada no tratamento das informações permite a identificação de valores nas mensagens publicadas de forma ágil, proporcionando compreender as subjetivas expostas, interpretando-as e classificando-as por sua polaridade.

Palavras-chave: Mineração de texto. Facebook. Análise de sentimentos.

ABSTRACT

Given that with the emergence of the Internet, people, companies and organizations started to stay connected also in this virtual world. The social networks that appear on the Internet provide the interactivity of its various users. Users of social networks, whether individuals or companies, can express, through comments, different feelings, opinions about what was published. This work proposes to explore the analysis of feelings in comments posted on Facebook, which was the social network chosen to deepen this work. In the course of the work will be presented the conception, implementation and evaluation of an approach to trend evaluation in comments of Facebook posts. The proposed approach aims at extracting a set of unstructured data from a Facebook page seeking to identify the trend, positive or negative of what is being discussed. It is also part of the scope of this work to apply the approach in a case study - DataMy Tool, as well as an analysis of the results obtained in the implementation of the approach and possible definition of improvements in it. Through the analysis of the results obtained it was verified that it was possible to identify the polarities in the comments processed. Through the results of the implementation of the approach, it was also possible to verify that the technique of text mining together with the analysis of feelings, used in the treatment of information allows the identification of values in the published messages in an agile way, providing an understanding of the subjective ones exposed, interpreting them and classifying them by their polarity.

Keywords: Text mining. Facebook. Sentiment analysis.

LISTA DE ILUSTRAÇÕES

Figura 1 – Dados divulgados pelo Facebook.....	18
Figura 2 – Etapas da abordagem proposta	23
Figura 3 – Representação do processo de recuperação da informação.....	25
Figura 4 – Ilustração do diagrama de casos de uso do sistema do DataMy	30
Figura 5 – CDU – Solicitar Consulta dos Comentários	31
Figura 6 – CDU – Responder Comentários	32
Figura 7 – CDU – Configurar Resposta Padrão.....	32
Figura 8 – Arquitetura do sistema do DataMy – Diagrama de pacotes	34
Figura 9 – Ilustração simbólica do tratamento dos dados.....	36
Figura 10 – Fragmento do código para coleta dos dados da publicação	37
Figura 11 – Fragmento do código para sumarização	38
Figura 12 – Fragmento do código para a tokenização.....	39
Figura 13 – Fragmento do código para stop-words	39
Figura 14 – Fragmento do código para avaliar os comentários	40
Figura 15 – Fragmento do código para analisar a polaridade nas mensagens.....	41
Figura 16 – Tabela das informações contidas após o tratamento dos dados textuais	42
Figura 17 – Técnicas para mineração de textos	44
Figura 18 – Comparação dos artigos abordados.....	46

LISTA DE ABREVIATURAS E SIGLAS

API	Application Programming Interface
CDU	Caso de Uso
CSS	Cascading Style Sheets
GPL	General Public License
HTML	Hyper Text Markup Language
JSF	Java Server Faces
MVC	Model view controller
PLN	Processamento de Linguagem Natural
SDK	Software Development Kit
SGBD	Sistema de Gerenciamento de Banco de Dados

SUMÁRIO

1	INTRODUÇÃO	12
1.1	OBJETIVOS	14
1.2	METODOLOGIA	15
1.3	ESTRUTURA DO TRABALHO	15
2	REFERENCIAL TEÓRICO	16
2.1	REDES SOCIAIS	16
2.1.1	Facebook	17
2.2	LINGUAGEM R	19
2.3	MINERAÇÃO DE TEXTO	19
2.3.1	Pré-processamento	20
2.3.1.1	Tokenização	20
2.3.1.2	Stop-Words	20
2.4	ANÁLISE DE SENTIMENTOS	21
2.4.1	Polaridades de Expressões Textuais	21
2.4.2	Dicionário Léxico de Polaridades	21
3	ABORDAGEM PROPOSTA	23
3.1	ARQUITETURA - COMPONENTES E RESPONSABILIDADES	23
3.1.1	Etapa 1: Coleta dos Dados	23
3.1.2	Etapa 2: Pré-processamento	23
3.1.3	Etapa 3: Análise de Sentimentos	24
3.1.4	Etapa 4: Expressão da Análise	24
3.2	RECUPERAÇÃO DE INFORMAÇÕES DO FACEBOOK	24
3.3	CLASSIFICAÇÃO E SUMARIZAÇÃO DAS INFORMAÇÕES COMO R	25
3.4	REFINANDO AS CONSULTAS A SEREM REALIZADAS	26
4	ESTUDO DE CASO: FERRAMENTE DATAMY	27
4.1	INTRODUÇÃO AO DATAMY	27
4.2	OBJETIVOS	29
4.3	FUNCIONALIDADES DO SISTEMA	29
4.3.1	Manter cadastro de usuário	30
4.3.2	Manter páginas do Facebook a serem monitoradas	30
4.3.3	Consultar comentários de publicação	31
4.3.4	Responder aos comentários	31
4.3.5	Manter respostas padronizadas	32
4.3.6	Manter token de acesso ao Facebook	33
4.3.7	Coletar dados	33
4.3.8	Tratar dados coletados	33
4.4	ARQUITETURA	33

4.4.1	Camada de Serviço do Facebook	34
4.4.2	Camada do Gerenciador do R	34
4.4.3	Camada de Persistência	35
4.4.4	Camada de Domínio	35
4.4.5	Camada de Visão	35
4.5	FUNCIONALIDADES DA LINGUAGEM R NA APLICAÇÃO	35
4.5.1	Coletando as Informações do Facebook com a Linguagem R	36
4.5.2	Pré-processamento - Sumarizando as informações com a Linguagem R	37
4.5.2.1	Sumarização	38
4.5.2.2	Tokenização	39
4.5.2.3	Stop-Words	39
4.5.3	Análise de sentimentos sobre os comentários com a linguagem R	40
4.5.4	Resultados	41
4.5.5	Tecnologias Utilizadas	42
5	TRABALHOS RELACIONADOS	44
5.1	TRABALHO RELACIONADO 1	44
5.2	TRABALHO RELACIONADO 2	45
5.3	LIMITAÇÕES DOS TRABALHOS RELACIONADOS	46
6	CONCLUSÃO, LIMITAÇÕES E TRABALHOS FUTUROS	47
6.1	CONCLUSÃO	47
6.2	LIMITAÇÕES DO TRABALHO	48
6.3	TRABALHOS FUTUROS	49
	REFERÊNCIAS	50

1 INTRODUÇÃO

Estudos realizados sobre os conceitos de redes sociais segundo Portugal (2007, p.1-3) vêm crescendo ao passar dos anos. Portugal (2007, p.3) descreve que a definição do conceito de rede social apareceu cedo na Sociologia e na Antropologia Social. No entanto, o termo “rede social” começou a ser utilizado entre os anos 30 e 40, no sentido metafórico. Durante a segunda metade do século XX, o conceito de rede social tornou-se, central na teoria sociológica, onde constitui um domínio de conhecimento específico (PORTUGAL, 2007, p.3).

No entanto, a concepção de rede social segundo Garton, Haythornthwaite e Wellman (1997) , diz que “[...] uma rede social é um conjunto de pessoas (ou organizações ou outras entidades sociais) conectadas por um conjunto de relações sociais, como amizade, trabalho em equipe ou troca de informação”.

Com o surgimento da Internet e a inclusão das redes sociais em plataformas *online*, tornou a comunicação de forma mais interativa no vínculo de relacionamento. Coz Fernández et al. (2012, p.61), descrevem através de seus estudos que “[...] as redes sociais virtuais permitem que os seres humanos se comuniquem de uma maneira global e dinâmica dando um caráter tangível a muitos dos aspectos intangíveis das redes sociais do mundo real”. Ainda segundo Coz Fernández et al. (2012, p.60), os seres humanos, por sua natureza são sociáveis, e com o surgimento da Internet, a virtualização permitiu que as redes sociais convencionais fossem mais acessíveis a todo mundo. Convergindo todos os tipos de redes sociais no mundo real em plataformas de redes sociais virtuais, acrescentando um caráter tangível que, muitas vezes, as redes sociais reais não têm. Conduzindo as pessoas a expor seu caráter social a quase todas as situações da vida, ou seja, os usos e costumes do mundo real para o mundo virtual.

Essas informações que são publicadas nas redes sociais virtuais, principalmente na plataforma do Facebook, na forma de: fotos, textos, vídeos ou *links*. Tais publicações geram dados não estruturados em resposta ao que está sendo publicado. Esses dados podem ser em formato textual ou ícones de diferentes tipos expressivos. O compartilhamento dessas publicações, permite aos amigos do usuário que publicou tenham acesso às informações. De modo que o compartilhamento desses informes sejam compartilhados de forma dinâmica para outros usuários de acordo com a relevância ou interesse do conteúdo no que está sendo publicado.

Através dessa forma interativa de publicação e compartilhamento, diversos usuários podem comentar as publicações dos seus amigos, ou aquelas sobre as quais eles reagiram. Através dos comentários um usuário expõe diferentes opiniões sobre todo o conteúdo propagado pelos usuários ligados a ele de alguma forma. Medhat, Hassan e Korashy (2014, p.1094) descrevem que, “os *sites* de redes sociais e *sites*

de *microblog* são considerados uma boa fonte de informação porque as pessoas compartilham e discutem suas opiniões sobre um determinado tópico livremente”. O problema dessas informações que são retornadas consiste no fator da subjetividade no que essencialmente as pessoas estão realmente expressando em seus comentários.

Teixeira e Azevedo (2011, p.55) salientam um ponto importante em seu artigo que, “caso o texto seja escrito de forma informal, poderá ser necessário um pré-processamento da frase, de modo a corrigir erros ortográficos e de pontuação que posteriormente iriam dificultar a procura de informação relevante”. Um ponto a ser considerado, é que, com a popularidade do Facebook, e a capacidade da plataforma de produzir e armazenar enormes volumes de conteúdo, que cada publicação possa gerar, é significativa. Portanto, uma análise abrangente nos dados textuais em relação aos comentários sobre o sentimento dos usuários, podem ajudar a entender o que as pessoas essencialmente pensam e sentem sobre o conteúdo divulgado por meio da publicação.

Eisenstein (2013, p.360-362) expõe outros possíveis problemas com relação à escrita informal nas redes sociais, quais sejam: analfabetismo, limitações de caracteres (no caso do Twitter), entradas de textos (referente aos diferentes tipos de formatação do teclado), pragmáticos (*emoticons* que são componentes expressivos não verbais) e variáveis sociais (podem ser consideradas como diferentes tipos de: idade, gênero, raça e localização). Além disso, Eisenstein (2013, p.360) conforme citado por Joner (2010) apresenta uma pesquisa sobre problemas ortográficos na Internet, que são: (i) as pessoas não têm certeza da grafia correta, (ii) é mais rápido, (iii) tornou-se a norma e (iv) as pessoas querem representar seus próprios dialetos e/ou sotaques.

O exame dos comentários em conteúdos postados, fornece um indício em relação às emoções de quem realizou o comentário. Uma forma de realizar essa análise, consiste na mineração textual, conjuntamente de técnicas relacionadas a análise de sentimentos que integram na extração das informações de textos para a análise da linguagem natural expressa nas redes sociais. Com o objetivo de automatizar a identificação da polaridade de um texto, classificando-os como positivos e negativos. De acordo com Souza, Pereira e Dalip (2017, p.82) conforme citado por Narayanan et al. (2009), “a análise de sentimentos tem por objetivo identificar e extrair de forma automática, as opiniões, sentimentos e emoções, expressados em um texto”.

Através da mineração textual, é possível extrair as informações importantes e não triviais. Dessa forma, a análise de sentimentos trata como uma função de classificação. No qual classifica a orientação de um texto em positivo ou negativo, permitindo o significado do que de fato querem enunciar. De acordo com Medhat, Hassan e Korashy (2014, p.1093) , “[...] o alvo da análise de sentimentos é encontrar opiniões, identificar os sentimentos que expressam e, em seguida, classificar sua polaridade”.

Os benefícios dessa análise estão relacionados à busca de informações específicas, através de processos para identificação de informações úteis no que está implícito no texto. Em redes sociais virtuais, essa investigação pode contribuir para melhorar o relacionamento de empresas com seus consumidores, gerando novos *insights* a partir das sugestões, elogios ou até mesmo críticas contidas nos comentários às postagens dessa empresa. O problema enfrentado pelas empresas, está no grande volume de informações que precisam ser analisadas. Esse problema pode levar a empresa em um mau atendimento, podendo ocasionar a perda de clientes e apresentar dificuldades de crescimento no mercado competitivo. O comportamento dos consumidores no meio *online* permite analisar o que eles pensam através do que eles estão comentando. O planejamento a essa estratégia permite mensurar o impacto em relação dos usuários à imagem da empresa e sua reputação, a fim de tomar ação ou percepção da situação frente aos seus consumidores.

Com o crescimento de palpites, avaliações e outras formas de expressão *online*, tem despertado o interesse em análise nos comentários postados nas redes sociais. O ponto de vista dos usuários que acessam a essas plataformas se transformou em informações de grande relevância para empresas e analistas que desejam estudar comportamentos e auxiliar na comercialização dos seus produtos ou serviços. Além de gerenciar suas reputações.

1.1 OBJETIVOS

Este trabalho tem como objetivo descrever a concepção, implementação e avaliação de uma abordagem para avaliação de tendência em comentários de postagens no Facebook.

Para o objetivo geral apresentado propõe-se os seguintes objetivos específicos:

- a) realizar estudos biográficos de técnicas e ferramentas para realização de mineração de textos utilizando a técnica de análise de sentimentos para interpretar o contexto da linguagem natural utilizada nas redes sociais na compreensão da máquina;
- b) definição de uma abordagem para extração de um conjunto de dados não estruturados de uma página do Facebook buscando identificar a tendência, positiva ou negativa, do que está sendo comentado;
- c) implementar e aplicar a abordagem em um estudo de caso - DataMy;
- d) analisar os resultados obtidos pela abordagem e possível definição de melhoramentos na mesma.

1.2 METODOLOGIA

O desenvolvimento deste trabalho consistiu, primeiramente, na busca de estudos biográficos para uma melhor compreensão de extração e tratamento de dados através das redes sociais. Para analisar os sentimentos e expressões que são postados através de comentários com linguagem natural publicada nas redes sociais.

Após o estudo biográfico, foram realizadas as definições para escolha das técnicas, com o objetivo de realizar a extração dos dados contidos nas publicações e o tratamento do mesmo nas postagens do Facebook. Utilizando o conceito da mineração textual, aplicando a técnica do pré-processamento para o tratamento dos dados textuais e a análise de sentimentos para a descoberta das polaridades nas mensagens contidas.

A próxima etapa foi a implementação da abordagem através de um estudo de caso. Foi desenvolvido o DataMy que trata-se de um sistema *web* que coleta, filtra e analisa os dados extraídos do Facebook para auxiliar os administradores de redes sociais, classificando os comentários das postagens em: positivas, negativas ou neutras, caso a mensagem não seja explícita.

Em seguida, foram analisados os resultados obtidos da implementação da abordagem, onde foram identificadas melhorias na abordagem e possíveis trabalhos futuros.

1.3 ESTRUTURA DO TRABALHO

Este trabalho está organizado da seguinte maneira: No Capítulo 1 é apresentada uma breve introdução ao tema abordado no trabalho. No Capítulo 2 é apresentado o referencial teórico, abordando conceitos sobre as redes sociais e mais especificamente o Facebook, bem como uma introdução à extração de informações através da linguagem R. Como também conceituando as técnicas de mineração textual em conjunto com a análise de sentimentos. No Capítulo 3 é apresentada a abordagem proposta, onde serão descritas e detalhadas as etapas de extração, coleta e análise de dados extraídos do Facebook, além do refinamento desses dados para a compreensão dessas informações. No Capítulo 4 é apresentado um estudo de caso de aplicação da abordagem através do desenvolvimento do projeto do DataMy, que tem o objetivo de auxiliar o administrador de redes sociais no acompanhamento da repercussão do que está sendo divulgado aos seus clientes. No Capítulo 5 são abordados os trabalhos relacionados, com o objetivo de contrapor os resultados alcançados nesse estudo. Por fim, o Capítulo 6 apresenta as considerações finais, bem como as limitações do trabalho, além de sugerir alguns possíveis trabalhos futuros.

2 REFERENCIAL TEÓRICO

Este Capítulo tem o objetivo de embasar conceitos que serão necessárias para adquirir o entendimento para o estudo aplicado.

2.1 REDES SOCIAIS

De acordo com Lozares (1996, p.108), as redes sociais são definidas por um conjunto de atores, que incluem indivíduos, organizações, comunidades ou sociedades globais conectadas por um conjunto de relações sociais. Nas plataformas *Web*, esse vínculo de relacionamento permite aos usuários o compartilhamento de informações e valores de forma interativa e dinâmica. Coz Fernández et al. (2012, p.61) enfatizam que “o caráter social dos seres humanos é inerente a quase todas as situações da vida, a vida cotidiana do mundo virtual nos levou a exportar nossos usos e costumes do mundo real para o mundo virtual”.

Coz Fernández et al. (2012, p.61) descrevem que

Redes sociais virtuais, são serviços prestados através da Internet que permitem aos usuários gerar um perfil, a partir do qual são disponibilizados dados públicos e informações pessoais e que fornecem ferramentas que permitem a interação com outros usuários, e localizá-los de acordo com as características publicadas em seus perfis.

As plataformas das redes sociais virtuais vêm oferecendo estruturas funcionais que permitem a divulgação de forma prática das informações, onde proporciona o compartilhamento desses valores por meio dos usuários dentro do vínculo de relacionamento.

Essa interatividade das relações nas redes sociais, ocorre de diversas formas, como por exemplo: (i) publicar algum conteúdo, (ii) compartilhar a publicação com outros usuários, (iii) apresentar alguma expressão através de *emoticons* ou curta referente alguma publicação, ou (iv) optar pela possibilidade de comentar em forma de texto alguma divulgação.

Consiste em um desafio a tarefa de interpretar o conteúdo e o contexto associado aos comentários, devido à subjetividade ou o tipo de grafia das mensagens. De acordo com Irfan et al. (2015, p.158), “[...] em *sites* de redes sociais, as pessoas geralmente usam linguagem não estruturada ou semiestruturada para comunicação”. Irfan et al. (2015, p.157) ressaltam a estrutura da escrita, citado por Sorensen (2009), onde relata que, “em todos os dias conversa sobre a vida, as pessoas não se importam com a grafia e a construção gramatical de uma sentença que pode levar a diferentes tipos de ambiguidades, como léxica, sintática ou semântica”.

A mineração de texto é uma possibilidade para solução do entendimento da linguagem natural utilizada nas mensagens publicadas nas redes sociais. Segundo Irfan

et al. (2015, p.158) citado por Aggarwal (2011), “técnicas de mineração de texto em conjunto com redes sociais podem ser usadas para encontrar opinião geral sobre qualquer assunto específico, padrões de pensamento humano e identificação de grupo”.

De acordo com Teixeira e Azevedo (2011, p.55),

As sondagens de opinião pública, normalmente feitas com recurso a chamadas telefónicas e entrevistas, podem ser complementadas ou substituídas por uma minuciosa análise das mensagens colocadas pelos utilizadores nas redes sociais, que estão disponíveis sem custos, recorrendo às suas APIs.

2.1.1 Facebook

O Facebook é um serviço de rede social, criado em 2004 por Mark Zuckerberg junto com colegas da Universidade Harvard. A princípio, seu lançamento foi limitado para os estudantes de Harvard, em seguida ampliando para outras universidades, e gradativamente se expandindo para o mundo.

A interação nesta rede social, permitem aos usuários criarem perfis, os quais podem conter fotos, informações pessoais, listas de interesses individuais. O Facebook também proporciona a troca de mensagens públicas e privadas por meio de publicações, onde são postados diferentes conteúdos com a finalidade de atrair a atenção daqueles que as visualizam. Através das publicações é possível o compartilhamento do conteúdo postado para outros usuários, para que, possam ter o acesso à essas informações. O Facebook também disponibiliza outras ferramentas, tais como o *feed* de notícias, com propósito de divulgar diversas informações de publicações que são compartilhadas de diferentes usuários.

Conforme Noyes (2018), os dados informados pelo Facebook, apresentam um crescimento constante de usuários em sua plataforma. Segundo Noyes (2018), o Facebook representa um enorme potencial diante dos outros canais de mídias sociais, mas está se tornando cada vez mais difícil se destacar da multidão, ou seja, satisfazer os seus usuários. Park (2018), citado por Zuckerberg (2018) relata que, “2017 foi um ano forte para o Facebook, mas também foi difícil”.

Figura 1 – Dados divulgados pelo Facebook

DESCRIÇÃO	INFORMAÇÕES	PERCENTAGEM
USUÁRIOS ATIVOS DIARIAMENTE (DAUs)	1,40 BILHÕES DE USUÁRIOS ACESSAM O FACEBOOK, MÉDIA DE DEZEMBRO DE 2017	UM AUMENTO DE 14% REFERENTE AO ANO ANTERIOR
USUÁRIOS ATIVOS MENSAS (MAUs)	2,13 BILHÕES DE USUÁRIOS ACESSAM O FACEBOOK, MÉDIA DE DEZEMBRO DE 2017	UM AUMENTO DE 14% REFERENTE AO ANO ANTERIOR
USUÁRIOS ATIVOS DIARIAMENTE POR DISPOSITIVOS MÓVEIS (Mobile DAU)	1,15 BILHÕES DE USUÁRIOS ACESSAM O FACEBOOK, MÉDIA DE DEZEMBRO DE 2017	UM AUMENTO DE 23% REFERENTE AO ANO ANTERIOR
BOTÕES DE CURTIR E COMPARTILHAR	A PARTIR DO ANO DE 2014, SÃO EXIBIDAS EM QUASE 10 MILHÕES DE SITES DIARIAMENTE	* * *

Fonte: Noyes (2018).

Segundo Park (2018), citado por Zuckerberg (2018), o Facebook está tomando iniciativa para que não seja apenas divertida, mas possibilite o bem-estar dos seus usuários, a fim de que haja um número significativo de pessoas que acessem sua plataforma.

Um fato importante é que a popularidade do Facebook não se dá apenas por usuários que usufruem de sua plataforma. Ou pela grande quantidade de curtidas e compartilhamento de informações, ou pelo seu volume de dados gerados, entre outros. Desenvolvedores também utilizam outros serviços disponibilizados pela plataforma do Facebook, tais serviços podem ser incorporados em sistemas tanto *Web*, quanto *mobile*. Um desses serviços disponíveis é o de autenticação de usuários, nesse caso a autenticação utilizando dados de usuários com contas no Facebook. Com esse recurso é possível autenticar usuários em *sites* ou aplicativos, sem a necessidade do preenchimento de formulários. De modo que o acesso às informações do usuário são solicitadas a *API* do Facebook através das permissões de acordo com o requerimento feito no sistema para autenticação do mesmo. Para que, assim que confirmada a autenticação, o sistema possa validar a entrada do usuário.

As requisições à *API* do Facebook podem ser realizadas com a criação de uma aplicação em sua plataforma, ou através da obtenção ao *token* de acesso diretamente na *Developers Graph API*, para que assim sejam realizadas as permissões para acesso às informações a serem consumidas. Também é disponibilizado um *SDK* com o qual é possível desfrutar do potencial dos recursos oferecido pelo Facebook por meio de *plugins*. Que simplifica as configurações para o acesso a sua *API*. As funcionalidades disponíveis com base nos valores das permissões são descritas através das funções da aplicação do sistema como também nos *tokens* de acesso.

2.2 LINGUAGEM R

Conforme Vance (2009) a linguagem de programação R foi desenvolvida por Robert Gentleman e Ross Ihaka no Departamento de Estatística da Universidade de Auckland, Nova Zelândia, em meados dos anos 90. Devido à necessidade de *software* livre que atuasse na área da estatística, sua modelagem foi criada sobre uma linguagem já existente da época, chamada de linguagem S, outro projeto de *software* estatístico. Desenvolvida em meados da década dos anos 70, S foi desenvolvido na Bell Labs anteriormente filiada a empresa AT&T, agora denominada Lucent Technologies, por John Chambers e colegas. Hoje, a linguagem R possui implementação independente da linguagem de programação S e atualmente é desenvolvida pela própria comunidade, onde Chambers é membro. Segundo Matloff (2009, p.1) a linguagem R funciona nas plataformas POSIX, Windows e MacOS.

O R é uma linguagem de programação e ambiente estatístico de código aberto GPL, para realizar análise de dados estatísticos e consequente geração de gráficos. A linguagem R fornece um conjunto de funções estatísticas. Segundo Matloff (2009, p.1-2) a linguagem R possui algumas características relevantes, tais como: (i) além de permitir operações estatísticas é uma linguagem de programação com conceitos gerais as demais linguagens, (ii) possui estrutura de programação funcional e orientada a objetos, (iii) a linguagem possui código aberto, permitindo o programador obter ajuda pela comunidade. Outros pontos podem ser consideráveis na linguagem R: modelagem linear e não linear, agrupamento de dados, cálculos matemáticos, entre outras funcionalidades.

A linguagem R é constituída por um *software*, que inclui um mecanismo de armazenamento de dados, além de um conjunto de operadores para realização de cálculos. Permite a programação e exibição dos dados, tanto a partir da linha de comando, quanto a partir de um ambiente de desenvolvimento integrado, denominado de *RStudio*.

2.3 MINERAÇÃO DE TEXTO

A mineração de texto é um processo automatizado para a obtenção de informações importantes de um texto. As ferramentas de mineração de texto disponibilizam um conjunto de métodos e técnicas de análise e extração de informações a partir de textos, frases ou palavras. De acordo com Nayak et al. (2016, p.16876), a “mineração de texto é o processo de busca ou extração de informações úteis dos dados textuais. Ele tenta encontrar padrões interessantes de grandes bases de dados”. Com isso, o propósito da mineração de texto baseia-se na recuperação de informações, extração de dados, aprendizado de máquina, estatísticas, entre outros. O objetivo da mineração de texto é de extração de conhecimentos valiosos, a partir de dados não estrutura-

dos. Segundo Irfan et al. (2015, p.158) citado por Sorensen (2009), “a mineração de texto é um processo de descoberta de conhecimento usado para extrair informações importantes e não triviais da linguagem natural”.

2.3.1 Pré-processamento

O pré-processamento é uma etapa importante da mineração de texto, visto que, uma vez realizada a coleta dos dados, os mesmos podem não estar adequados ou em formatos compreensíveis. Segundo Nayak et al. (2016, p.16875), “a limpeza de dados converte os dados brutos preenchendo valores ausentes, suaviza dados ruidosos, identifica erros e remove valores discrepantes e resolve as inconsistências”. Com isso, torna-se necessário a aplicação de técnicas para extração, seleção, limpeza e redução dos volumes de dados. Ou seja, visa eliminar o grande volume de dados que não se adéquam a compreensão da informação.

Conforme Meyer, Hornik e Feinerer (2008, p.23),

Os textos de entrada em seu formato bruto nativo podem ser um problema ao analisá-los com métodos de mineração de texto, uma vez que podem conter muitas palavras irrelevantes (como e/ou) ou podem ser formatados inconvenientemente. Portanto, o pré-processamento, ou seja, a aplicação de métodos para limpeza e estruturação do texto de entrada para análise posterior, é um componente central em estudos práticos de mineração de texto.

2.3.1.1 Tokenização

A tokenização é o primeiro passo do pré-processamento, cujo objetivo é extrair unidades mínimas de um texto. Nesse processamento são identificadas e separadas as expressões em unidades mínimas extraídas a partir de um texto livre. Cada unidade mínima é chamada de *token*, que corresponde a uma palavra do texto analisado. O reconhecimento e identificação dos *tokens* é realizado após a eliminação dos elementos separadores das expressões, tais como: espaços, vírgulas, pontos, entre outros. Conforme Teixeira e Azevedo (2011, p.60), “[. . .] o processo de tokenização realiza a divisão do texto em *tokens*, de modo a facilitar o tratamento posterior de cada *token* de forma independente”.

2.3.1.2 Stop-Words

A identificação de *stop-words* é realizada depois da tokenização. Nesta etapa serão identificados os *tokens* que serão desconsiderados. Por tanto, as *stop-words* são palavras consideradas irrelevantes dentro do conjunto de palavras tokenizadas. Essa etapa se encarrega de eliminar os *tokens* que não possuem nenhuma relevância semântica, sendo assim consideradas úteis apenas para compreensão geral textual. A

remoção das palavras consideradas como *stop-words* incluem elementos, tais como: preposições, artigos, adjetivos, numerais, pronomes, advérbios e pontuações. Conforme Irfan et al. (2015, p.159) citado por Shekar & Shoba (2009), “a fase de remover o *stop-word* melhora a eficácia e a eficiência do processamento de texto porque o número de palavras no documento é reduzido”.

2.4 ANÁLISE DE SENTIMENTOS

A análise de sentimentos se baseia em técnicas capazes de extrair informações subjetivas a partir de linguagem natural utilizada nos textos, tais como: sentimentos e opiniões. As informações contidas nos textos podem ser classificadas em dois principais fatores: os fatos e as opiniões. Os fatos são expressões objetivas que procuram oferecer informações precisas ao leitor, ou seja, transmitem informações concretas sobre algo. Opiniões são geralmente expressões subjetivas que descrevem os sentimentos, avaliações particulares ou pessoais. As opiniões expressam a visão pessoal do autor a respeito de algum assunto. Através dos elementos textuais é possível a interpretação da expressão quanto ao seu caráter, sendo considerada a sua polaridade, que são classificadas em mensagens, positivas, negativas ou neutras. De acordo com Souza, Pereira e Dalip (2017, p.82) referente a análise de sentimentos onde descrevem que, “[...] esse tipo de análise foi desenvolvido para que fosse possível captar informações sobre determinados assuntos e extrair um pensamento positivo, negativo ou neutro, conforme a polarização das palavras contidas nos textos analisados”.

2.4.1 Polaridades de Expressões Textuais

Através da análise de sentimentos mediante a extração da subjetividade encontrada em textos de linguagem natural. Permite a identificação da polaridade por meio dos sentimentos expressado no texto. Algumas técnicas permitem categorizar o emocional de uma frase, classificando-as em positivamente, negativamente ou neutras; caso a expressão não seja explícita. Portanto, a polaridade de uma análise de sentimentos é a categorização de como a expressão pode ser compreendida, em formato de classificação (positivas, negativas ou neutras). Segundo Souza, Pereira e Dalip (2017, p.82) citado por Gonçalves (2015), “[...] a polarização das palavras é a representação do grau de positividade, negatividade ou neutralidade emocional de uma frase”.

2.4.2 Dicionário Léxico de Polaridades

O objetivo da aplicação da análise de sentimentos é poder identificar através das mensagens subjetivas, tais como: opiniões, avaliações e emoções; a sua objetividade.

No entanto, para que seja realizado o reconhecimento das polaridades contidas nas mensagens, é necessária uma base que possua indicadores propícios para esse reconhecimento. Com isso, Wilson, Wiebe e Hoffmann (2005), descreve em seu artigo que “[...] uma abordagem típica da análise de sentimentos é começar com um léxico de palavras positivas e negativas e frases. Nestes léxicos, as entradas são marcadas com sua polaridade prévia a priori”. Essas mensagens são consideradas como “pistas” contida no arquivo *subjclueslen1-HLTEMNLP05.tff* utilizado para o reconhecimento das polaridades. A elaboração desse arquivo foi coletada de várias fontes, sendo a maioria dos dados contidos de acordo com Wilson, Wiebe e Hoffmann (2005) por Riloff e Wiebe (2003), e alguns foram selecionados manualmente.

Com isso, o arquivo de acordo com Wilson, Wiebe e Hoffmann (2005) possui “mais de 8.000 pistas de subjetividade. Pistas de subjetividade são palavras e frases que podem ser usadas para expressar estados, ou seja, eles têm usos subjetivos”, no qual cada linha contida no arquivo, contém uma pista de subjetividade. Permitindo assim, a realização da análise de sentimentos para que a polaridade dos sentimentos possa ser extraída do texto.

3 ABORDAGEM PROPOSTA

Este Capítulo tem o objetivo de apresentar a abordagem proposta. A concepção da proposta de abordagem aqui apresentada se baseou em estudos biográficos sobre as técnicas de extração e tratamento dos dados. Essas técnicas têm o intuito em obter de forma automatizada a compreensão dos comentários postados. Processamento esse visando obter informações relevantes do que está sendo extraídos e processado. O objetivo desta abordagem é extrair informações subjetivas descritas nos textos, com a finalidade de obter o conhecimento da informação através do sentimento expresso. O cerne da abordagem está na extração, a partir de uma dada rede social, de mensagens postadas em linguagem natural para obtenção da sua polaridade.

3.1 ARQUITETURA - COMPONENTES E RESPONSABILIDADES

A abordagem proposta foi inicialmente dividida em quatro etapas, as quais estão ilustradas na Figura 2. A seguir são descritas individualmente todas as etapas da mesma.

Figura 2 – Etapas da abordagem proposta



Fonte: Elaboração própria em 2018.

3.1.1 Etapa 1: Coleta dos Dados

É a etapa inicial da abordagem, e tem como objetivo a extração dos dados de uma rede social. Nesta etapa são necessárias as definições claras dos tipos de dados que devem ser coletados. Esses procedimentos são métodos utilizados para extração de informações brutas necessárias para a preparação e tratamento do volume textual. Ou seja, são coletados insumos em uma fonte de informação para que em outras etapas sejam realizadas a construção do conhecimento a fim de extrair algum tipo de ciência sobre os dados coletados.

3.1.2 Etapa 2: Pré-processamento

É a etapa após a coleta dos dados, e tem como objetivo prover técnicas para a captação, preparação, organização e tratamento da massa textual. Esta etapa engloba a análise inicial dos dados brutos, além de operações necessárias para seleção de dados relevantes ao objetivo final. Nesta etapa, o conteúdo dos dados é decomposto

em termos, os menos significativos são descartados, os que possuem mais repetições são classificadas como sem valor, restando apenas os termos relevantes. Os resultados dessa etapa servem para facilitar a identificação do conhecimento da informação.

3.1.3 Etapa 3: Análise de Sentimentos

Após a etapa de pré-processamento, o resultado serve de entrada para a etapa de análise de sentimentos. Nesta etapa são utilizadas técnicas automatizadas de extração para a descoberta das polaridades expressas nos textos. Essa análise aplica-se a partir da subjetividade de linguagem natural por diferentes aspectos: sentimentos, opiniões ou avaliações. Nesta etapa é realizada a identificação da polaridade do texto analisado, seja positiva, negativa ou neutra. De modo geral, o objetivo da análise de sentimentos é o de compreender o posicionamento de uma determinada pessoa em relação com alguma postagem realizada na rede social. Com o propósito de identificar as polaridades do que está sendo contextualizado.

3.1.4 Etapa 4: Expressão da Análise

Os métodos e técnicas aplicados nas etapas anteriores permitem extrair dados brutos em informações úteis expressas através das mensagens textuais postadas em uma dada rede social. Com isso, é possível identificar através de dados ríspidos, informações compreensíveis e o reconhecimento da polaridade do que está sendo comentado referente a uma dada publicação. Esses tipos de técnicas processadas pelas etapas anteriores possibilitam agilidade na análise sobre o conteúdo expresso na rede social. Esta etapa é a soma dos resultados das etapas anteriores. Com isso, o objetivo desta etapa é proporcionar a objetividade do que está sendo exposto nos comentários, identificando o reconhecimento da polaridade do que está sendo comentado em uma dada publicação.

3.2 RECUPERAÇÃO DE INFORMAÇÕES DO FACEBOOK

As redes sociais, em especial o Facebook, possuem um grande volume de usuários que conseguem gerar um grande volume de dados. Por meio de interações que permitem através de publicações, ocasionar uma reação aos demais dentro do seu vínculo de amizade. As publicações, curtidas, compartilhamentos e comentários que os usuários possam expressar, produzem as suas diferentes expressões sentimentais, tais como: alegria, tristeza, entusiasmo ou reclamações, entre outras. Com relação ao que está sendo exposto.

O grande volume de dados pode ocasionar um excessivo esforço para a recuperação das informações. Para permitir o processamento posterior dos mesmos,

os dados são armazenados em banco de dados. Como trata-se de dados dinâmicos o seu armazenamento é de extrema importância. Conforme ilustrada na Figura 3, ao ser solicitada a requisição dos dados coletados no Facebook, os dados passam pelo tratamento e posteriormente armazenado em um banco de dados. Para que ao ser requerida as informações dos mesmos, o processo de busca obtenha rapidez as informações.

Figura 3 – Representação do processo de recuperação da informação



Fonte: Elaboração própria em 2018.

A finalidade na recuperação da informação constitui-se em processos que abranjam um conjunto de mecanismos para que permitam que as informações relevantes sejam perceptíveis pelo usuário. Na recuperação das informações, o sistema precisa dispor de informações para atender as requisições solicitadas. Com isso, o objetivo do sistema é recuperar informações (contidas nos comentários) no que possa ser relevante para quem o acesse.

3.3 CLASSIFICAÇÃO E SUMARIZAÇÃO DAS INFORMAÇÕES COM O R

O processo de sumarização textual envolve métodos para encontrar os dados importantes no texto. A linguagem R permite agilizar esse processo de classificação através de procedimentos específicos aplicados ao tratamento dos dados. As informações relevantes do texto são conservadas, descartando informações triviais. O objetivo da sumarização é a redução do volume textual conservando a importância do contexto da informação.

Na etapa de pré-processamento, a sumarização dos dados é realizada através da extração e limpeza dos termos, a qual permite que cada palavra relevante seja preservada, conservando a importância da informação principal do conteúdo. Essa fase é dividida em duas, conforme abaixo:

- a) **Tokenização** – permite decompor os termos de uma entrada em palavras individuais. Os delimitadores para a tokenização são espaço em branco e pontuações;
- b) **Stop-words** – permite a limpeza de alguns *tokens* dos termos encontrados, onde são removidos: preposições, artigos, pronomes, numerais, pontuações, advérbios.

3.4 REFINANDO AS CONSULTAS A SEREM REALIZADAS

O refinamento das informações após o tratamento pode ser classificadas segundo a polaridade que as mensagens apresentam. A identificação dos sentimentos é composta por palavras subjetivas que possuem diferentes formas, tais como: opinativas, sentimentais ou avaliativas descritas nas redes sociais. A análise, é realizada através de comparações das palavras com um dicionário, o qual determina a predominância da polaridade nos comentários. A análise individual das palavras possibilita a classificação da mensagem como: positiva, negativa ou neutra.

Em relação às empresas, o refinamento dessas informações permite a percepção geral das opiniões expressas pelos consumidores. Favorecendo o conhecimento do rumo da empresa através das consultas, refletidas nesse tipo de análise, para serem utilizados em tomadas de decisões.

4 ESTUDO DE CASO: FERRAMENTE DATAMY

Visando uma validação inicial da abordagem proposta, a mesma foi aplicada em um estudo de caso, caracterizado pelo desenvolvimento de um sistema *Web* para análise de sentimentos sobre os comentários postados pelos consumidores às publicações nas *fanpages* de uma dada empresa. Foram aplicadas as técnicas descritas na abordagem proposta, para as funcionalidades da aplicação no desenvolvimento do *software*. O propósito desta aplicação é auxiliar o trabalho de administradores/gerentes de redes sociais com relação aos seus consumidores. Bem como, identificar a credibilidade da empresa através da análise dos comentários postados. Serão descritos em mais detalhes os objetivos da aplicação, sua arquitetura, casos de uso e funcionalidades do sistema com a linguagem R.

4.1 INTRODUÇÃO AO DATAMY

As empresas utilizam as redes sociais para interagir com seus clientes, bem como divulgar informações de publicidades que possam despertar interesses em relação aos seus produtos ou serviços. Tais informações são publicadas em páginas empresariais, com o intuito de que suas divulgações sejam acessadas pelos consumidores, propiciando o baixo custo sobre o investimento e o macro de grandes visualizações.

Algumas dessas empresas usam as redes sociais para avaliar a satisfação do cliente em relação aos produtos ou serviços oferecidos, como também para tirar dúvidas específicas sobre os seus produtos ou serviços. Já outras organizações apesar de terem este tipo de contato com o consumidor, não fazem uso desse recurso para avaliar o pós-venda.

Esse tipo de ação permite à empresa avaliar o sentimento de contentamento ou desapontamento resultante da oferta dos seus produtos e serviços em relação às expectativas das pessoas.

Para isso, avaliar os sentimentos dos clientes no momento de pós-venda, no que diz respeito sobre o que os usuários expressam suas opiniões, avaliações ou sentimentos às publicações da empresa nas redes sociais, permite auxiliar na estratégia da tomada de decisão. Com isso, o foco da análise proposta pela ferramenta está em postagens no formato de texto expressas através dos comentários.

Um fator importante a ser considerado é a avaliação da credibilidade da empresa por meio dos comentários postados relativos às publicações realizadas pela empresa. Outro importante fator está na grande quantidade de comentários associados a cada publicação, bem como a grande variedade das informações contidas nos comentários nas publicações. Com isso, empresas utilizam a figura de um administrador/gerente de redes sociais, com a função de ler todos os comentários publicados e

dar um retorno, sempre que necessário. O papel do administrador/gerente de redes sociais, está em ser um intermediador entre a marca da empresa e os seus consumidores. No qual, através dos comentários descritos nas páginas, auxiliam a empresa para um bom atendimento. Um ponto importante a ser considerado é o tempo gasto para que cada mensagem seja correspondida. Tornando assim, o trabalho manual e naturalmente lento, dado que para cada comentário específico necessita haver uma resposta correspondente.

Dado esse contexto, é de fundamental importância o trabalho do administrador/gerente de redes sociais. Melhorias nesse perfil de serviço consegue promover um melhor relacionamento entre empresa e consumidor. Entretanto, o mau atendimento pode trazer grandes prejuízos com relação à vazão dos consumidores junto a marca. Portanto, qualquer ferramenta que possa auxiliar esse serviço, pode representar uma garantia de que os consumidores possam ter um melhor atendimento.

Nesse contexto, a identificação da ausência de uma ferramenta eficaz, rápida e precisa para avaliação e análise desse grande volume de informações, motivou o desenvolvimento do DataMy, um *software Web* para coleta, tratamento, interpretação e visualização desses dados.

O DataMy é um *software Web* voltado para solução de *Big Data*, que emprega técnicas descritas na abordagem proposta em sua infraestrutura. A fim de que possibilite uma análise otimizada do atendimento ao serviço ou produto do cliente, postado através das redes sociais, no caso o Facebook. Os resultados dessas análises são verificados e retornados ao usuário de acordo com as suas solicitações exigidas nos comentários descritos nas publicações. A fim de otimizar e agilizar de forma ágil e competente para um melhor atendimento.

O *software* do DataMy permite que as informações coletadas sejam gerenciadas e retornadas ao sistema com aspectos de *cards*, que categorizam as polaridades existentes no contexto das informações do que o usuário está essencialmente expressando em sua mensagem. Com isso, os *cards* são caracterizados em três formas, tais como: positivos, negativos e neutros. Outra funcionalidade existente do sistema, é proporcionar que os comentários sejam retornados de forma eficiente com mensagens padronizadas ou editáveis, com a finalidade de otimizar e agilizar o trabalho manual pelo administrador/gerente de redes sociais.

A princípio o sistema do DataMy trabalha apenas com a rede social do Facebook, devido ao alto crescimento de usuários e o incentivo ao *marketing* digital que a plataforma fornece para a criação de páginas empresariais.

4.2 OBJETIVOS

O sistema do DataMy tem como objetivo geral a implementação dos conceitos da abordagem proposta. Para realização da análise nos comentários postados na plataforma do Facebook por meio das publicações, com a finalidade de auxiliar e otimizar o trabalho do administrador de redes sociais.

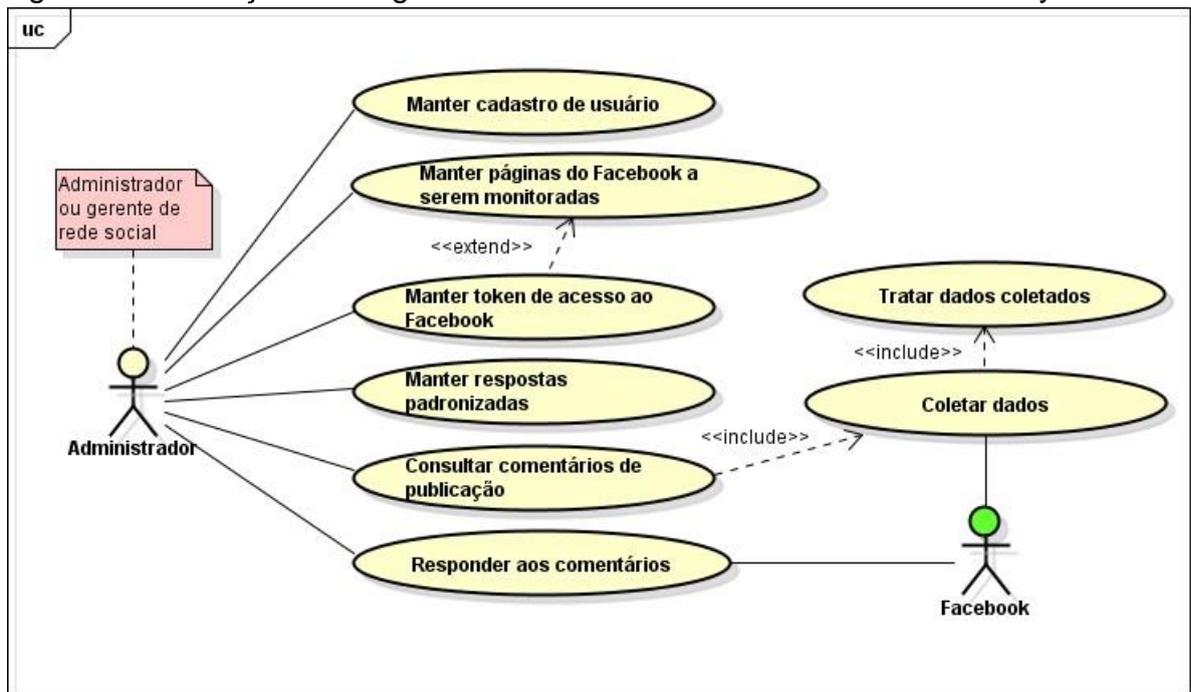
Para o objetivo geral apresentado propõe-se os seguintes objetivos específicos:

- a) coletar as informações de linguagem natural através dos comentários realizados nas publicações veiculadas através da *fanpage* da empresa;
- b) aplicar as etapas da abordagem proposta no sistema do DataMy, para analisar e interpretar o sentimento expresso nos comentários;
- c) expor as tendências dos comentários analisados: positiva, negativa ou neutra e apresentar em formas de *cards* no sistema, para que o administrador/gerente de redes sociais, visualize e examine de forma precisa e otimizada o que está sendo comentado na publicação;
- d) permitir que os comentários analisados e categorizados nos *cards* sejam de fácil acesso, para que o administrador/gerente de redes sociais possa retornar de forma ágil à solicitação do cliente, através de mensagens padronizadas ou editáveis.

4.3 FUNCIONALIDADES DO SISTEMA

O DataMy possui o objetivo de melhorar os processos de trabalho dos administradores/gerentes de redes sociais, bem como disponibilizar uma ferramenta para avaliar a credibilidade da empresa através da *fanpage* empresarial. Os serviços oferecidos pelo sistema se baseiam na análise dos sentimentos dos comentários relativos às publicações realizadas da empresa no Facebook. As funcionalidades do sistema estão ilustrados conforme a Figura 4.

Figura 4 – Ilustração do diagrama de casos de uso do sistema do DataMy



Fonte: Elaboração própria em 2018.

4.3.1 Manter cadastro de usuário

Caso não esteja cadastrado no sistema, o possível usuário poderá inserir informações relativas aos dados solicitados para submetê-las ao sistema, possibilitando a efetivação do cadastro. Esses dados são informações básicas para autenticação no sistema.

A implementação do cadastro foi realizada apenas para autenticação no sistema do DataMy. Porém, os dados do usuário não vinculam com as informações coletadas sobre a rede social do mesmo.

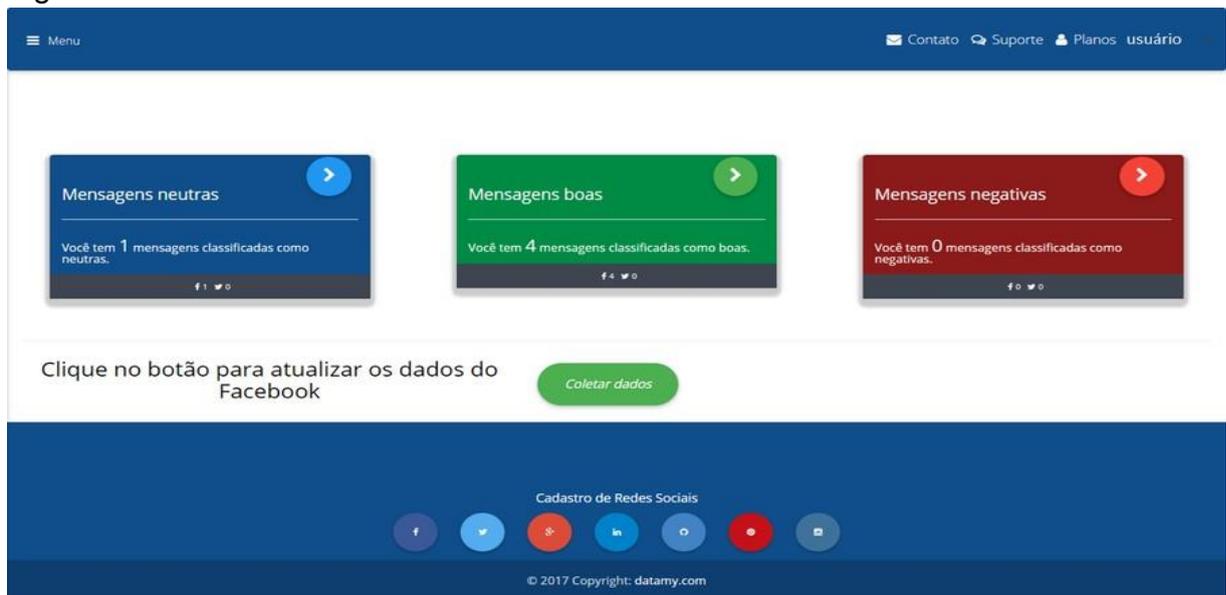
4.3.2 Manter páginas do Facebook a serem monitoradas

O usuário poderá inserir suas informações para se cadastrar junto ao sistema do DataMy usando seus dados das redes sociais, a princípio com o Facebook. Essas informações permitirão a coleta dos dados através da *API* que está disponível na plataforma do Facebook e o armazenamento dos mesmos no banco de dados de nosso sistema para controle de acesso. Para que assim, o sistema possa monitorar os dados em tempo real e realizar as chamadas à *API* de acordo com as requisições. No entanto, não foi finalizado o botão de autenticação que coleta de forma automatizada as informações dos dados na *API* do Facebook.

4.3.3 Consultar comentários de publicação

O usuário do sistema deve estar autenticado no sistema para realizar os respectivos casos de uso e deve ter pelo menos uma publicação na *fanpage* da empresa. Após a autenticação, o sistema exibe um botão para coleta dos comentários relativos a uma dada publicação (esse botão realiza a busca das informações já processadas e contidas no banco de dados e não na realização de uma nova consulta a *API* do Facebook). Os *cards* são classificados em três categorias: mensagens negativas, neutras ou positivas. O administrador/gerente de redes sociais após a coleta, visualiza a quantidade de comentários relativos à publicação. Como também o sistema organiza os comentários de acordo com as expressões expostas nos *cards*, conforme ilustrado na Figura 5. O sistema permite ao administrador/gerente visualizar ou não os comentários publicados e responder de acordo com as exigências.

Figura 5 – CDU – Solicitar Consulta dos Comentários



Fonte: Elaboração própria em 2018.

4.3.4 Responder aos comentários

O administrador/gerente, após solicitar a coleta e análise dos dados, pode visualizar as mensagens como também informações relevantes do mesmo após ser comentado na publicação. Proporcionando ao usuário do sistema selecionar a mensagem para atender à solicitação de acordo com requisitado pelo Facebook, através de um botão responder. Conforme é ilustrada pela Figura 6. Ao selecionar a opção de responder, o usuário do sistema, do tipo administrador/gerente poderá corresponder a requisição do solicitante.

Figura 6 – CDU – Responder Comentários



The screenshot shows the 'Comentários do Facebook' section of the Data My system. It features a table with the following columns: 'Mensagem do post', 'Reações', 'Publicado por', 'Contador de comentários', and 'Mensagem do cliente'. Each row represents a comment and includes a green 'RESPONDER' button.

Mensagem do post	Reações	Publicado por	Contador de comentários	Mensagem do cliente	
Bom dia! Às 18H estaremos funcionando!	1	Pizza da Casa Delivery	6	Oba!	RESPONDER
Bom dia! Às 18H estaremos funcionando!	1	Pizza da Casa Delivery	6	Ótima pizza da região... recomendo!	RESPONDER
Bom dia! Às 18H estaremos funcionando!	1	Pizza da Casa Delivery	6	Bom!	RESPONDER

Fonte: Elaboração própria em 2018.

4.3.5 Manter respostas padronizadas

O sistema disponibiliza configurar mensagens de respostas padronizadas, possibilitando responder individualmente, ou em lote, às mensagens selecionadas. Esta funcionalidade pode ser visualizada de acordo com a ilustração da Figura 7, que permite ao usuário do sistema, garantir o atendimento de forma ágil e dinâmica ao retorno dos comentários.

Figura 7 – CDU – Configurar Resposta Padrão



The screenshot shows a modal window titled 'Responder ao comentario de Maxwell Nicollas'. The comment text is 'Comentario: Ótima pizza e entrega rápida ... recomendo muito!'. Below the comment, there are three radio button options for selecting a response: 'Hehe!', 'Entre em contato pelo telefone 84 991234567', and 'Obrigado por nos avaliar, sua opinião é importante para nós!'. At the bottom of the modal, there are two buttons: 'Publicar' and 'Marcar como visto', and a link 'Ir para a publicação'. A 'RESPONDER' button is also visible in the background.

Fonte: Elaboração própria em 2018.

O usuário ao publicar a mensagem, o sistema registra o atendimento e a mensagem não será mais exibida no sistema. Caso a mensagem não tenha relevância, sem a necessidade de responder ao cliente, o sistema permite que o usuário visualize a mensagem e possibilite a opção de marcar a mensagem como lida. Ao realizar essa opção, o sistema registra o atendimento e a mensagem não será mais exibida. O sistema permite também ao usuário acessar diretamente a publicação, onde é acessada diretamente a página do Facebook correspondente à publicação em questão.

4.3.6 Manter token de acesso ao Facebook

O sistema receberá o *token* de acesso para realizar as requisições a *API* do Facebook. Para a obtenção dos dados referentes a página do usuário pelo processo de coleta das informações, a fim de realizar a análise no sistema do DataMy. A princípio a obtenção do *token* é realizada de forma manual pela *developers graph api*, devido a não finalização do botão de autenticação que permite o acesso automático a esse dado.

4.3.7 Coletar dados

O usuário do sistema solicita a coleta dos dados no Facebook de acordo com as informações contidas na página. A coleta permite o monitoramento que ocorre de forma automatizada e em tempo real.

4.3.8 Tratar dados coletados

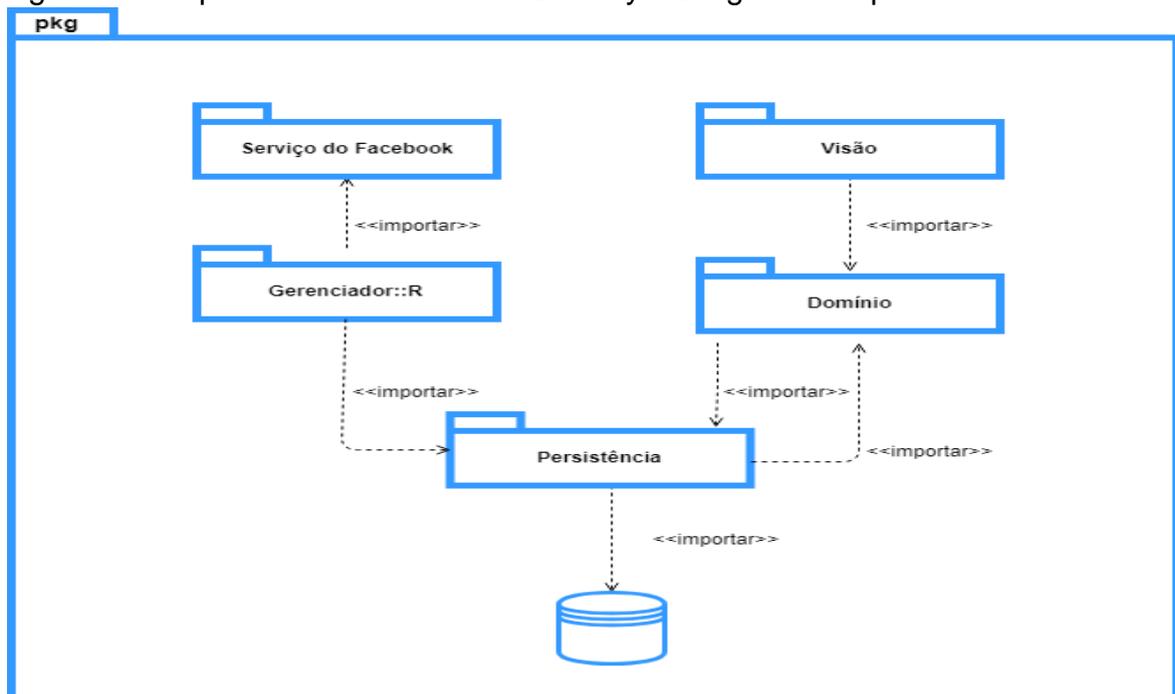
Após a coleta dos dados realizado pelo sistema, será feito um processo de tratamento dos dados, cujo objetivo se trata de separar dados importantes de outros que não possuem valor. Após esse tratamento do sistema, serão utilizados algoritmos capazes de comparar os textos descritos, para uma analogia do que está sendo dito e o que possa ser encaminhado como uma solução de retorno ao usuário.

4.4 ARQUITETURA

A arquitetura do DataMy possui funcionalidades inter-relacionadas entre si pelos seus componentes.

O sistema está estruturado em camadas lógicas de modo a permitir a separação das responsabilidades do *software*, conforme ilustrado na Figura 8.

Figura 8 – Arquitetura do sistema do DataMy – Diagrama de pacotes



Fonte: Elaboração própria em 2018.

A seguir são apresentados a descrição das camadas e suas respectivas responsabilidades.

4.4.1 Camada de Serviço do Facebook

A camada de serviço do Facebook é responsável por permitir que o sistema acesse as informações do usuário através do *feed* de noticiais e na página de perfil do Facebook, por meio do *token* de acesso. Isso se dá através da *Graph API*, a qual possibilita requisições aos dados de um usuário do Facebook. O usuário necessita autorizar o acesso aos dados específicos, possibilitando desta forma o consumo dessas informações pelo sistema.

4.4.2 Camada do Gerenciador do R

A camada do gerenciador do R é camada principal do sistema do DataMy, no qual realiza a coleta dos dados brutos para a posteriormente proceder o tratamento dos mesmos. O processo de refinamento se dá após a coleta dos dados relativos a um serviço/produto através do *feed* de notícias da empresa. Isso é realizado de forma automática através da análise por meio da técnica de mineração textual para interpretação do contexto em linguagem natural.

Esse processo é realizado por meio da técnica de pré-processamento, a qual visa sumarizar os dados brutos, removendo os elementos irrelevantes no contexto para em seguida ser realizada a análise de sentimentos.

Com o resultado do pré-processamento, as frases são analisadas e inseridos rótulos nas palavras de acordo com o que está sendo expressado através do sentimento para reconhecimento da sua polaridade. Esses rótulos são identificadores que pesam a polaridade como: positiva, negativa e neutra. Esse tratamento é realizado para facilitar a compreensão no que está sendo dito nos comentários postados, para que assim, possam ser compreensíveis a objetividade das mensagens descritas nos comentários. O resultado dessa análise é então armazenado no banco de dados.

4.4.3 Camada de Persistência

A camada de persistência, é responsável pelo armazenamento e recuperação das informações no armazenamento persistente – banco de dados. Tais informações são os resultados produzidos pelo gerenciador do R. Faz parte da responsabilidade dessa camada possibilitar o acesso, inserção e atualização das informações.

4.4.4 Camada de Domínio

A camada de domínio é responsável pela implementação de toda a lógica da aplicação. Essa camada, sempre que necessário solicita os serviços da camada de persistência. Esta camada é composta pelas classes que fazem parte do domínio do problema.

4.4.5 Camada de Visão

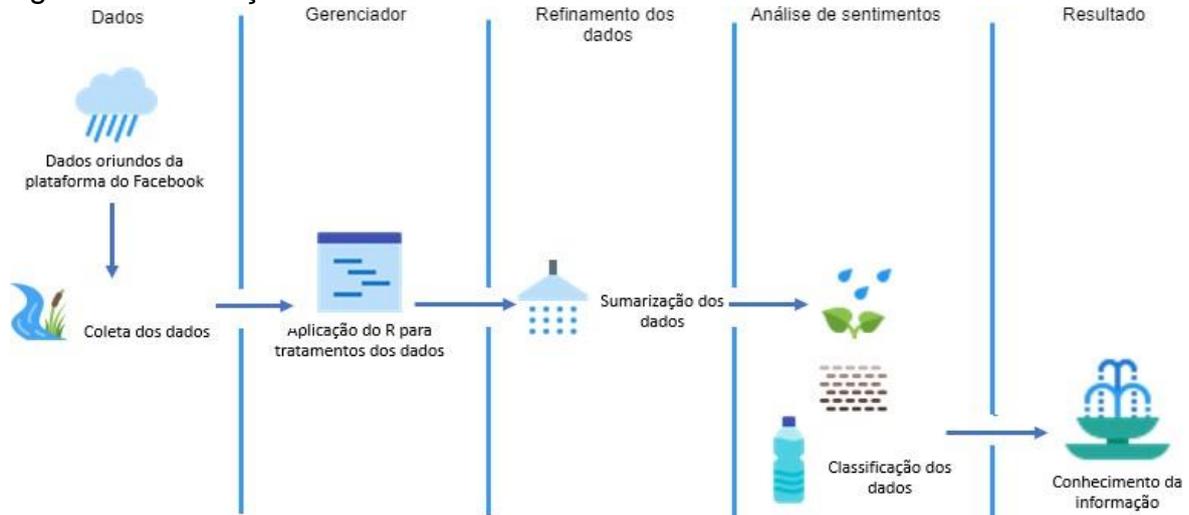
A camada de visão é a camada que dispõe da responsabilidade de apresentação do sistema e interação com o usuário por meio dos componentes visuais vinculado na *interface*. A partir dos componentes, o sistema permite a comunicação das solicitações do usuário com a lógica da camada de domínio de forma dinâmica aplicado ao sistema. Através dessa camada, os usuários possuem acesso às informações contidas pelo tratamento dos dados coletados e refinados pela mineração textual em conjunto com a análise de sentimentos. Como também realizam modificações dos dados já armazenados no banco de dados de acordo com as necessidades das solicitações.

4.5 FUNCIONALIDADES DA LINGUAGEM R NA APLICAÇÃO

Nesta seção serão detalhadas as técnicas utilizadas no tratamento dos dados, bem como os códigos correspondentes na linguagem R. A Figura 9 apresenta uma

visão geral do processo de análise dos dados coletados a partir da plataforma do Facebook.

Figura 9 – Ilustração simbólica do tratamento dos dados



Fonte: Elaboração própria em 2018.

Nas próximas seções, estão detalhadas as etapas do processamento dos dados realizados pelo DataMy. Também estão ilustrados os trechos de código correspondentes escritos com a linguagem R.

4.5.1 Coletando as Informações do Facebook com a Linguagem R

A primeira atividade do processamento consiste do acesso ao Facebook e a coleta dos dados de interesse (comentários associados às publicações de uma dada empresa) e o seu respectivo tratamento através da linguagem R. O acesso aos dados em questão se dá através da *API* do Facebook. Isso se dá através do pacote *RFacebook* da linguagem R, o qual permite a interação de uma aplicação escrita em R com o Facebook. Para que a aplicação em R possa acessar à *API* do Facebook é necessário que seja definido um *token* de acesso. Este *token* possui validade de no máximo duas horas e pode ser renovado através da *developers graph api*. O *RFacebook* possui uma série de funções, as quais permitem a extração de informações a serem utilizadas nas análises disponibilizadas pela linguagem.

A Figura 10 ilustra o código da função *get_fb_dataset*, o qual recebe como parâmetro do *script main* as informações do *token* de acesso bem como a lista (*array*) das páginas a serem consultadas. Incluso no pacote do *RFacebook* contém alguns métodos que coletam as informações do Facebook, tais como: o *getPage*, *getPost*. Que permitem a utilização da extração dos dados contidos nas páginas.

Figura 10 – Fragmento do código para coleta dos dados da publicação

```

1  get_fb_dataset <- function(token, arrayDePaginas){
2    require(Rfacebook)
3
4    token <- token
5    pages <- arrayDePaginas
6
7    posts <- getPage(arrayDePaginas[2], token = token, n = 10, reactions = T, api = "v2.8")
8    comments <- getPost(posts$id[1], token = token, n.comments = 10, api = "v2.8")
9    comments <- as.data.frame(comments, stringsAsFactors = F)
10
11   return(comments)
12
13 }

```

Fonte: Elaboração própria em 2018.

A chamada `require(RFacebook)` permite o carregamento de todas as funções consistentes no pacote para o acesso aos métodos contidos em sua funcionalidade. O objetivo desse pacote é permitir a interação da linguagem R com o Facebook. A fim de executar comandos que extraíam dados das publicações e páginas apresentada nessa rede social. Para a execução das funções é necessária a definição da variável `token` de acesso e da lista de nomes das páginas que serão alvos da coleta. Nas linhas 7 a 8, são realizados os métodos para a coleta dos dados. Esses dados extraídos do Facebook, são informações contidas tanto na página, como nas publicações (a variável `n` dentro de ambos os métodos, delimitam a quantidade de publicações contidas nas páginas quanto às quantidades de comentários que serão extraídos das publicações). O resultado da extração dos dados é passado pelo método `as.data.frame` ilustrado na linha 9, que permite que cada elemento seja separado por coluna no quadro de dados. Isto é, a criação de uma tabela contendo todos os dados coletados, conforme ilustrado na Figura 16. Por fim, os dados agregados e estruturados contidos na variável `comments` servem de retorno da função `get_fb_dataset` para o processo de tratamento dos dados coletados às funções posteriores.

4.5.2 Pré-processamento - Sumarizando as informações com a Linguagem R

Na realização da atividade de pré-processamento, o uso da tokenização permite que o texto seja separado por unidades de palavras representadas pelo termo `token`. Que corresponde a uma palavra contida no texto. A atividade de pré-processamento continua com a sumarização dos dados que já passaram pelo processo de tokenização que permite identificar entre um conjunto de dados, informações relevantes para o conhecimento.

A atividade de pré-processamento no que refere a tokenização envolve a separação das palavras através da remoção dos espaços vazios, separando-as em formas de `strings` para simplificar a manipulação dos dados. Em seguida, as palavras consideradas como conectivos e auxiliares são eliminadas dada a irrelevância das mesmas no

texto extraído. Essa atividade se dá através da aplicação de uma função que remove as palavras de acordo com repetições ou classificações dentro do contexto, tais como: preposição, artigo, numerais, pontuações, entre outros. A sumarização permite extrair apenas as informações não triviais que serão consideradas na análise do contexto do comentário.

4.5.2.1 Sumarização

A sumarização para esse contexto, está no processo que permite eliminar informações triviais contidas no texto. Para que, através desse procedimento possibilite a apresentação das informações relevantes para uma melhor análise e compreensão dos dados mais importantes compostas no texto.

A Figura 11 ilustra o código da função *tokenize_message*, o qual recebe como parâmetro os dados coletados pela função *get_fb_dataset*. O carregamento do pacote *stringr* permite a manipulação das *strings*. A variável *only_text* permite o acesso ao conteúdo que contém apenas as mensagens apresentadas na publicação por meio da extração dos dados processado pela função anterior.

Figura 11 – Fragmento do código para sumarização

```

14 tokenize_message <- function(dataset){
15
16     require(stringr)
17
18     only_text <- dataset$comments.message
19
20     text_without_punctuation <- gsub(pattern = "[[:punct:]]", "", only_text)
21     text_without_controlChar <- gsub(pattern = "[[:cntrl:]]", "", text_without_punctuation)
22     text_low <- stringr::str_to_lower(text_without_controlChar)
23     text_without_digit <- gsub(pattern = "[[:digit:]]", "", text_low)
24     ...
25
26
27
28
29
30
31
32 }

```

Fonte: Elaboração própria em 2018.

O objetivo do trecho do código é o tratamento das *strings* desnecessárias compostas nas mensagens. O método *gsub()* da linguagem R permite a manipulação de uma variável *string* ou vetores de *strings* através da remoção de dados triviais. Com isso, a utilização desse método permite a limpeza de dados irrelevantes do contexto da mensagem, tais como: a remoção de pontuações ou símbolos contidos nos comentários ((i) “ (ii) # (iii) \$ (iv) %(v) & (vi) ’ (vii) () (viii) * (ix) + (x) , (xi) – (xii) . (xiii) / (xiv) : (xv) ; (xvi) < (xvii) = (xviii) > (xix) ? (xx) @ (xxi) [] (xxii) ^ (xxiii) _ (xxiv) ‘ (xxv) {} (xxvi) ~). Como também elementos que não são representados por único símbolo (\n ou \r) como

caracteres de controle. Por fim, todos os numerais que possam conter nas mensagens publicadas são removidos. A utilização do argumento *pattern* indica os dados que serão removidos. Caso contrário, não remove nenhum dado específico dentro do método. Também é utilizado o método *str_to_lower* que converte as *strings* em letras minúsculas para um melhor tratamento das informações contidas. Tais técnicas de manipulação possibilitam o enxugamento das informações contidas nas mensagens.

4.5.2.2 Tokenização

O processo de tokenização é responsável por extrair unidades pequenas de textos. Cada uma dessas unidades é denominada de *token*, normalmente corresponde a uma palavra do texto.

Figura 12 – Fragmento do código para a tokenização

```
23 ...
24 tokenize <- str_split(text_without_digit,"[[:space:]]")
25 ...
```

Fonte: Elaboração própria em 2018.

A Figura 12 ilustra o trecho do código que, permite o tratamento de um conjunto de palavras para separação dos mesmos. O processo de quebra do texto em *tokens* se dá através do método *str_split()*, a qual permite a separação dessas unidades mínimas através da identificação dos espaços em branco contido no texto.

4.5.2.3 Stop-Words

Os *stop-words* são palavras ou sentenças comumente utilizadas que são consideradas insignificantes para um conjunto de resultados a ser apresentado no texto.

Figura 13 – Fragmento do código para stop-words

```
24 ...
25 tokenize <- gsub(pattern = stopwords::stopwords(language = 'pt'),"",tokenize)
26 ...
```

Fonte: Elaboração própria em 2018.

O trecho de código da Figura 13 ilustra a eliminação de palavras que são consideradas irrelevantes no conteúdo do texto. A função *stopwords()* tem o objetivo de comparar as palavras desnecessárias contidas no texto, com uma lista. Essa função permite que os termos comparados e achados na lista sejam removidos, tornando o texto mais enxuto para uma maior precisão no conhecimento da informação do texto. As palavras contidas na lista são: artigos, advérbios, pronomes, numerais, entre outros.

A remoção dessas palavras permite a redução dos dados para facilitar os próximos processamentos a serem realizados com os dados.

4.5.3 Análise de sentimentos sobre os comentários com a linguagem R

A análise de sentimentos é uma técnica de extração de subjetividades em um texto em linguagem natural, tais como opiniões ou sentimentos. A finalidade da análise de sentimentos é obter a polaridade de um texto ou uma frase de forma automática. A análise do conteúdo textual possibilita a classificação das informações de acordo com suas polaridades, sejam elas positivas, negativas ou neutras caso o contexto não seja compreensível.

A atividade de análise de sentimentos recebe o resultado do pré-processamento com o objetivo de categorizar a informação contida nos comentários. A categorização ocorre por meio de um dicionário, onde são comparados os *tokens* contidos nas mensagens, para avaliar a maior incidência da polaridade existente no contexto. Uma vez avaliada a polarização da mensagem, a mesma é classificada como positiva, negativa ou neutra.

É utilizado um dicionário para balizar essa avaliação. O proposto para esse estudo de caso refere-se ao dicionário citado pelo referencial teórico, utilizado também pela Azure em sua infraestrutura. Sua estrutura contém mais de oito mil palavras para o uso na análise de sentimentos.

A Figura 14 ilustra o trecho de código, da função *avaliar_comentarios* que recebe como parâmetro o resultado realizado pelo pré-processamento em conjunto com os dados contidos pelo dicionário léxico.

Figura 14 – Fragmento do código para avaliar os comentários

```

1  avaliar_comentarios <- function(dataset, dicionario){
2
3    for(i in 1:nrow(dataset)){
4      token <- tokenize_message(dataset[i,])
5      classificacao <- sentimental_analisys(dataset_token = token,
6                                           dataset_lexico = dicionario)
7      dataset$avaliacao[i] <- classificacao
8
9    }
10   return(dataset)
11
12 }

```

Fonte: Elaboração própria em 2018.

As funções *tokenize_message* e *sentimental_analisys* são responsáveis pela análise de sentimentos. A função *tokenize_message()* permite listar os *tokens* contidos nas mensagens, para serem analisados pelas palavras compostas no dicionário. A

classificação desses *tokens* é realizado mediante a solicitação da função chamada de *sentimental_analisys*.

A Figura 15 ilustra o trecho de código, da função *sentimental_analisys* que recebe como parâmetro o conjunto de *tokens* das mensagens, como também o conjunto dos dados contidos pelo dicionário léxico.

Figura 15 – Fragmento do código para analisar a polaridade nas mensagens

```

36 sentimental_analisys <- function(dataset_token, dataset_lexico){
37
38   require(dplyr)
39
40   total_match <- semi_join(dataset_lexico, dataset_token, "word")
41
42   positivo <- nrow(filter(total_match, sentimento == "positive"))
43   negativo <- nrow(filter(total_match, sentimento == "negative"))
44   neutro <- nrow(filter(total_match, sentimento == "neutral"))
45   +
46   nrow(filter(total_match, sentimento == "both"))
47
48   maior_incidencia <- which.max(c(positivo, negativo, neutro))
49
50   avaliacao <- classificacao_string(maior_incidencia)
51
52   return(avaliacao)
53 }

```

Fonte: Elaboração própria em 2018.

A função *sentimental_analisys* utiliza o método *semi_join()* conforme ilustrada na linha 40, que realiza a contabilização da quantidade de *tokens* classificados como positivos, negativos ou neutros com base nas palavras existentes no dicionário. O método *which.max()* ilustrada na linha 48 permite identificar o maior valor de incidência com relação à polaridade.

Por fim, a finalidade dessa função é distinguir os *tokens* de acordo com a sua polaridade. Ao final da análise de sentimentos, é avaliada a maior incidência das polaridades na mensagem. Com o objetivo de classificar a mensagem como um todo de acordo por sua polaridade, seja: positiva, negativa ou neutra. Permitindo assim, a identificação do sentimento expresso através do que essencialmente o usuário/consumidor está informando.

4.5.4 Resultados

Essas informações após serem coletadas e analisadas possuem a estruturação de acordo ilustrado na Figura 16. Com a finalidade de, em seguida armazenar os

resultados no banco de dados. Para que ao ser realizada a solicitação ao banco de dados sejam feitas as chamadas a essas informações pela camada de domínio e visão.

Figura 16 – Tabela das informações contidas após o tratamento dos dados textuais

	post.from	post.from	post.mess	post.creat	post.type	post.link	post.id	post.likes	post.comr	post.share	comment:	comment:	comments.mes	comment:	comment:	comment:	comment:	avaliacao
1	512713372	Carrefour	O Carrefoi	2018-03-01	photo	https://w	512713372	216	29	16	#N/D	#N/D	Realizei uma co	2018-03-1	0	2	164849915	positive
2	512713372	Carrefour	O Carrefoi	2018-03-01	photo	https://w	512713372	216	29	16	#N/D	#N/D	Obrigada eu tar	2018-03-01	0	1	164849915	neutral
3	512713372	Carrefour	O Carrefoi	2018-03-01	photo	https://w	512713372	216	29	16	#N/D	#N/D	Obrigada, faço	2018-03-01	1	1	164849915	negative
4	512713372	Carrefour	O Carrefoi	2018-03-01	photo	https://w	512713372	216	29	16	#N/D	#N/D	Nossa o site e u	2018-03-01	0	1	164849915	negative
5	512713372	Carrefour	O Carrefoi	2018-03-01	photo	https://w	512713372	216	29	16	#N/D	#N/D	Não consigo ser	2018-03-01	0	1	164849915	positive
6	512713372	Carrefour	O Carrefoi	2018-03-01	photo	https://w	512713372	216	29	16	#N/D	#N/D	Sou grata por fa	2018-03-01	0	1	164849915	negative
7	512713372	Carrefour	O Carrefoi	2018-03-01	photo	https://w	512713372	216	29	16	#N/D	#N/D	Faço parte dess	2018-03-01	0	1	164849915	negative
8	512713372	Carrefour	O Carrefoi	2018-03-01	photo	https://w	512713372	216	29	16	#N/D	#N/D	Eu Wadson Jonz	2018-03-11	0	1	164849915	positive
9	512713372	Carrefour	O Carrefoi	2018-03-01	photo	https://w	512713372	216	29	16	#N/D	#N/D	Carrefour Brasil	2018-03-01	2	0	164849915	positive
10	512713372	Carrefour	O Carrefoi	2018-03-01	photo	https://w	512713372	216	29	16	#N/D	#N/D	Parabéns a toda	2018-03-01	0	0	164849915	positive

Fonte: Elaboração própria em 2018.

Para esse experimento, foram coletados dez comentários de uma página empresarial de um hipermercado a partir de uma dada publicação, permitindo assim uma análise bem simplificada para a identificação das suas polaridades contidas nas mensagens.

Ao ser realizado todo o processo do tratamento dos dados coletados para a análise. Os resultados foram analisados de forma manual a fim de obter a validação dos dados analisados. No entanto, algumas limitações na análise dos dados permitiram verificar que alguns resultados, no qual cerca de cinquenta por cento não foram satisfatórios quando aplicadas à classificação de suas emoções.

4.5.5 Tecnologias Utilizadas

O desenvolvimento do DataMy adotou uma arquitetura baseada na utilização do padrão MVC voltado para *Web*, e um módulo auxiliar desenvolvido na linguagem R. A implementação do sistema contou com as seguintes tecnologias específicas:

- a) **HTML5 e CSS3** – Usado para construção das páginas *Web*, onde o HTML é usado para estruturar os conteúdos nas páginas e o CSS é utilizada para adicionar estilo nas páginas *Web*;
- b) **Bootstrap** – *Framework* usado para auxiliar na construção de páginas responsivas como também adicionar elementos de *interface*;
- c) **JavaScript** – Linguagem de programação utilizada para deixar conteúdo das páginas *Web* dinâmicas;
- d) **JQuery** – Biblioteca JavaScript que interage em páginas *Web*;
- e) **Java EE com JSF** – *Framework* Java *Web* responsável pelo controle das funcionalidades do sistema;

- f) **Linguagem R** – Linguagem de programação utilizada para coletar as informações, filtrar e analisar os dados coletados através de dicionários e bibliotecas da própria linguagem, promovendo mineração das informações em conjunto com a análise de sentimentos. A fim de identificar informações valiosas dos comentários postados e separação dos mesmos;
- g) **PostgreSQL** – Sistema de gerenciamento de banco de dados utilizado para o armazenamento de informações coletados nos *posts* e dados dos usuários cadastrados no sistema.

5 TRABALHOS RELACIONADOS

Neste capítulo, são apresentados projetos existentes que abordam diferentes conceitos sobre análises textuais, com o objetivo de embasar mais o conhecimento, para uma melhor compreensão. Os trabalhos citados, servem para destacar o que já foi feito e os resultados obtidos sobre o problema em questão. Esses trabalhos possuem utilidade de soma ou são utilizados para fins de comparação. É importante ressaltar que a área de análises textuais através das redes sociais é de grande amplitude, e estudos relacionados são de grande vália.

5.1 TRABALHO RELACIONADO 1

Irfan et al. (2015, p.157) apontam através da pesquisa realizada, uma análise de diferentes técnicas de mineração de texto, a fim de descobrir as variações dos modelos textuais a partir de *sites* de redes sociais, pelo qual os usuários transmitem diferentes valores sociais através da comunicação. Segundo Irfan et al. (2015, p.157) citado por Sorensen (2009), “os *sites* fornecem um meio muito poderoso para comunicação entre indivíduos que leva a aprendizagem mútua e compartilhamento de conhecimentos valiosos”. Irfan et al. (2015, p.158) enfatiza que, nas redes sociais, a linguagem utilizada pelas pessoas não são estruturadas. Conforme Irfan et al. (2015, p.157) citado por Sorensen (2009), “Em todos os dias conversa sobre a vida, as pessoas não se importam com as grafias e a construção gramatical, precisa de uma sentença que pode levar a diferentes tipos de ambiguidades, como lexical, sintática e semântica”.

Conforme os problemas relacionados, a pesquisa realizada de Irfan et al. (2015, p.168) fornece a compreensão de diferentes técnicas de mineração de textos e suas aplicações nas redes sociais.

Figura 17 – Técnicas para mineração de textos

DIFERENTES TÉCNICAS PARA MINERAÇÃO DE TEXTOS		
PRÉ-PROCESSAMENTO NA MINERAÇÃO DE TEXTO	SELEÇÃO DE RECURSOS (SR)	
	EXTRAÇÃO DE RECURSOS (ER)	
CLASSIFICAÇÃO DE TEXTO BASEADA NA APRENDIZAGEM DE MÁQUINAS	ALGORITMO ROCCHIO	
	ALGORITMO DE APRENDIZAGEM BASEADO EM INSTÂNCIAS	
	ÁRVORE DE DECISÃO E MÁQUINA DE VETOR DE SUPORTE	
	REDES ARTIFICIAIS NEURAI	
	ALGORITMOS GENÉTICOS	
CLASSIFICAÇÃO DE TEXTO BASEADA NA ONTOLOGIA ABORDAGEM HÍBRIDA		
MINERAÇÃO DE TEXTO USANDO O CLUSTERING	AGRUPAMENTO HIERÁRQUICO	
	CLUSTERING PARCIAL	K-MEAN, K-MEDOID, C-MEAN, AND C-MEDOID
		ALGORITMO DE SINGLE-PASS
		ALGORITMO PROBABILISTICO
	BASE SEMÂNTICO CLUSTERING	

Fonte: Elaboração própria em 2018.

A investigação dos avanços no campo de análise de texto em redes sociais ainda é um desafio, que depende totalmente de conjuntos de dados a ser analisados e da natureza do problema a ser investigado. Outro ponto importante que é citado, está nas técnicas de mineração que podem ser utilizadas para dados textuais gramaticalmente incorretos.

Ao analisar o respectivo trabalho, compreende-se que o foco do mesmo está baseado na análise das técnicas de mineração de textos nas redes sociais. Bem como realizar análise no estilo gramatical, sintático, semântico, lexical em que são escritos nas páginas de redes sociais, ou seja, a linguagem natural com intuito de obter o sentimento das mensagens postadas. Dados esses, que fundamentasse no trabalho realizado nesse estudo. Outro ponto apresentado no artigo, está nos tipos de mineração textuais e seus procedimentos criados em diferentes algoritmos. Entretanto, a escolha de algoritmos eficiente e eficaz, capaz de minerar fornecendo resultados precisos nos dados não estruturados, torna-se desafiadora quando as informações textuais não são estruturas de acordo com a convenção gramatical. De acordo com Irfan et al. (2015, p.157) diz que, “portanto, a extração de padrões lógicos com informações precisas dessa forma não estruturada é uma tarefa crítica a desempenhar”.

5.2 TRABALHO RELACIONADO 2

No artigo descrito por Teixeira e Azevedo (2011), os autores apresentam uma técnica de análise sentimental aplicada em redes sociais, em conjunto com a abordagem de técnicas de pré-processamento textual e processamento de linguagem natural, conforme descrito no artigo anterior. O pré-processamento permite um tratamento refinado das informações não estruturadas de linguagem natural, para que aprimorando as informações, seja feita o tratamento da análise de sentimentos. Citando duas redes sociais para experimento, tais como: Facebook e o Twitter, devido à abrangência e popularidade de usuários. Tendo em vista as correlações significativas nas percentagens de mensagens positivas e negativas.

As coletas de dados abordadas no artigo, são informações disponíveis nas redes sociais. Nos quais essas informações representam comentários publicados nas redes sociais, pelo fato de ser opiniões dos usuários que comentam sobre os filmes de cinemas lançados no mercado. Após a coleta desses dados, os autores introduzem o conceito de análise sentimental para o tratamento das informações, como também é apresentado o problema referente a esse tratamento, sendo detalhado a abordagem, expondo uma breve explicação da solução desenvolvida, na qual apresentam diferentes níveis para extração e verificação dos dados coletados. Por fim, apresentam os resultados obtidos na análise.

Ao analisar o artigo, temos que o tema abordado através de um estudo de caso,

no qual a análise de dados, se dá através dos comentários divulgados nas redes sociais. Um ponto importante do artigo, está no detalhamento que o estudo e elaboração do projeto foi abordado, até chegar nos resultados obtidos. Essas informações contribuem para o contexto do trabalho, bem como no estudo aplicado ao desenvolvimento. Em contraponto, por se tratar de um grande volume de dados, o tratamento específico não assegura que todos os textos extraídos nas redes sociais, possam ter sido analisados pela técnica abordada.

5.3 LIMITAÇÕES DOS TRABALHOS RELACIONADOS

No artigo, conforme Irfan et al. (2015), a implementação de técnicas de mineração de texto para pesquisadores nas redes sociais torna-se um desafio. Segundo Irfan et al. (2015, p.167) diz que, “nas redes sociais os dados textuais podem ser grandes, turbulentos e dinâmicos”. Conforme Irfan et al. (2015, p.167), a interpretação de expressões devido aos *emoticons* por expressar conceito e/ou emoções específicas ainda é um desafio.

No segundo artigo de Teixeira e Azevedo (2011), a investigação da pesquisa possui limitações, no qual, o processo de análise implementada no projeto, tais como: a coleta de dados (comentários e *emoticons*) e no seu tratamento, não são infalíveis. Visto que, há a possibilidade da existência de frases e termos que não foram encontradas na base de dados, ou seja, não podem terem sido avaliados através da análise de sentimentos a sua polaridade. O tratamento de grande volume de dados, torna o desafio com maior complexidade para a análise de dados, para que possam ser obtidos os resultados desejáveis.

Na Figura 18 está uma ilustração referente aos artigos abordados.

Figura 18 – Comparação dos artigos abordados

RELAÇÃO DE COMPARAÇÃO ENTRE ASSUNTOS ABORDADOS NOS ARTIGOS						
	TRABALHO RELACIONADO 1 A SURVEY ON TEXT MINING IN SOCIAL NETWORKS		TRABALHO RELACIONADO 2 ANÁLISE DE OPINIÕES EXPRESSAS NAS REDES SOCIAIS		COLETANDO E ANALISANDO INFORMAÇÕES DO FACEBOOK COM A LINGUAGEM R	
	SIM	NÃO	SIM	NÃO	SIM	NÃO
TRATA DA EXTRAÇÃO DE DADOS	X		X		X	
TRATA DO PROCESSAMENTOS DOS RESULTADOS	X		X		X	
DEFINE UMA ABORDAGEM		X	X		X	
VALIDA A ABORDAGEM ATRAVÉS DE UM ESTUDO DE CASO		X	X		X	
PERMITE CONFIGURAR PARA QUE PERIODICAMENTE POSSAM FAZER AS CONSULTAS		X	ATRAVÉS DE NOVAS CHAMADAS AS APIs		ATRAVÉS DE NOVAS CHAMADAS AS APIs	
CONCEITUA E/OU REALIZA O TRATAMENTO COM PROCESSAMENTO DE LINGUAGEM NATURAL	X		X			X
ABORDA MAIS DE UM TIPO DE TRATAMENTO TEXTUAL NAS REDES SOCIAIS	X			X		X

Fonte: Elaboração própria em 2018.

6 CONCLUSÃO, LIMITAÇÕES E TRABALHOS FUTUROS

Este Capítulo apresenta uma síntese das conclusões obtidas a partir da finalização deste trabalho. Oportunamente são discutidas as limitações identificadas por este estudo. Por fim, são apresentadas sugestões para trabalhos futuros.

6.1 CONCLUSÃO

Este trabalho, encarou o desafio de examinar, de modo estatístico, as particularidades da classificação automática de comentários relativos às publicações em redes sociais. Para tanto, foram aplicados conceitos e técnicas de análise de sentimentos em conjunto com técnicas de mineração textual. A concepção da proposta de abordagem apresentada, baseou-se em estudos biográficos sobre as técnicas para extração e tratamento dos dados. Para obter de modo automatizado a compreensão dos comentários sobre as informações subjetivas descritas nos textos, com o intuito de extrair o sentimento expresso.

Através da proposta abordada, foram aplicadas as técnicas para as funcionalidades da aplicação no desenvolvimento do DataMy. O propósito desse sistema é auxiliar o trabalho de administradores/gerentes de redes sociais com relação a atender as solicitações descritas pelos seus consumidores. Bem como, identificar a credibilidade da empresa através da análise dos comentários postados. Entretanto, a aplicação não identifica erros ortográficos, textos ambíguos e a identificação de *emoticons* contidos nos textos. Proporcionando assim, alguns resultados inconsistentes realizados pela análise. Visando atingir melhorias para essa abordagem, estudos aprofundados por meio de técnicas para esse tipo de tratamento, torna-se necessário para analisar de forma precisa o resultado dos sentimentos expressos. A reparação a esses problemas permitem a identificação das emoções e a sua clareza. Possibilitando identificar de forma minuciosa as tendências das polaridades nas expressões.

Foi utilizado um dicionário léxico como base para interpretar e classificar as palavras de acordo com suas polaridades. Classificando os comentários processados como: positivos, negativos ou neutros.

Como o sistema do DataMy empenha-se na exibição de dados a partir de uma análise preditiva de informações coletadas por meio dos comentários postados através de uma dada rede social, a princípio o Facebook. O uso de um *dashboard* analítico poderá apresentar as informações de forma fácil e intuitivo para compreender com agilidade melhor as oportunidades que a empresa tem através da análise, como também os riscos encontrados e possibilitar elaborar estratégias para o melhoramento dos resultados. Todas essas informações poderão ser facilmente analisadas possibilitando uma melhor tomada de decisões para a empresa. Com isso, o uso de um *dashboard* são fundamentais para dirigir as mudanças necessárias e possibilitar melhorias em

uma empresa através das informações estatísticas.

O aumento dos estudos na análise de sentimentos ressalta a sua importância para as organizações. Com o propósito de auxiliar as organizações a tomada de decisão. Como também, apoiar as organizações para tarefas de melhorias no atendimento ao cliente, a fim de otimizar e agilizar o encargo do administrador/gerente de redes sociais.

6.2 LIMITAÇÕES DO TRABALHO

A análise de sentimentos sobre os comentários relacionados às determinadas publicações, são realizadas em função da polaridade. Os comentários analisados podem ser classificados como positivos, negativos ou neutros quando os usuários não são explícitos no texto dos mesmos. No entanto, essa análise fica limitada a uma compreensão mais aprofundada dos sentimentos expressos pelos usuários. Através das técnicas abordadas o trabalho inspeciona o texto fornecido para a identificação emocional da opinião dominante em um texto.

Ao realizar a análise através do estudo de caso, observou que frases que possuíam uma certa extensibilidade considerável, poderiam dificultar o reconhecimento da polaridade. Uma vez que, durante a escrita de texto extenso, dificulta a compreensão do texto deixando o entendimento imperceptível, tornando a mensagem ambígua.

Outro problema que fugiu ao escopo definido para este trabalho, consiste na identificação de erros ortográficos na grafia dos comentários. O algoritmo proposto para realização da análise não processa a correção e o devido tratamento. Ao analisar palavras e não verificar a sua existência no dicionário léxico, as mesmas não são avaliadas. Proporcionando assim, em algumas mensagens a omissão da compreensão da informação, sobre o que o autor está querendo expressar.

Outra limitação identificada para este trabalho está no fato do não processamento da linguagem não-verbal contida no texto como, por exemplo, a análise dos *emoticons*. O usuário pode combinar o uso de *emoticons* para expressar diferentes expressões humanas, tais como: raiva, alegria, dentre outras. A sua inclusão permite uma melhor identificação da informação inclusa no texto.

Um fator que pode ser destacado como outra limitação para este trabalho, está em uma abordagem para a identificação dos termos presentes, como: o sarcasmo e a ironia, que possuem suas peculiaridades linguísticas empregadas nos textos escritos em uma dada publicação. Onde o autor da mensagem pretende oferecer um discurso em um sentido conotativo do que realmente quer expressar. A identificação desses termos permite diferenciar e detectar o reconhecimento real da polaridade ao que realmente está sendo expresso no discurso enunciado através dos comentários escritos nas publicações.

6.3 TRABALHOS FUTUROS

O trabalho realizou uma análise explorativa das técnicas de análise de sentimentos e de mineração textual nos comentários publicados na plataforma do Facebook.

Como trabalhos futuros, pretende-se: (i) estudar conceitos sobre técnicas de mineração de texto para aplicação de corretor ortográfico ou identificadores de abreviação de palavras, com objetivo de auxiliar na identificação de palavras, evitando que os termos desconhecidos sejam identificadas; (ii) estudar conceitos de análise sobre *emoticons*, visando o melhoramento da aplicação para aprimorar a precisão da análise. Para que através da análise dos *emoticons* nas mensagens publicadas, o algoritmo possa identificar os sentimentos transmitidos por meio dessa expressão. Permitindo assim, que o resultado seja mais coerente no sentimento contido na frase; (iii) estudar a técnica de processamento de linguagem natural (PLN), para fornecer o tratamento mais aprofundado dos dados coletados. Com objetivo de melhorar o reconhecimento do contexto, através da análise morfológica, sintática e semântica. Com a finalidade de extrair informações relevantes e interpretar os sentidos do conteúdo, bem como analisar os sentimentos contidos nos textos processados; (iv) comparar os resultados obtidos pela aplicação da abordagem com uma análise manual, a fim de contabilizar o tempo gasto de retorno nas mensagens e a análise realizada sobre o mesmo, como também, verificar se os resultados atingidos pela aplicação estão coerentes sobre o que realmente os usuários estão expressando nas mensagens; (v) explorar algoritmos mais eficientes que contextualize a análise de sentimentos, com o intuito de verificar a objetividade da precisão de ambas as técnicas para o aperfeiçoamento da aplicação; (vi) integrar as técnicas utilizadas diretamente na plataforma do Facebook para outras redes sociais. Permitindo assim, a agregação do vínculo da empresa a outras plataformas, não limitando o uso da aplicação.

REFERÊNCIAS

- COZ FERNÁNDEZ, J. R. et al. Evaluación de la Privacidad de una Red Social Virtual. *Revista Ibérica de Sistemas e Tecnologias de Informação*, Porto, n. 9, p. 59 – 73, jun. 2012. Disponível em: <http://www.scielo.mec.pt/scielo.php?script=sci_arttext&pid=S1646-98952012000100006&lng=pt&nrm=iso>. Acesso em: 23 mar. 2018.
- EISENSTEIN, J. What to do about bad language on the internet. 2013. Disponível em: <<http://www.aclweb.org/anthology/N13-1037>>. Acesso em: 06 abr. 2018.
- GARTON, L.; HAYTHORNTHWAITE, C.; WELLMAN, B. Studying Online Social Networks. *Journal of Computer-Mediated Communication*, [S.l.], v. 3, p. 75 – 106, jun. 1997. Disponível em: <<https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1083-6101.1997.tb00062.x>>. Acesso em: 29 mar. 2018.
- IRFAN, R. et al. A survey on text mining in social networks. *The Knowledge Engineering Review*, [S.l.], v. 30, n. 2, p. 157 – 170, mar. 2015. Disponível em: <<http://dx.doi.org/10.1017/S0269888914000277>>. Acesso em: 13 mar. 2018.
- LOZARES, C. La teoria de redes sociales. *Universitat Autònoma de Barcelona*, Barcelona, v. 48, p. 103 – 126, 1996. Disponível em: <<http://papers.uab.cat/article/view/v48-lozares/pdf-es>>. Acesso em: 23 mar. 2018.
- MATLOFF, N. The Art of R Programming. 2009. Disponível em: <<http://heather.cs.ucdavis.edu/~matloff/132/NSPpart.pdf>>. Acesso em: 31 maio 2018.
- MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, [S.l.], v. 5, p. 1093 – 1113, dez. 2014. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S2090447914000550>>. Acesso em: 06 abr. 2018.
- MEYER, D.; HORNIK, K.; FEINERER, I. Text Mining Infrastructure in R. *Journal of Statistical Software*, [S.l.], v. 25, n. 5, p. 1 – 54, 2008. Disponível em: <<http://epub.wu.ac.at/3978/1/textmining.pdf>>. Acesso em: 18 mar. 2018.
- NAYAK, A. S. et al. Survey on Pre-Processing Techniques for Text Mining. *International Journal Of Engineering And Computer Science*, [S.l.], v. 5, p. 16875 – 16879, jun. 2016. Disponível em: <<http://ijecs.in/index.php/ijecs/article/download/2019/1868>>. Acesso em: 21 abr. 2018.
- NOYES, D. The Top 20 Valuable Facebook Statistics. 2018. Disponível em: <<https://zephoria.com/top-15-valuable-facebook-statistics/>>. Acesso em: 24 mar. 2018.
- PARK, M. Facebook Reports Fourth Quarter and Full Year 2017 Results. 2018. Disponível em: <<https://www.prnewswire.com/news-releases/facebook-reports-fourth-quarter-and-full-year-2017-results-300591468.html>>. Acesso em: 24 mar. 2018.
- PORTUGAL, S. Contributos para uma discussão do conceito de rede na teoria sociológica. *Oficina do CES*, [S.l.], n. 271, p. 1 – 36, mar. 2007. Disponível em: <<http://www.ces.uc.pt/publicacoes/oficina/271/271.pdf>>. Acesso em: 26 mar. 2018.

SOUZA, K. F. de; PEREIRA, M. H. R.; DALIP, D. H. UniLex: Método Léxico para Análise de Sentimentos Textuais sobre Conteúdo de Tweets em Português Brasileiro. *Abakós - Instituto de Ciências Exatas e Informática*, Minas Gerais, p. 79 – 96, maio 2017.

Disponível em: <<http://periodicos.pucminas.br/index.php/abakos/article/view/P.2316-9451.2017v5n2p79/11702>>. Acesso em: 02 abr. 2018.

TEIXEIRA, D.; AZEVEDO, I. Análise de opiniões expressas nas redes sociais. *Revista Ibérica de Sistemas e Tecnologias de Informação*, Porto, n. 8, p. 53 – 65, dez. 2011.

Disponível em: <<http://www.scielo.mec.pt/pdf/rist/n8/n8a06.pdf>>. Acesso em: 15 mar. 2018.

VANCE, A. Data Analysts Captivated by R's Power. 2009. Disponível em: <https://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?pagewanted=1&_r=3&emc=eta1#>. Acesso em: 31 maio 2018.

WILSON, T.; WIEBE, J.; HOFFMANN, P. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. 2005. Disponível em: <<https://people.cs.pitt.edu/~wiebe/pubs/papers/emnlp05polarity.pdf>>. Acesso em: 26 jun. 2018.