Um *Data Warehouse* Textual em Língua Portuguesa: Estudo de caso do sentimento dos usuários do Twitter durante a eleição de 2018

Instituto Federal de Educação, Ciência e Tecnologia Catarinense – Campus Camboriú
Caixa Postal 2016 – 88.340-055 – Camboriú – SC – Brasil

Jonathan Vinícius Suter, Rodrigo Ramos Nogueira,Tatiana Tozzi, Daniel Fernando Anderle, Rafael de Moura Speroni

{jonathan.vinicius.suter, wrkrodrigo, tatitozitt}@gmail.com, {daniel.anderle, rafael.speroni}@ifc.edu.br

Resumo. *As redes sociais cada dia mais causam impacto no cotidiano das pessoas e organizações, neste contexto, o Twitter, no qual um usuário escreve uma expressão com até 280 caracteres e outras pessoas podem ver ou compartilhar novamente essa mesma expressão ou a sua própria. Este artigo apresenta um trabalho que tem como objetivo consumir o grande repositório de dados que é o Twitter, e, a partir dele criar um Data Warehouse no qual é possível analisar os textos, as expressões contidas. Nesta proposta, são incluídos métodos de pré-processamento dos textos. Também para enriquecer essa base com a análise de sentimento, além do projeto do banco de dados a proposta inclui um método classificador para os textos, utilizando aprendizado de máquina, que é capaz de predizer um sentimento relacionado a um Tweet, seja ele positivo, neutro ou negativo.*

Abstract. *Social networks are increasingly impacting everyday people and organizations, in this context Twitter, in which a user writes an expression with up to 280 characters and other people can see or share that phrase again or their own. This paper presents a work that aims to consume the great data repository that is Twitter, and from it to create a Data Warehouse in which it is possible to analyze the texts, the expressions contained. In this proposal, methods of preprocessing texts are included. Also to enrich this base with the analysis of feeling, in addition to the project of the database the proposal includes a classifier method for the texts, using machine learning, that is able to predict a feeling related to a Tweet, be it positiva, neutral or negative.*

1. Introdução

Desde o início da Web, o volume de dados que estão nos repositórios na rede mundial tem crescido de forma exponencial, atualmente são cerca de 200 milhões de sites ativos na Internet1, dos quais, apenas a rede social Twitter gera, em média, 500 milhões de postagens por dia. Tal explosão de dados, levou a um estudo do IDC (Institute Data Corporation) que estima que até 2020 serão gerados 44 zettabytes de dados em todo mundo. No entanto, com o crescente aumento de dados de maneira escalável fazem com que os métodos tradicionais de exploração desses dados têm se tornado inadequados,

segundo (CAMILO; SILVA, 2009). Desta forma, torna-se necessário não apenas a revisão dos métodos atuais, mas principalmente a criação de novos métodos de exploração, mais rápidos e precisos.

Com o crescimento da utilização das redes sociais, os conteúdos compartilhados demonstram características associadas ao perfil de cada usuário, principalmente seus interesses e opiniões sobre os mais diversos assuntos. No contexto da expressão de pensamentos e compartilhamento desses sentimentos, as redes sociais são repositórios gigantes de dados a respeito dos mais variados assuntos, objetos, pessoas, comportamentos. No caso do Twitter, o poder de se definir um raciocínio em poucos caracteres a torna singular neste aspecto, sendo um facilitador da dispersão de idéias.

Devido ao volume de dados, torna-se humanamente impossível analisar as expressões de ideias, sendo necessário o desenvolvimento de um mecanismo que seja capaz de captar esses dados, analisar e indicar quais os sentimentos, emoções estão sendo transmitidos pelas pessoas através dos Tweets.

Para JUNQUEIRA (2018), entre diversas aplicações em um conjunto linguístico baseado em textos do Twitter, se destacam as pesquisas que exploram a análise de sentimento. O processo de análise de sentimentos consiste na abordagem computacional que, com a utilização de técnicas de processamento de linguagem natural e aprendizagem de máquina, tem o objetivo de julgar textos a fim de determinar sentimentos e opiniões presentes em frases. Análise de sentimentos também é comumente conhecida por vários outros termos, tais como: extração de opinião, mineração sentimento, análise de subjetividade, análise afetiva, análise de emoções e mineração de opinião.

Em redes sociais, a análise de sentimentos é utilizada para verificar a polaridade de opiniões e pensamentos dos usuários, ou seja, se as opiniões e pensamentos são positivos ou negativos. Assim, a análise de sentimentos se tornou campo de interesse de vários setores, funcionando como ferramenta de *feedback* sobre o que as pessoas pensam, segundo (CAVALCANTE, 2017).

A análise multidimensional da rede social pela perspectiva do sentimento pode ser útil em diversos contextos desde marcas avaliando seu produto,até mesmo como no caso do objeto de estudo, avaliar o cenário político. Deste modo, este artigo apresenta as etapas da criação de um Data Warehouse alimentado com dados da rede social Twitter e efetuar o enriquecimento semântico partir do sentimento dos dados extraídos.

2. Trabalhos Relacionados

A análise de sentimentos é uma sub-área da inteligência artificial em ascensão tendo diversas aplicações, principalmente no marketing de produtos e político. Sabendo da ampla utilização do Twitter para armazenar dados e expressar sentimentos, o mesmo tem sido amplamente empregado como fonte de caso de estudo para diversos trabalhos nesta área. JUNQUEIRA (2018), realizou a coleta de 988.512 textos do Twitter, os rótulos foram inseridos manualmente, posteriormente foram avaliados os métodos de aprendizado de máquina, onde o melhor método foi o SVM com uma acurácia de 95,7%.

Um estudo de ARAÚJO, Mateus et. al (2016) sobre o funcionamento dos métodos de análise de sentimento no contexto das redes sociais. Utilizando duas bases com dados de redes sociais, foi feita comparação do funcionamento entre oito métodos de classificação de sentimentos e quais os resultados da análise. Utilizando a mineração de dados no Twitter, CORREA (2017) fez a extração e análise dos sentimentos dos filmes indicados ao Oscar de 2017 utilizando o algoritmo de classificação naive Bayes, baseado no teorema de Thomas Bayes, classificando os Tweets em relação ao seu conteúdo como positivo, negativo e/ou neutro. Efetuando a extração de Tweets relacionados a três categorias de notícias, NASCIMENTO, Paula et al. (2012) efetuaram a análise de sentimento, se o sentimento em relação a elas era positivo ou negativo. Utilizando classificadores, com aprendizado supervisionado, mediu-se qual era mais eficaz para este tipo de tarefa, para textos em português, especificamente.

LOCHTER et. al (2014) identificaram a necessidade de qualificadores de textos mais eficazes para auxiliar na medição da polaridade dos mesmos, dada a grande quantidade de abreviações, gírias e símbolos utilizados nas redes sociais. Para isto, utilizou-se dicionários semânticos e ontologias para auxiliar na elaboração de um comitê de classificadores que detectam automaticamente os métodos de classificação de linguagem natural mais eficazes nessa tarefa. Por sua vez, MORAES et. al. (2015) coletaram Tweets em português que foram postados durante a partida entre Brasil e Alemanha na Copa do Mundo FIFA de 2014 para identificar a polaridade. A classificação dos Tweets foi feita a mão e após, foram mostrados os resultados, a quantidade de Tweets positivos, negativos e/ou neutros.

AGUIAR et al. (2018), mediram a capacidade de um comitê de algoritmos de aprendizado de máquina para análise de sentimento em redes sociais com a língua portuguesa, usando como estudo de caso a rede social Twitter. Concluiu-se que em alguns casos, os outros algoritmos e o comitê obtiveram desempenho equivalente na mesma tarefa. TAVARES et al. (2017) apresentam uma solução de Business Intelligence para facilitar a extração de informações da rede social Twitter por organizações, efetuando a etapa de ETL, extraindo informações através do reconhecimento de entidades nomeadas, a descoberta de conhecimento em texto , inserindo os dados em uma nova base e permitindo a análise gráfica dos dados.

Para minerar dados da rede social Twitter, TREVISAM (2015) desenvolveu uma ferramenta para recuperação inteligente de dados para que pudesse efetuar a sumarização e posterior análise dos dados para que se possa extrair informações a partir desta base dados, a respeito de algum evento no mundo real.

3. Metodologia

Para PIZZANI et al. (2012), uma pesquisa bibliográfica tem vários fins, para aprimorar-se o conhecimento a respeito do assunto abordado e as tecnologias relacionadas, também para auxiliar na definição do escopo do que será desenvolvido. Por isso a primeira etapa desta pesquisa foi dedicada ao levantamento bibliográfico para se obter a fundamentação teórica sobre o que é Data Warehouse, o que é a análise de sentimento, os métodos de classificação por aprendizado de máquina, bem como os trabalhos já desenvolvidos na mesma linha de estudo (estado da arte).

Esta pesquisa também se enquadra como pesquisa tecnológica de acordo com JUNIOR et al. (2014), pois o produto final é conjunto de arquitetura, software, complementado de um conjunto de dados. Para o desenvolvimento desta etapa foi realizado a extração dos dados, mediante ao emprego de um Web Crawler, que busca os Tweets através da API disponibilizada pelo próprio Twitter. Então, os Tweets são classificados de acordo com seu conteúdo, se eles têm conteúdo positivo, neutro ou negativo relativo ao assunto.

A condução do desenvolvimento foi realizado tendo como base a arquitetura criada com base na arquitetura de um Data Warehouse de KIMBAL (2011). A Figura 1 mostra a arquitetura proposta por esta aplicação de Data Warehouse. Inicialmente efetuada a coleta dos textos assim como o pré processamento, compondo a etapa de ETL. Finalmente, após os dados pré-processados e limpos podem ser realizadas consultas OLAP para explorar o cubo de dados. As etapas da arquitetura são descritas em maior nível de detalhamento na sequência.

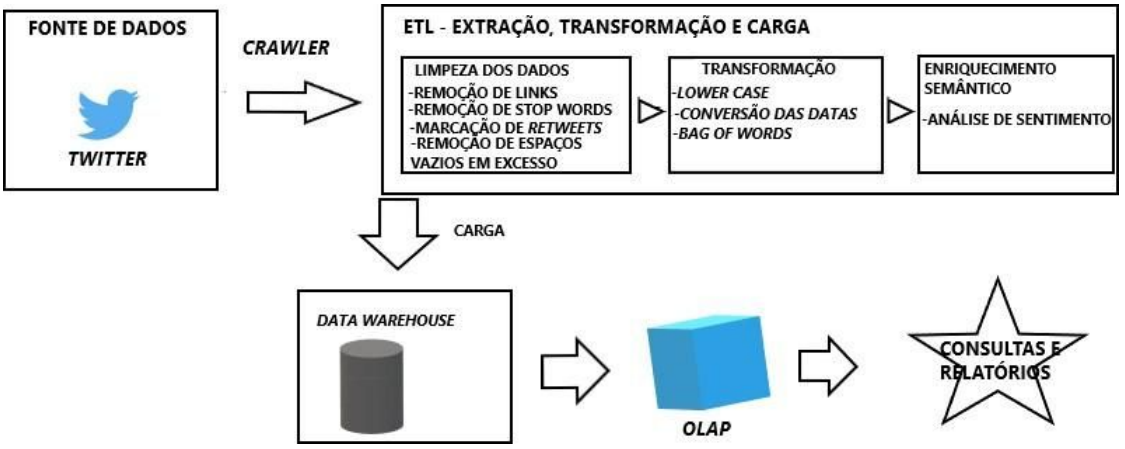

Figura 1. Arquitetura utilizada para coleta e Data Warehousing

Acoplado à etapa de extração da ETL, a coleta é feita por um *Web Crawler*, desenvolvido utilizando a linguagem de programação *Python*, na versão 3.6. As requisições ocorrem através do uso da biblioteca "TwitterSearch 1.0.210". Uma vez optando-se por coletar textos em língua portuguesa, sobre as eleições de 2018.

A etapa da limpeza de dados é essencial para o armazenamento de textos, pois é nela que são removidos os dados desnecessários, que além de ocupar espaço em disco, podem atrapalhar o desempenho dos métodos computacionais que utilizam os dados armazenados. Na etapa de limpeza desenvolvida durante a arquitetura Data Warehouse deste projeto foram considerados os seguintes fatores que foram removidos dos textos coletados:

a) Existência de imagens, bitmaps, gifs e etc. no meio dos textos, sendo necessária a retirada dos mesmos para que possam ser inseridos na base.

b) Retweets: devido às limitações que a API do Twitter impõe, não há como desconsiderá-los entre as requisições, diminuindo a variação dos textos e criando a necessidade de tratar os textos com essa marcação.

c) Links no meio dos Tweets. Exemplo:

d) Sequência de caracteres que são reconhecidos como de “escape” pelo compilador ou que prejudicam a construção do SQL para inserção do texto na base (sequências do tipo “\n” e aspas simples no meio da cadeia de texto)

e) Espaços vazios em “excesso”. Exemplo:

f) Remoção de Stop-words.

Com os textos já limpos, seleciona-se a data do registro e é efetuada sua formatação para que possa ser inserida na base. A partir disso, os dados do Tweet estão preparados para que o mesmo possa “quebrado” e se efetue a Bag of Words. Com os dados do Tweet, as palavras são quebradas pelo script e inseridas na base de dados multidimensional. Caso a palavra já exista na base, é apenas atualizada sua frequência.

E assim, tem se um documento com os termos e sua frequência em cada Tweet e com uma consulta, sua frequência na base como um todo.

O banco de dados multidimensional armazena os textos dos *Tweets* que foram padronizados e limpos. O modelo multidimensional descrito na Figura 2 representa os dados armazenados neste artigo. No qual a tabela fato são os textos curtos (tweets) que são analisados por suas dimensões (sentimento, tempo, palavra).

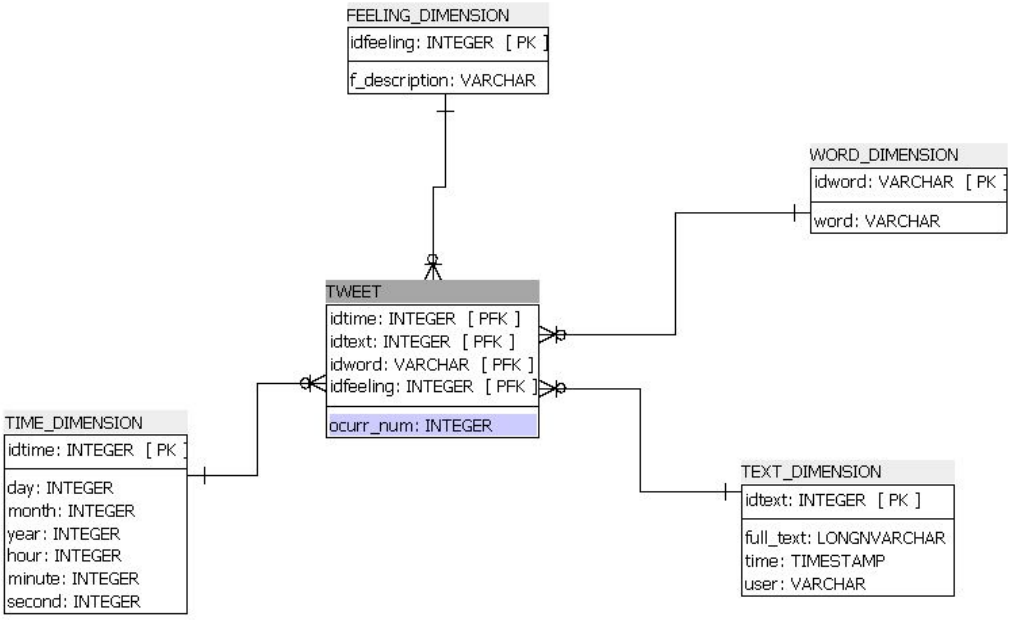

Figura 2. Modelo Multidimensional de textos e sentimentos

Foram coletados 108893 Tweets entre os meses de julho e outubro, referentes à hashtag “eleicoes2018”. Após as etapas de coleta, preparação dos textos e enriquecimento semântico e, ao efetuar o treinamento do algoritmo de classificação, usando o conjunto de dados para treinamento com 1300 tweets classificados manualmente.

4. Resultados e Discussões

O ambiente desenvolvido, que visa essa integração por meio da proposta de uma arquitetura de Data Warehouse, bem como os materiais e métodos utilizados em seu desenvolvimento.

O objetivo do ambiente desenvolvido é fornecer um conjunto de dados consistentes e limpos, na forma de um conjunto de dados em um modelo multidimensional de texto do Twitter rotulados com sentimentos, de tal maneira, que possam ser consumidos aplicações externas e usuários. Sendo assim, o ambiente foi desenvolvido baseado em uma arquitetura que visa proporcionar:

- um modelo multidimensional que armazene o conjunto de textos, sentimentos e a característica temporal dos Tweets com dados em em tempo real;
- anotações semânticas de maneira dinâmica no ambiente;
- a exploração de qualquer cubo de dados de consultas multidimensionais requisitadas por aplicações e usuários.

A arquitetura proposta por esta aplicação foi desenvolvida com em um processo de ETL denominado *ETQ (Extract, Transform, Query*), que realiza consultas dinâmicas em variadas fontes de dados (principalmente dados oriundos da Web, etc.), gerando um painel visual para atender às demandas dos usuários e principalmente uma *API (Application Programing Interface)* para responder às demandas online de aplicações. Assim, os resultados do processamento *OLAP* podem integrar dados de todas as fontes relevantes às consultas realizadas.

Então, após o teste, foram qualificados os demais tweets da base e assim, explorando as dimensões do *Data Warehouse,* pode-se obter os resultados de palavras e ocorrências mostrados pelo Quadro 1.

Palavra	Identificador	Quantidade
eleições2018	93	51458
bolsonaro	80	24424
candidato	3	10559
haddad	79	9726
diz	97	8184
presidente	230	7188
contra	93	7160
eleições	341	7125
sobre	39	6443

Quadro 1. As dez palavras com maior número de ocorrências

Pode-se observar que naturalmente, o termo usado para a pesquisa dos Tweets é o que tem mais ocorrências, este pode ser desconsiderado no momento. Entretanto, a segunda palavra mais citada entre os textos é “bolsonaro”. O segundo termo mais citado é “candidato” e o terceiro é “haddad”, indicando primariamente que estes foram os candidatos mais citados.

Tendo como objetivo fazer uma análise mais objetiva sobre as eleições, foram selecionados os nomes dos candidatos e consultados os mesmos. O Quadro 2 ilustra o resultado de menções por candidato.

Código identificador	Candidato	Quantidade de citações
80	Bolsonaro	24424
79	Haddad	9726
333	Ciro	4510
545	Alckmin	1897
2524	Amoêdo	192
4640	Daciolo	1819
889	Meirelles	588
1052	Marina	1817
7067	Alvaro	139
2487	Boulos	1098
2979	Vera	61
2534	Eymael	13
13512	Goulart	78
Total	-	46362

Quadro 2. Quantidade de menções diretas por candidato

Como é possível ver no Quadro 2, entre o total de Tweets coletados, houveram 46362 com citações a candidatos à presidência. O candidato mais citado entre os Tweets foi Jair Bolsonaro, com 24424 citações em Tweets, cerca 52,68% do total de citações; Em contrapartida, o candidato com menos citações na base é o José Maria Eymael, com apenas 13: cerca de 0,028% do total de citações. Esta consulta pode ser utilizada como uma base para uma análise de repercussão de cada candidato. Demonstra numericamente quais candidatos estiveram mais à vista dos eleitores.

A análise multidimensional também permite a associação entre dimensões de um Data Warehouse. O Quadro 3 mostra a nome dos candidatos e as menções feitas à eles em relação aos sentimentos.

Candidato	Ruim	Neutro	Bom
Bolsonaro	7946	14796	1279
Haddad	3255	5681	465
Ciro	1826	1936	632
Alckmin	1163	618	81
Amoêdo	117	46	27
Daciolo	401	868	314
Meirelles	227	312	40
Marina	763	897	149
Alvaro	41	72	26
Boulos	590	251	249
Vera	14	46	1
Eymael	6	6	1
Goulart	23	55	0
Total	16372	25584	3264

Quadro 3. Quantidade de menções por sentimento

A primeira análise a ser feita é que, os candidatos que estavam à frente do pleito primeiro receberam um grande volume de tweets negativos e positivos. Sendo que do

mesmo modo, é possível observar que nenhum candidato obteve mais citações boas que ruins. Este dado reflete a polarização e o ódio muitas vezes noticiado durante a campanha[1], sendo que o sentimento geral entre os tweets foi ruim, e que nenhum candidato conseguiu obter uma grande aprovação dos eleitores, comprovado pelo grande número de abstenções na eleição de 2018. (refletindo de certa forma, muito bem como se encerrou a disputa eleitoral).

O melhor resultado, não era o esperado no início da pesquisa, partiu justamente da ocorrência de termos e sua exploração multidimensional. Um vez que o quando ordenamos os candidatos pelo seu número de citações na rede social, o ranking é muito próximo do resultado das eleições em primeiro turno O Gráfico 1 elucida tais resultados, no qual quando comparados com dados do TSE (2018) a única diferença é que os candidatos Guilherme Boulos e Marina Silva obtiveram mais citações do que votos.

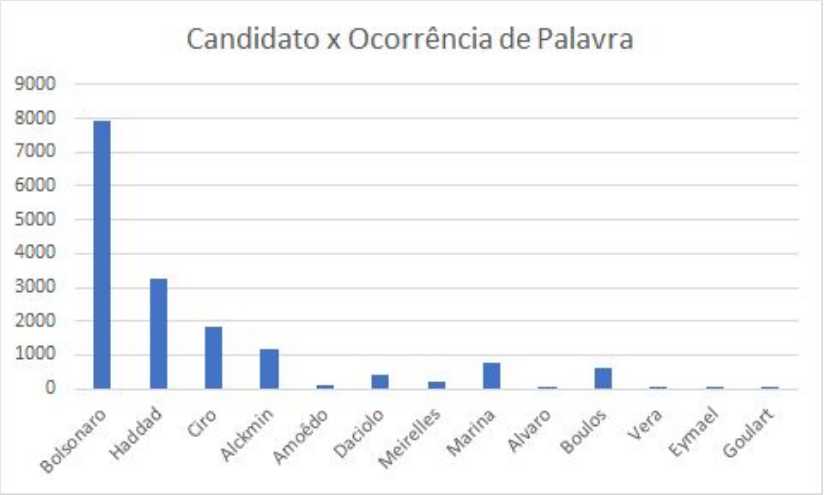

Gráfico 1. Menções por candidato no primeiro turno

6. Considerações finais

O tratamento e análise de textos escritos por pessoas, que possuem pouca ou nenhuma revisão, ainda mais em um espaço de informalidade como o Twitter, podem trazer desafios, tanto com os dados em si quanto com o sentido que eles possuem. Assim, a inserção de uma etapa para classificação dos textos como parte da ETL se tornou essencial para automatizar essa tarefa, que pode ser bastante morosa para um humano. Desta forma, o pré-processamento dos textos para que os mesmos possam entrar na base de dados já limpos e qualificados permite ao usuário se preocupar apenas com o processo analítico dos dados, e desta forma, extrair informações e relatórios, como proposto.

Apesar das limitações que a API do Twitter impõe, ainda é possível criar aplicações interessantes, usando os métodos corretos para a estrutura e análise dos

[1] "Eleições 2018 levam ódio e desavença às relações - Política - Estadão." 30 set. 2018, Disponível em: https://politica.estadao.com.br/noticias/eleicoes,eleicoes-2018-levam-odio-e-desavenca-as-relacoes,70002525774. Acesso em: 18 mar. 2019.

dados. A exploração do modelo multidimensional do Data Warehouse são só alguns exemplos do que pode ser feito.

As consultas efetuadas e os dados extraídos, foram capazes de demonstrar bem o sentimento dos eleitores a respeito das eleições como um todo e dos candidatos. Muita indiferença dos eleitores em relação às eleições; grande parte das pessoas que possuíam algum sentimento em relação aos candidatos, levaram para o Twitter o sentimento geral sobre os políticos: desaprovação, seja por ações ou ideologias de cada. O fato é que, a amostra deste estudo e sua análise é coerente até certo ponto com os fatos verificados no mundo real, gerando a necessidade de melhorias na aplicação com um todo.

Referências

AGUIAR, Erickson et. al. Análise de Sentimento em Redes Sociais para a Língua Portuguesa Utilizando Algoritmos de Classificação. 2018.

ANDRADE, Carina et. al. O Twitter como Agente Facilitador de Recolha e Interpretação de Sentimentos: Exemplo na Escolha da Palavra do Ano. In: 15ª Conferência da Associação Portuguesa de Sistemas de Informação. 2015.

ARAÚJO, Mateus et. al. Métodos para análise de sentimentos no Twitter. In:Universidade Federal de Minas Gerais. 2016.

CAMILO, Cássio et al. Mineração de dados: conceitos, tarefas, métodos e ferramentas. In: Instituto de Informática. Universidade Federal de Goiás. 2009. p 12- 15.

CAVALCANTE, Paulo Emílio Costa. Um dataset para análise de sentimentos na língua portuguesa. 2017.

CORRÊA, Igor. Análise de sentimentos expressos na rede social Twitter em relação aos filmes indicados ao Oscar 2017. In: Universidade Federal de Uberlândia. 2017.

JUNIOR, V. F., WOSZEZENKI, C., ANDERLE, D. F., SPERONI, R., NAKAYAMA, M. K. (2014). A pesquisa científica e tecnológica. Espacios, 35(9).

JUNQUEIRA, Kássio TC; DA ROCHA FERNANDES, Anita Maria. Análise de Sentimento em Redes Sociais no Idioma Português com Base em Mensagens do Twitter. Anais do Computer on the Beach, p. 681-690, 2018.

KIMBALL, Ralph; ROSS, Margy. The data warehouse toolkit: the complete guide to dimensional modeling. John Wiley & Sons, 2011.

MANSMANN, Svetlana. Building a Data Warehouse for Twitter Stream Exploration.In: University of Konstanz, Germany. 2012.

MORAES, Silvia et. al. 7x1•PT: um Corpus extraído do Twitter para Análise de Sentimentos em Língua Portuguesa. 2015.

NASCIMENTO, Paula et. al. Análise de sentimento de tweets com foco em notícias. 2012.

NOGUEIRA, Rodrigo R. Newsminer: um sistema de datawarehouse baseado em texto de notícias.In: Universidade Federal de São Carlos. 2017

PIZZANI, Luciana et. al. A arte da pesquisa bibliográfica na busca do conhecimento. In: Revista Digital de Biblioteconomia e Ciência da Informação.

TAVARES, Jonatas et al. Soluções de BI 2.0 para Análise de Dados a partir do Twitter®: Eleições 2014.2017.

TREVISAN, Allan. MINERAÇÃO DE TEXTOS NO TWITTER .2015

TSE. Disponível em <http://divulga.tse.jus.br/oficial/index.html>. Acesso em 10 dez. 2018.