**ENRIQUECIMENTO SEMÂNTICO COM ANÁLISE DE SENTIMENTO NA ETAPA DE ETL**

**Estudo de Caso das Eleições de 2018**

*Leonardo Croda[[1]](#footnote-0),Jonathan Suter,[[2]](#footnote-1) Rodrigo Nogueira[[3]](#footnote-2),Daniel Anderle[[4]](#footnote-3)*

**RESUMO**

O twitter é uma das redes sociais mais utilizadas do mundo, no qual um usuário escreve uma expressão com até 280 e diariamente são discutidos os assuntos mais relevantes em todo o mundo. Consumindo dados dessa rede social, os textos foram classificados manualmente em sentimentos (bons, ruins e neutros). A partir dos dados classificados, foi desenvolvido um sistema de data warehouse, que acopla um algoritmo de machine learning. E assim foram qualificados os demais textos, 107393, totalizando os 108693 tweets. A partir dos dados coletados e classificados, foi possível fazer uma análise dos dados da eleição presidencial de 2018.

**Palavras-chave**: Machine Learning, Análise de Sentimento, Data Warehousing, ETL.

**INTRODUÇÃO**

Dentre diversas aplicações em um *corpus* linguístico baseado em textos do *Twitter,* se destacam as pesquisas que explorar 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 (JUNQUERA,2017).

As redes sociais tem grande importância para a sociedade está relacionada ao fato de que as mesmas possuem grande potencial de compartilhamento de informação. Sendo assim, os dados extraídos de uma rede social, podem ser utilizados para o auxílio na tomada de decisão de determinado assunto de cunho estratégico, para uma corporação ou até mesmo um indivíduo (TOMAÉL ,2005).

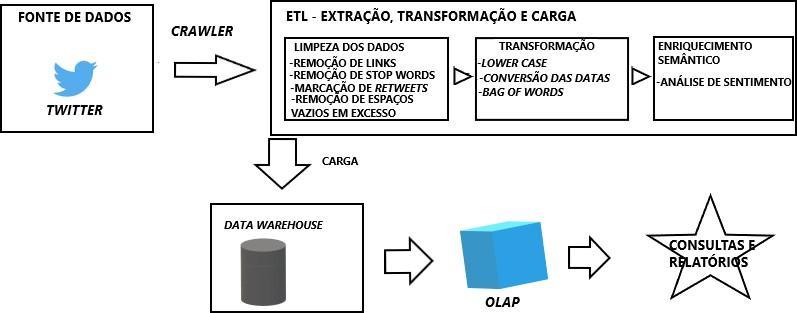
No entanto, por mais interessante que seja a aplicação de aprendizado de máquina para extração de sentimento, o grande desafio no emprego de técnicas de aprendizado de máquina é que 80% de todo o esforço computacional é gasto na etapa de pré-processamento de dados (LOSARWAR, 2012). O desenvolvimento de uma ferramenta que faça a coleta dos dados, realize a limpeza, normalize os mesmos e guarde-os em uma estrutura definida, além de diminuir o esforço nesta etapa, ainda facilita a utilização destes dados por terceiros, permitindo ao usuário que foque-se em sua atividade principal de análise destes dados.

A partir dessa problemática, essa pesquisa tem como objetivo o desenvolvimento de um Data Warehouse alimentado com dados em tempo real da rede social *Twitter*, sob o qual foram coletados e analisados os textos sobre a eleição de 2018.

**PROCEDIMENTOS METODOLÓGICOS**

Uma vez que o produto final da pesquisa é um conjunto de arquitetura de software, complementada de um conjunto de dados, esta pesquisa se enquadra como pesquisa tecnológica (JUNIOR et al. 2014). O desenvolvimento teve como base na arquitetura de um *Data Warehouse* de KIMBAL(2011). A Figura 1 mostra a arquitetura proposta por esta aplicação de *Data Warehouse*. Para que ocorra o armazenamento dos *Tweets* para posterior uso nas consultas, é 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.

**Figura 1.** Fluxo de funcionamento da aplicação

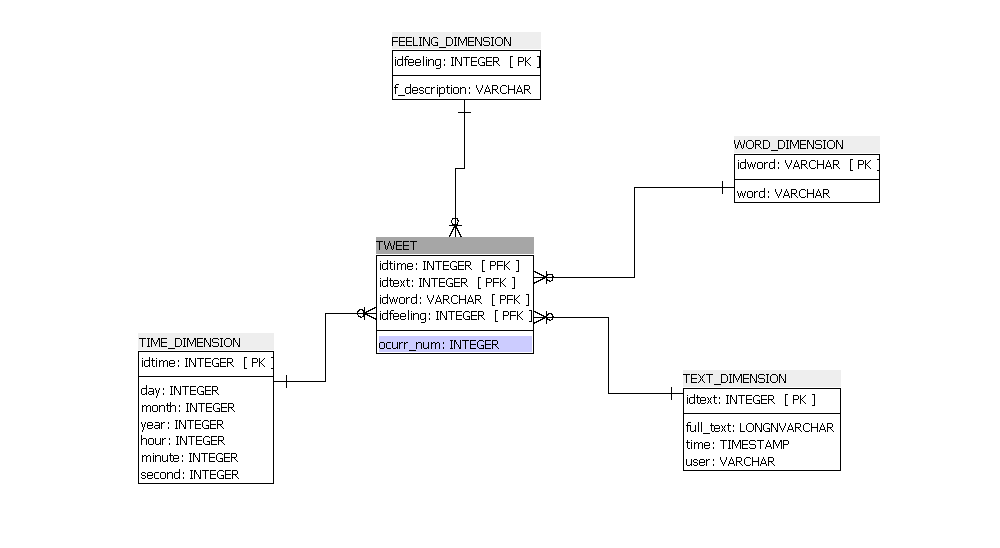


Fonte: Os autores

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 modelo multidimensional, mostrado pela Figura 2, foi implementado com o SGDB PostgreSQL, versão 10.2.1 em 64 bits, com o pgAdmin4. Este Data Warehouse é um ROLAP (Relational On-line Analytical Processing), pois seus dados derivam de uma base dados relacional, são uma fração selecionada de dados de uma base relacional, reorganizada. A tabela “TWEET” é a tabela fato as outras são as dimensões, a tabela que “une as demais”. A tabela “TEXT\_DIMENSION” é a tabela que armazena os textos dos Tweets, o usuário e o momento de criação do Tweet. Na tabela “TIME\_DIMENSION” ficam inseridos todas as combinações dos segundos, minutos, horas dias, meses e o ano que existem entre os meses de Julho e Outubro. “WORD\_DIMENSION” é a tabela em que ficam registrados os termos extraídos dos textos dos Tweets. A tabela “FEELING\_DIMENSION” é responsável por armazenar os sentimentos, no caso, “Positivo”, “Neutro” e “Negativo”.

**FIGURA 2.** Modelo multidimensional desenvolvido



Fonte: Os autores.

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

**RESULTADOS**

Esta seção tem como objetivo ilustrar alguns dos resultados obtidos que podem ser explorados a partir do modelo multidimensional desenvolvido. Para a execução das consultas, foi gerada uma base com dados coletados entre setembro e outubro de 2018 utilizando como termo de busca “eleicoes2018”. Vale ressaltar que a API não confere acesso total à base de dados da rede social e há limite de coleta por dia. A começar pela palavra de maior menção.

**Tabela 2.** Ranking de palavras com maior número de ocorrências

|  |  |
| --- | --- |
| **Palavra** | **Quantidade** |
| eleições2018 | 51458 |
| bolsonaro | 24424 |
| candidato | 10559 |
| haddad | 9726 |
| diz | 8184 |
| presidente | 7188 |
| contra | 7160 |
| eleições | 7125 |
| sobre | 6443 |

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.

Ao analisar as menções diretas por candidato, os valores foram Bolsonaro:24424, Haddad:9726, Ciro:4510, Alckmin:1897, Daciolo:1819, Marina: 1817, Boulos:1098, Meirelles: 588, Amoêdo: 192, Álvaro:139, Goulart: 78,Vera:61, Eymael:13.

No primeiro momento, é possível observar que nenhum candidato obteve mais citações boas que ruins, refletindo que o sentimento geral entre os *tweets* foi ruim, e que nenhum candidato conseguiu obter uma grande aprovação dos eleitores.Com base nos textos. O candidato que obteve a maior quantidade de citações com sentimento “bom” foi o Bolsonaro. Entretanto, também foi o candidato que obteve a maior quantidade de citações classificadas como “ruim”.

Ao comparar os resultados com os da eleição ( Disponível em <<https://especiais.gazetadopovo.com.br/eleicoes/2018/resultados/votacao-candidatos-presidente-brasil/>>), Obtém-se certa equivalência entre os resultados extraídos do *Data Warehouse* e as intenções de voto, apesar de muitas divergências, há de se considerar ainda que a base possui muitas citações qualificadas como neutras, podendo ocorrer maior distribuição para as citações com sentimento “ruim” e “bom”.

**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 dados. As consultas efetuadas para explorar o 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**

JUNIOR, Vanderlei FREITAS et al. **A pesquisa científica e tecnológica**. Espacios, v. 35, n. 9, 2014.

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; CASERTA, Joe. **The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data**. John Wiley & Sons, 2011.

LOSARWAR, V.; JOSHI, D. M. **Data preprocessing in web usage mining**. In: Proceedings of International Conference on Artificial Intelligence and Embedded Systems. New Asia, Singapura: [s.n.], 2012. p. 15–16.

TOMAÉL, Maria et. al. **Das redes sociais à inovação.**Ciência da Informação. Brasília: 2005. Volume 34, numero 2. p. 93-104.

1. Cursando Bacharelado em Sistemas de Informação, Instituto Federal Catarinense – Campus Camboriú, lccroda@gmail.com [↑](#footnote-ref-0)
2. Egresso do Bacharelado em Sistemas de Informação, Instituto Federal Catarinense – Campus Camboriú, jonathan.vinicius.suter@gmail.com [↑](#footnote-ref-1)
3. Mestre em Ciência da Computação, professor do Instituto Federal Catarinense – Campus Camboriú, rodrigo.nogueira@ifc.edu.br [↑](#footnote-ref-2)
4. Doutor em Engenharia do Conhecimento, professor do Instituto Federal Catarinense – Campus Camboriú, daniel.anderle@ifc.edu.br [↑](#footnote-ref-3)