Aplicação para Coleta de Mensagens no Twitter e Análise de Sentimentos e Emoções

Levi F. Teixeira[1], Matheus Quirino A. dos Santos[2], Ivan C. A. de Oliveira[3]

Universidade Cruzeiro do Sul (UNICSUL) Campus São Miguel Paulista
Caixa Postal 225 – 08060-070 – São Paulo – SP – Brazil

[1]Departamento de Ciência da Computação
Universidade Cruzeiro do Sul (UNICSUL) – São Paulo, SP – Brazil

leeviife@gmail.com - matheusc84a@gmail.com

Abstract. *Based on the high number of textual publications on Twitter and the growing number of users, this work assesses the feeling and emotions expressed by the user in their Tweets based on the hashtags #nubank, #santander and #itau. For this, a conceptual map was constructed to obtain the concepts associated with the theme and delimit the scope of the work, a literature review was carried out that allowed finding techniques for collecting messages on Twitter, and treating them with Language Processing Natural; machine learning techniques for the generation of predictive model such as the Support Vector Machine that detects the feeling, positive, negative or neutral and in the identification of emotions using the Naive Bayes that identified sadness, happiness and anger. Finally, the application architecture and its development in the Python language and the Angular framework for the front end were elaborated.*

Resumo. *Tendo como base a alta quantidade de publicações textuais no Twitter e o crescente aumento de usuários, este trabalho avalia o sentimento e as emoções expressas pelo usuário em seus Tweets com base nas hashtags #nubank, #santander e #itau. Para isso, foi realizada a construção de um mapa conceitual para obter os conceitos associados ao tema e delimitar o escopo do trabalho, foi efetuada a revisão de literatura que permitiu encontrar técnicas de coleta de mensagens no Twitter, e tratando-as com Processamento de Linguagem Natural; técnicas de aprendizado de máquina para a geração do modelo preditivo como o Suport Vector Machine que detecta o sentimento, positivo, negativo ou neutro e na identificação das emoções usando o Naive Bayes que identificou tristeza, felicidade e raiva. Por fim, foi elaborada a arquitetura da aplicação e o seu desenvolvimento na linguagem Python e o framework Angular para o front end.*

Palavras-chave: *Análise de Sentimentos. Twitter. Análise de Emoções. Aprendizado de Máquina. Coleta de Mensagens.*

1. INTRODUÇÃO

A partir dos avanços dos meios de comunicação e a evolução das redes sociais, os meios de expressar-se têm sido cada vez mais frequentes e fundamentais, tornando visível a opinião pública sobre um certo produto ou serviço. O Twitter, gera uma massiva coleção de *tweets* crescente dia a dia e destaca-se por opiniões expressas em modo textual, fundamentais para uma mineração de sentimentos e emoções eficiente (KRIKORIAN, 2013).

Uma das estratégias para lidar-se com uma grande quantidade de texto é o uso de sua mineração, que neste caso visa extrair palavras que sintetizam as emoções mais frequentes. Para servir de base a isto, a coleta de dados no Twitter, é feita por meio de sua própria Application Programming Interface (API), esta, utiliza *hashtags* como referência. Os sentimentos são identificados pelo Processamento de Linguagem Natural (PLN). Esta identificação é feita através de palavras chaves presentes, que compõem um padrão emocional, que por sua vez formam os sentimentos.

Algumas das técnicas de processamento de texto muito utilizadas, são fundamentais no tratamento de ruídos nos documentos, como *stopwords* e *normalization*. Também são utilizadas algumas técnicas de *Machine Learning* (ML), como: *Naive Bayes* (NB) e *Suport Vector Machine* (SVM).

Este artigo foi organizado em seções, conforme descrição a seguir. Na seção 2, é abordada de forma breve a metodologia utilizada no desenvolvimento deste trabalho. Alguns conceitos associados como a coleta de texto, sua mineração e bases de dados a serem utilizadas, são apresentados com a fundamentação teórica na seção 3. A seção 4 apresenta a base de dados e a seção 5 é denotada pelo desenvolvimento e definição da proposta. Por fim, constam as considerações finais e indicações dos próximos passos deste trabalho na seção 6.

2. METODOLOGIA

Durante o desenvolvimento do projeto, foram realizadas as seguintes atividades:

- Criação de perguntas para auxiliar no desenvolvimento do projeto.
- Pesquisa bibliográfica
- Elaboração de um Mapa Conceitual.

- Definição da proposta da aplicação.
- Desenvolvimento e implementação das técnicas de PLN e ML.
- Métricas de erro como *accuracy*, f1 e *recall*
- Testes e análises de resultados.

A primeira etapa serviu para nortear o desenvolvimento do projeto com base na seguinte pergunta: “Quais as técnicas e algoritmos de mineração de texto podem ser necessárias para análise de sentimentos e emoções no Twitter?”. Na segunda etapa, foi possível fazer a pesquisa bibliográfica, o qual permitiu identificar conjuntos de textos relacionados a análise de sentimento, algoritmos e técnicas utilizadas.

Na próxima etapa foi construído um mapa conceitual baseado em livros, trabalhos acadêmicos e artigos coletados, que auxiliou na organização do conhecimento e nas prioridades definidas referentes à análise de sentimentos e emoções.

A partir do conhecimento obtido, foi possível definir a proposta do projeto, determinar a porcentagem de satisfação sobre um produto e seus sentimentos base, estes, gerados por emoções em destaque nos *tweets*. Sendo assim, um ator principal que avalia, de acordo com o seu sentimento sobre o produto.

Para o desenvolvimento da ferramenta foram utilizadas as técnicas de pré-processamento PLN e Term Frequency – Inverse Document Frequency (TF-IDF) e algoritmos de ML que foram obtidos a partir da fundamentação teórica.

Foram utilizadas técnicas de análise de erro como *accuracy*, f1 e *recall* para medir a assertividade do algoritmo de análise de sentimentos e emoções criado.

A partir da análise dos resultados foram realizados testes no mesmo período do desenvolvimento do projeto, permitindo uma melhor correção das falhas no decorrer de sua elaboração. Com base nisso, viu-se a necessidade de utilizar técnicas e métodos adicionais para solução dos problemas de balanceamento encontrados.

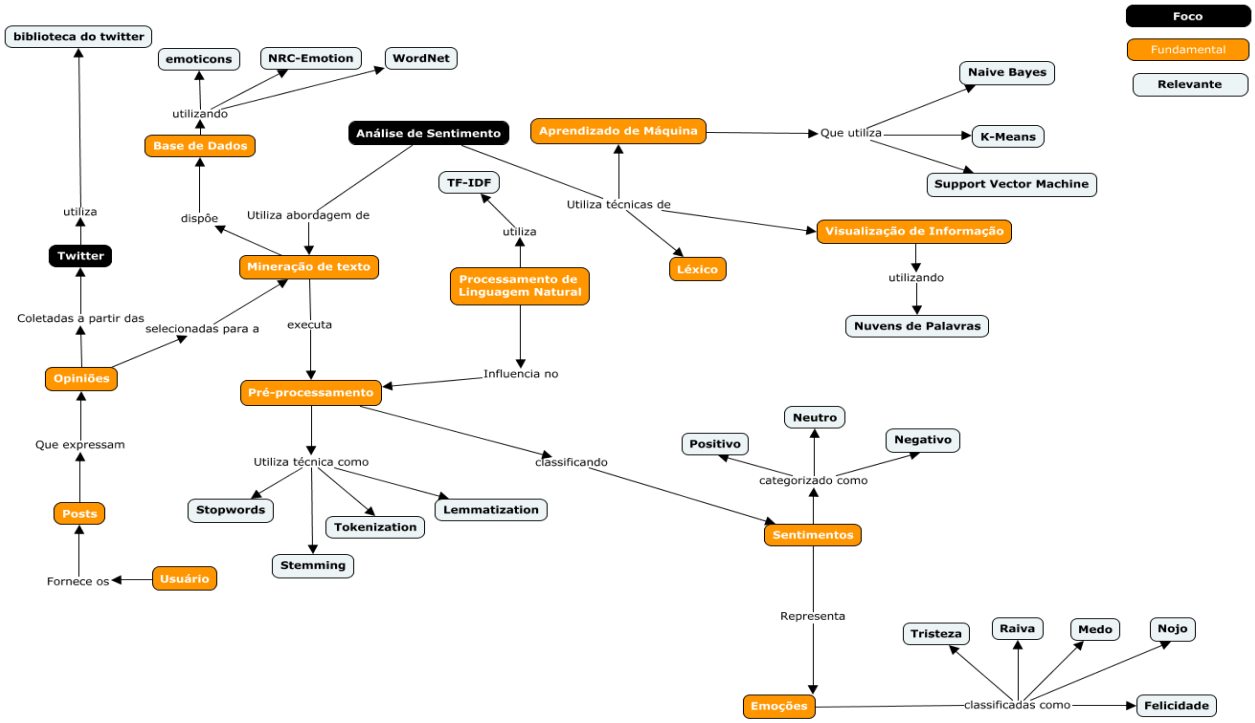
3. FUNDAMENTAÇÃO TEÓRICA

3.1 Mapa Conceitual

Uma das estratégias para nortear o trabalho foi dispor de um mapa conceitual que permitiu organizar os conceitos do assunto. O mapeamento entre os conceitos forneceu uma visão geral a qual foi de grande ajuda para definir o escopo do projeto.

O mapa foi divido em cores para apresentar os assuntos, separados conforme ilustra a Figura 1.

Figura 1. Mapa Conceitual



Fonte: Elaborado pelo Autor.

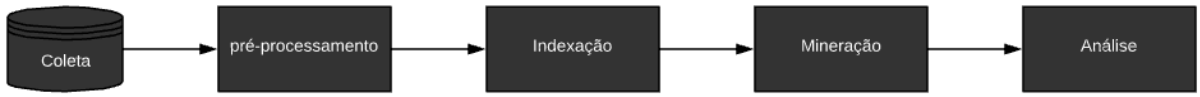
Representado por “foco” na cor preta, “fundamental” na laranja e “relevante” na cinza claro. Os conceitos apresentados possuem a finalidade de mostrar somente as técnicas mais relevantes.

3.3 Mineração de Texto

Cabral, Bruno e Corado (2015), definem Mineração de texto ou *Knowledge Discovery from Text* (KDT) como uma ampliação das técnicas de mineração de dados, porém utilizando de fluxos procedimentais. Para Carvalho Filho (2014), é uma técnica de processo de extração de padrões ou conhecimentos não triviais obtidos de bases de dados não estruturados.

O processo de mineração de texto pode ser dividido por etapas para a sua aplicação, sendo mostrado a seguir um modelo proposto por (ARANHA; PASSOS, 2006), e os processos que precisam ser seguidos na Figura 2.

Figura 2. Processo de Mineração de Textos



Fonte: Elaborado pelo Autor.

Na etapa de coleta, são construídas operações que funcionam em tempo real sem interrupção, numa grande escala de volume de dados para consultar ou extrair conhecimentos, foi utilizado o pré-processamento para limpar os ruídos ou informações não necessárias, implementando como exemplo técnicas de PLN. Depois identificou-se os dados na etapa de indexação, extraindo dos arquivos coletados. Em seguida, algoritmos foram aplicados para encontrar padrões de extração de conhecimento. Por fim, foi feita a análise a partir do conhecimento adquirido.

3.4 Pré-processamento

Normalmente os dados coletados em documentos de texto e redes sociais, apresentam muitos ruídos, o que prejudica a qualidade do conhecimento obtido. Um dos desafios de uma boa análise de sentimentos e emoções em redes sociais, é lidar com erros gramaticais dos usuários e dialetos próprios das redes, fazendo necessária a utilização de técnicas eficientes para resolver estes problemas.

A partir do momento que se possui uma base de dados para utilizar, sendo o seu texto bruto chamado de *corpus,* aplica-se técnicas que podem ser utilizadas nesta etapa, e assim retira-se

algumas palavras repetidas, sobrando apenas o que é chamado de léxico, (CAPOBIANCO,2016). Ademais, Santos (2018) utiliza de técnicas de PLN para a captura de informações relevantes, como a análise de sentimento.

3.4.1 PLN

A PLN faz uso de um conjunto de técnicas que permitem à máquina manipular o processo de linguagem com nós humanos. Capobianco (2016) mostra que é possível lidar com elementos linguísticos, como estruturas gramaticais e análise de frequência de palavras, sendo esta última muito útil para a análise de sentimentos com o algoritmo TF-IDF.

3.4.1.1 TF-IDF

TF-IDF é um algoritmo utilizado pelo pré-processamento de dados, para analisar a frequência em que uma palavra é referida ao decorrer do texto. Assim é possível calcular a relevância perante a repetição de um termo no documento (VILLARROEL, 2020).

A execução do TF-IDF começa com a descoberta do número de vezes que uma palavra aparece no texto, dividido pela quantidade de palavras no total (1).

$$tf(i) = \frac{\sum Pi}{\sum Pd} \quad (1)$$

- A somatória de ocorrências representada como P*i* utilizando a palavra *i*, dividido pela quantidade total de palavras *Pd.*

Verificou-se a frequência inversa, pelo cálculo logaritmo da quantidade total de palavras, divido pelo número de vezes em que aparece a mesma (2).

$$idf(i) = log(\frac{|D|}{di}) \quad (2)$$

- Logaritmo de *D*, sendo a quantidade de palavras do *corpus*, e *Di*, a frequência da mesma

Obtém-se a resposta (3).

□□□□(□) = □□(□) ∗ □□□(□) (3)

Conforme o cálculo da frequência tende ao valor total do texto, o resultado fica cada vez mais perto de 1. Assim se torna possível verificar a importância da palavra no texto (CAPOBIANCO, 2016).

3.4.2 Técnicas de pré-processamento

Rani e Kumar (2017) definem algumas técnicas utilizadas na etapa de pré-processamento, sendo elas:

- ***Stopwords*:** Técnica que retira palavras que não influenciam na análise de texto, como preposições, artigos, conjunções entre outras classes gramaticais. A aplicação desta técnica visa melhorar a performance do processamento (SANTOS,2018).
- ***Tokenization*:** Quebra o texto quando tiver encontrado um espaço entre as palavras. Cada separação é chamada de *token* (RANI; KUMAR,2017).
- ***Stemming*:** Visa transformar as palavras em sua forma raiz, usando como exemplo "movido" e "movimento" que vira "mover" (SANTOS, 2018).
- ***Normalization*:** Visa normalizar o conteúdo abreviado utilizados nas redes sociais e identificar as gírias (RANI; KUMAR, 2017).

3.5 Tipos de Abordagem

Medhat, Hassan e Korashy (2014), apresentam duas abordagens para a análise de sentimentos e emoções, sendo elas: ML e análise léxica.

3.5.1 Machine Learning

Segundo Thayná e Rossi (2017), ML é uma subárea da Inteligência Artificial (IA) que visa adquirir de maneira automática, possibilitar que o computador aprenda padrões com as experiências acumuladas a partir de soluções bem-sucedidas de problemas passados.

Becker e Tumitan (2013) definem que o ML fornece uma análise de treinamento ao qual o modelo preditivo resulta na etapa de aprendizagem realizada com métricas, como acurácia (capacidade do modelo de prever corretamente) e precisão (número de instâncias previstas corretamente em uma dada classe).

ML aborda duas formas para executar um treinamento, sendo elas: modelo supervisionado e modelo não supervisionado.

a) **Aprendizagem Supervisionado:** Trabalha de forma indutiva, recebendo treinamento pré-classificado sobre um corpus rotulado com classificações como positivo, negativo ou neutro usando como referência análise de sentimentos.

b) **Aprendizagem Não-supervisionado:** com foco em problemas de agrupamento utilizando de algoritmos com exemplos não rotulados extraindo as características do conjunto (THAYNÁ; ROSSI, 2017).

3.5.1.1 Algoritmos de Machine Learning

No processo de mineração de opiniões nota-se um predomínio no uso de classificação supervisionada utilizando um conjunto de algoritmos: NB e SVM.

a) ***Naive Bayes*:** O NB é um algoritmo simples de classificação probabilístico que se baseia principalmente no teorema de Bayes. Ele é frequentemente utilizado como base na classificação de textos por ser rápido e fácil de implementar.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \quad (4)$$

Para Capobianco (2016), o teorema se baseia na probabilidade de um evento que venha a ocorrer dado um conhecimento a priori relacionado ao evento. Sendo assim, B representa um evento que ocorreu previamente e A, um evento que depende de B, para calcular a sua probabilidade.

b) ***Support Vector Machine*:** O SVM separa classes de busca distintas e analisa os pontos de cada grupo mais próximo de outras classes, baseado na distância entres elas. Escolhendo a melhor separação usando como comparação a distância do hiperplano (linha que separa as classes) com a maior margem.

3.6 Análise de sentimento

A análise de sentimento é um campo da linguística computacional em que o objetivo é descobrir a avaliação que o criador do texto expressa sobre tal tópico ou produto ao qual se refere. Para Liu (2010), essa área é útil tanto para aplicações voltadas para empresas e organizações, quanto para indivíduos. Seu objetivo não é categorizar os textos por tópicos, e sim classificá-los

com base na emoção contido em determinado documento. Normalmente, a classificação é baseada em sentimentos positivos e negativos. Segundo Pang e Lee (2008), a análise, de uma forma mais abrangente, é usada para o tratamento computacional de opinião, e subjetividade em textos.

Turney (2002) propôs vários tipos de sentimento para classificar os *tweets* usando *hashtags* e *emojis* como rótulos. Além disso, Barbosa e Feng (2010) propuseram uma abordagem de duas etapas para classificar os sentimentos dos *tweets* usando SVM, classificadores com características abstratas.

Rani e Kumar (2017), definem análise de sentimentos como sendo um processo em que, a partir do uso de técnicas de mineração de texto e PLN, é possível identificar e classificar as opiniões dos usuários em trechos de textos ou documentos, onde estes podem expressar diferentes polaridades de sentimentos, sendo eles: positivo, neutro ou negativo. Um dos grandes desafios está relacionado ao sarcasmo e a ironia, que estão muito ligados ao domínio do contexto da aplicação e ao idioma (BECKER; TUMITAN, 2013; SERRANO-GUERRERO et al., 2015).

Além dos sentimentos, é possível detectar com a mineração de texto as emoções utilizando algoritmos de ML ou com dicionários léxicos. Para Guedes (2014), há diversas teorias no campo da computação que estudam com auxílio de correntes da psicologia a relação das palavras com alguma emoção específica, sendo elas: Raiva, tristeza, felicidade, alegria e nojo.

3.7 Coleta de dados

Os termos são extraídos a partir de uma análise de frequência das palavras ou frases em cada documento e em todo o conjunto de informação. Nas técnicas linguísticas, os termos de indexação são extraídos utilizando técnicas de PLN, como, por exemplo, análise lexical, morfológica, sintática e semântica.

3.7.1 Twitter

Uma rede de grande valor agregado de pensamentos críticos é o Twitter, de caráter altamente social que permite que você poste mensagens com 140 caracteres ou menos; essas mensagens são chamadas de *tweets*. Tais recursos produzem um grande volume de dados diariamente, como por exemplo o Twitter, que produz cerca de 500 milhões de *tweets* por dia (KRIKORIAN, 2013).

O acesso à dados de mídias sociais tornou-se muito difícil, principalmente pelo seu uso indevido. Contudo, para a realização dos experimentos foi possível coletar dados de postagens reais de brasileiros no Twitter no ano de 2017, que totalizaram 8.199 registros. Segundo O'Connor et al. (2010), a opinião expressa pelos usuários em redes sociais tem sido monitorada também por governos e empresas na busca pela compreensão do que uma população pensa sobre um determinado produto ou serviço.

Vista a popularidade do Twitter, a análise de sentimentos em *tweets* tem atraído mais atenção (PARIKH; MOVASSATE, 2009; BARBOSA; FENG, 2010; TURNEY, 2002), seguida da abordagem de aprendizagem de máquina para análise sentimento de *tweets*.

3.8.1.1 Bibliotecas referentes ao Twitter

Utilizou-se como uma das principais bases documentais de desenvolvimento, a documentação oficial da API do Twitter (http://dev.twitter.com/doc). Tendo em vista trabalhar essas opiniões críticas baseadas no sentimento dos usuários notou-se a possibilidade de extrair estes dados usando uma biblioteca simples para a API web do Twitter, que está disponível por meio de um pacote chamado Twitter (GITHUB) e pode ser instalado com o módulo “easy_install” do Python.

Assim, busca-se apresentar um algoritmo de análise de sentimentos e emoções de *tweets* com o auxílio do algoritmo de PLN e o Naïve Bayes, implementando este sistema sob a linguagem de programação Python, juntamente com a utilização da biblioteca NLTK.

3.9 Trabalhos relacionados

O IBM Watson, plataforma aberta, que utiliza modelos de ML a serem implementados em modelos de negócio, oferece também a possibilidade de utilizar algumas ferramentas já desenvolvidas nas plataformas de cloud da Google, Microsoft e AWS. Para uso de aplicações, um dos serviços que Watson utiliza é o *“Watson Natural Language Understanding”* serviço que desfruta da extração de metadados contidos nos textos, sendo possível detectar sentimentos, emoções e funções semânticas utilizando a PLN. (IBM, 2021).

4. Base de dados

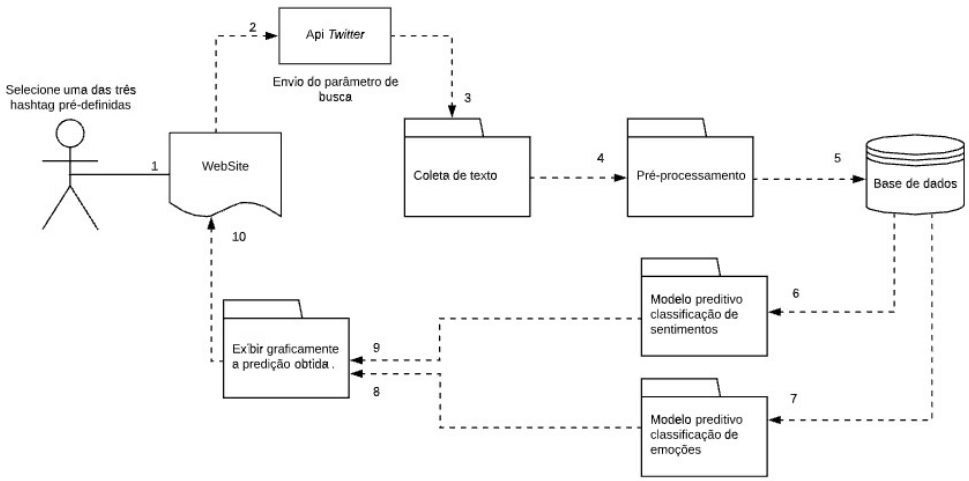
Utilizou-se para o desenvolvimento deste projeto uma base de dados pública disponibilizada no GitHub (https://github.com/stacktecnologias/stack-repo), fornecida pelo governo do estado de Minas Gerais (MG). Esta base, é composta por 8199 linhas classificadas em “positivo”, “negativo” e “neutro”, serviu como base para fundamentação da base de dados do projeto.

5. Proposta de solução e desenvolvimento

Este trabalho tem como objetivo desenvolver uma aplicação que realiza a avaliação dos sentimentos e emoções disponibilizadas no Twitter, sobre cinco hashtags, sendo elas: #nubank, #santander, #itau, #bradesco, #caixaeconomica; partindo da coleta dos textos, o seu tratamento e posteriormente rotulagem e classificação dos sentimentos com o NB e SVM.

A seguir é apresentado um diagrama de arquitetura para complementação do projeto na Figura 3.

Figura 3: Diagrama de arquitetura



Fonte: Elaborado pelo Autor

Expressa-se por meio deste diagrama arquitetura, o qual compõem o fluxo para obtermos o objetivo especificado. Consiste na escolha de uma das três *hashtags*, dentre elas *#nubank, #santander, #itau*, #bradesco, #caixaeconomica, a ser usada na coleta de dados no passo 1. Em seguida o passo 2 ilustra o uso deste valor na API do Twitter e a coleta do texto é realizada no

passo 3. Logo após a coleta, o texto é tratado em um pré-processamento no passo 4 e armazenamento no passo 5. A classificação destes dados armazenados é treinada no passo 6 e 7. Por fim as polaridades resultantes são expressas graficamente aos usuários no passo 8 e 9. O retorno dos resultados na página web ocorre em seguida no passo 10 finalizando o fluxo.

5.1 Classificação de sentimentos e emoções

Para a execução desta proposta foi necessário criar um algoritmo, que permite coletar os dados do Twitter e posteriormente armazena-los em um *Data Set*. Na primeira etapa, o algoritmo executa o filtro dos *tweets* utilizando como parâmetro de busca *hashtags* com nomes de instituições bancárias, na linguagem Português do Brasil (PT-BR). Na segunda etapa, o algoritmo guarda os dados em um dicionário que posteriormente, usando a biblioteca Pandas, é convertido para um *DataSet*.

Com a coleta feita, classificou-se manualmente os *tweets* com os sentimentos (positivo, negativo e neutro) e com as emoções (tristeza, raiva e felicidade).

Definiu-se estas emoções a partir da intepretação do texto coletado e se era passível de ser classificado. Entretanto, alguns dados da coleta não foram possíveis de serem classificados por não expressarem nenhuma emoção, assim reduzindo consideravelmente a base de dados total para as emoções. Ao tratarmos o conteúdo de referência ao tema bancário, não encontramos tanta diversidade de emoções sólidas, como o nojo e alegria. Não foi encontrado nenhum *tweet* classificado como “nojo” e no caso da “alegria” os poucos que foram encontrados, foram classificados como “felicidade” para não favorecer ao desbalanceamento da base própria.

Com a coleta feita, os ruídos do texto foram retirados usando expressões regulares eliminando assim os *links* e números contidos nas postagens, como ilustra Figura 4, que utiliza das *hashtags* #bradesco e #santander.

Nota-se que em muitos itens as Emoções estão com a expressão **NaN**, significando como emoção não classificada.

Figura 4. Base de dados *hashtags* #bradesco e #santander

: 0	Twitter	Hashtag	Sentimentos	Emoções
0	@juliette ChenofoBia mimimi até quando isso co...	bradesco	negativo	tristeza
1	#QueVergonhaItaú e que venha futuras investiga...	bradesco	negativo	raiva
2	Com aval de NY, Ibovespa avança em sessão de a...	bradesco	neutro	NaN
3	Assim não é possível. Bloquearam mais uma vez ...	bradesco	negativo	triste
4	ATUALIZADA! Se liga como ficou a tabela do #Br...	bradesco	neutro	NaN
...	...	...	...	...
84	Alguém passa ódio por nunca conseguir ser aten...	santander	negativo	raiva
85	Rumo deixa carteira ESG do Santander em julho;...	santander	neutro	NaN
86	#UmPassoMudaTudo e as exposições #SpaceAdventu...	santander	neutro	NaN
87	O movimento sindical enviou ao Santander, uma ...	santander	neutro	NaN
88	Hora de investir da renda fixa? Veja o tesouro...	santander	neutro	NaN

Fonte: Elaborado pelo Autor.

Foram coletados no total, 1430 *tweets* que após a lapidação que removeu os itens redundantes e inconsistentes obteve-se o número de 923 itens conforme representa a Figura 5.

Figura 5. Base de dados própria

	Twitter	Hashtag	Sentimentos	Emoções
0	@AnaMartaLimaos funcionarios do @itau tem que ...	itau	negativo	raiva
1	@thiadm os funcionarios do @itau tem que apren...	itau	negativo	raiva
2	@Julioliveira_ os funcionarios do @itau tem qu...	itau	negativo	raiva
3	@Martaclari os funcionarios do @itau tem que a...	itau	negativo	raiva
4	@edegarfaria @itau os funcionarios do @itau te...	itau	negativo	raiva
...	...	...	...	...
919	Alguém passa ódio por nunca conseguir ser aten...	santander	negativo	raiva
920	Rumo deixa carteira ESG do Santander em julho;...	santander	neutro	NaN
921	#UmPassoMudaTudo e as exposições #SpaceAdventu...	santander	neutro	NaN
922	O movimento sindical enviou ao Santander, uma ...	santander	neutro	NaN
923	Hora de investir da renda fixa? Veja o tesouro...	santander	neutro	NaN

Fonte: Elaborado pelo Autor.

Utilizou-se das mesmas técnicas de remoção de ruídos e de dados duplicados, na base pública disponibilizada pelo governo de MG, totalizando 5750 *tweets*, possibilitando assim, a

união com a base de dados própria.

5.2 Resultados obtidos

Foram feitos testes utilizando três conjuntos de dados, que foram utilizados para encontrar o melhor resultado possível, baseados na acurácia e na matriz de confusão. Segregou-se a base própria e a base do governo de MG em dois vetores, um de sentimentos, outro com os textos coletados no Twitter e também se utilizou a base coletada para separação de dois vetores, sendo eles de emoções e textos coletados no Twitter.

Primeiramente foi feito o uso do TF-IDF, sendo o mais preciso para lidar com o processamento de texto, para valorar todas as palavras obtidas em frequencias numéricas. Tais frequências numéricas possibilitam a criação de uma matriz que a partir dos dados armazenados possibilitam o cálculo do algoritmo. A matriz armazena o peso gerado a partir do valor numérico de cada palavra do *tweet* coletado.

Foi utilizada a biblioteca *Sklearn* para aplicar os algoritmo TF-IDF no processo acima, e para os algoritmos SVM e NB na criação do modelo preditivo. Utilizou-se também a biblioteca NLTK para criação de métodos de pré-processamento com base em *stopwords* e *tokens*.

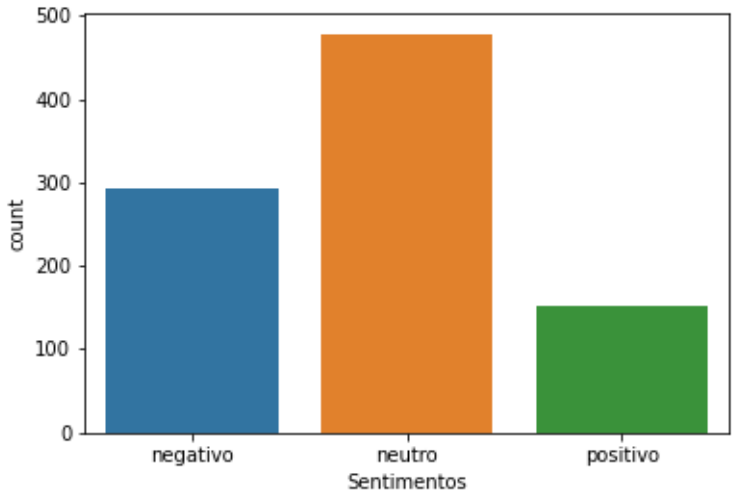
Dentre os parênteses, o primeiro valor significa o primeiro *tweet* encontrado, o segundo valor representa a frequência numérica das palavras deste *tweet*. Fora dos parênteses, o valor representa o resultante do cálculo feito e transformado, sendo ele a escala de probabilidade utilizada no NB.

5.2.1 Balanceamento da base de dados

Antes de treinar o algoritmo, notou-se que a base apresenta desbalanceamento, ou seja, com uma desigualdade de dados entre as classes, sendo assim necessário o balanceamento para não existir problemas com falsos positivos, negativos ou neutros.

Foi analisada a base coletada para presente trabalho para sentimentos e emoções e a base fornecida para os sentimentos. É possível verificar na base pessoal que existe uma presença predominante da classe Neutro sob as outras, como mostra a Figura 6.

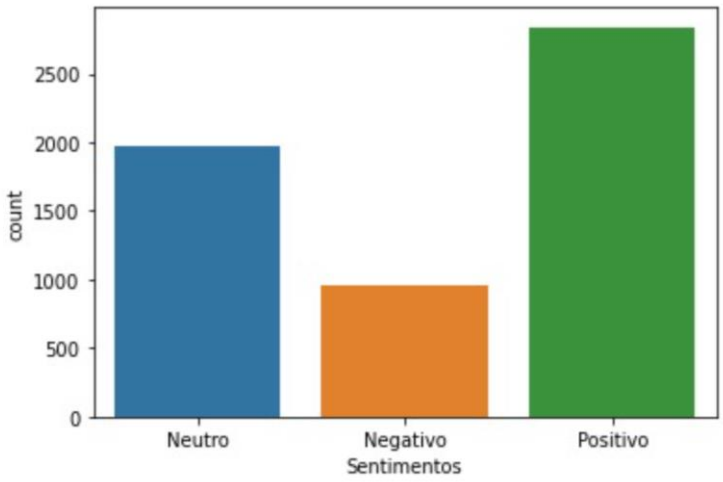
Figura 6. Sentimentos da base de dados coletada



Fonte: Elaborado pelo Autor.

Já na base fornecida, não temos uma certa uniformidade na proporção dos dados com um grande diferencial na classe Positiva, como demonstra a Figura 7.

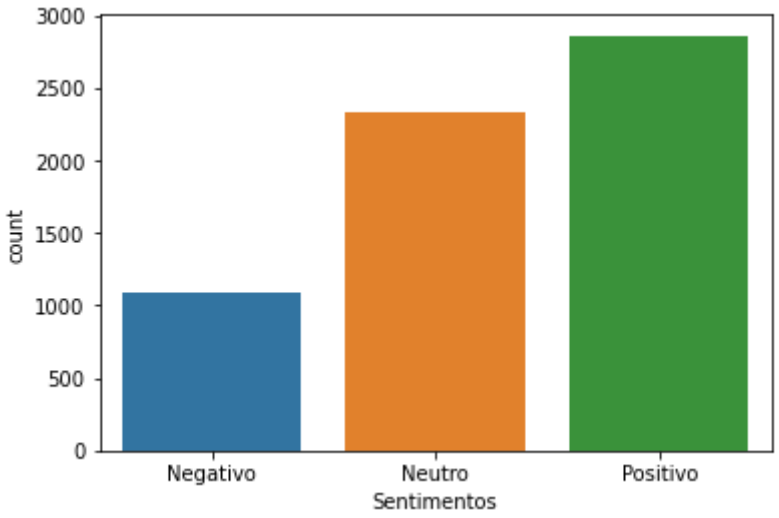
Figura 7. Sentimentos da base de dados do governo de MG.



Fonte: Elaborado pelo Autor.

Nota-se que ao juntar a base coletada com a fornecida é possível notar uma mudança, sendo agora o sentimento positivo o mais presente, conforme a Figura 8.

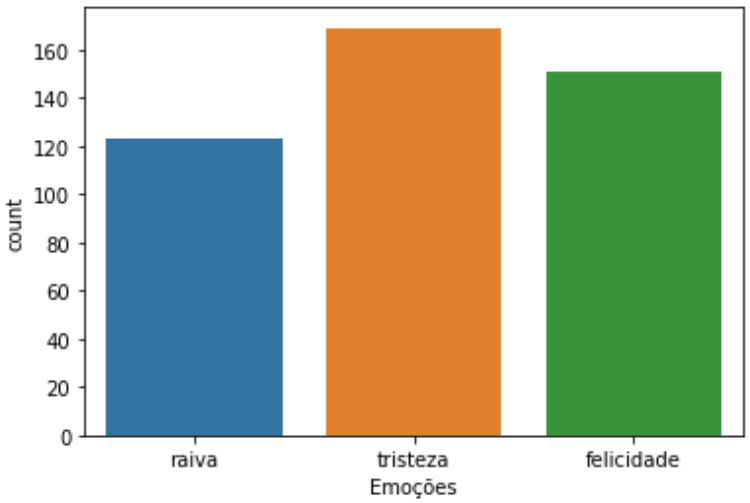
Figura 8. Sentimentos da base de dados unificada



Fonte: Elaborado pelo Autor.

É possível verificar que diferentes dos sentimentos, na base de emoções não há uma discrepância entre as classes, sendo as suas separações melhor distribuídas, conforme a Figura 9.

Figura 9. Base de dados própria das emoções



Fonte: Elaborado pelo Autor.

O algoritmo *NearMiss*, é um método de subamostragem, que consiste em escolher exemplos a partir da distância da classe majoritária para exemplos da classe minoritária. Ao aplica-lo foi possível tornar a base mais equilibrada e com uma quantidade menor de resultados falsos.

É feito um cálculo entre as classificações de sentimentos maiores com as menores, as instancias das classes maiores que possuem distancia mais curta são selecionados com as menores e caso haja m instancias da classe menor, o retorno do algoritmo será m*n da classe maior (BARBOSA; MIRANDA; MELO; SILVA, 2019).

Logo após é executada uma função chamada train_test_split que separa a base em treino e teste a partir de quatro parâmetros, as *features* sendo elas os dados já transformados, as classes de sentimento (positivo negativo e neutro) e as de emoções (felicidade, tristeza e raiva), além do valor que define a separação de 70% para treino e 30% para o teste.

5.2.2 Resultados

Ao longo dos testes foram obtidos os seguintes resultados em três modos extração, sendo eles a base mista, composta pela base própria somada à base do Governo de MG e ambas as bases separadas.

Nota-se que teve um aumento de acurácia ao comparar a utilização do algoritmo NB para os sentimentos ao utilizar técnicas de pré-processamento. Já com a utilização do SVM a acurácia foi maior sem as técnicas.

Ao analisar as emoções, foi possível indentificar que a acurácia foi melhor utilizando as técnicas de pré-processamento (*stopword* e *tokenization*) no SVM, porém no NB a medida de acurácia rmse foi melhor sem elas, conforme ilustra a Figura 10.

Figura 10. Base própria e do governo de MG

Base própria e base do Governo de Minas Gerais				
	NB sentimento	NB sentimento com stopword e token	SVM com stopword e token	SVM de sentimento
Acurácia:	0.816077954	0.834348356	0.84409257	0.851400731
	SVM emoção	SVM de emoção com stopword e token	NB emoção com stopword e token	NB emoção
Acurácia:	0.806451613	0.827956989	0.838709677	0.870967742

Fonte: Elaborado pelo Autor.

É possível verificar que a acurácia apresentou o mesmo padrão de resultado que a base conjunta, com uma única diferença de que os valores foram menores possivelmente ocasionados pelo tamanho da base ser significativamente menor, conforme mostra a Figura 11.

Figura 11. Base de dados coletada

	Base própria			
	NB sentimento com stopword e token	NB sentimento	SVM sentimento com stopword e token	SVM sentimento
Acurácia:	0.816077954	0.552631579	0.561403509	0.614035088

Fonte: Elaborado pelo Autor.

Novamente o padrão de acurácia se repete, no entanto os resultados foram melhores, pois estão somados aos resultados da base própria, soma esta que torna base do governo desbalanceada. Concluído a partir dos resultados das tabelas acima, conforme representa a Figura 12.

Figura 12. Base do governo de MG

	Base do Governo de Minas Gerais			
	NB sentimento	NB sentimento com stopword e token	SVM sentimento	SVM sentimento com stopword e token
Acurácia	0.878151261	0.883753501	0.915966387	0.910364146

Fonte: Elaborado pelo Autor.

Baseando-se nos melhores resultados encontrados, obtivemos as matrizes de confusão das bases. O algoritmo SVM possuiu maior acurácia para os sentimentos, e o NB para as emoções.

Pode-se notar na matriz abaixo que houve uma perda considerada no tamanho total dos dados a partir do balanceamento, apesar disso, tivemos uma menor quantidade de falsos negativos, positivos e neutros, conforme a Figura 13.

Figura 13. Matriz de confusão dos sentimentos

Predito	Negativo	Neutro	Positivo	All
Real				
Negativo	219	19	0	238
Neutro	19	217	2	238
Positivo	1	23	214	238
All	239	259	216	714

Fonte: Elaborado pelo Autor.

O mesmo comportamento aconteceu com as emoções, com uma única observação de que esta foi a base que mais sofreu diminuição de dados devido ao balanceamento, conforme exibe a Figura 14.

Figura 14. Matriz de confusão das emoções

Predito	felicidade	raiva	tristeza	All
Real				
felicidade	23	7	1	31
raiva	2	29	0	31
tristeza	0	2	29	31
All	25	38	30	93

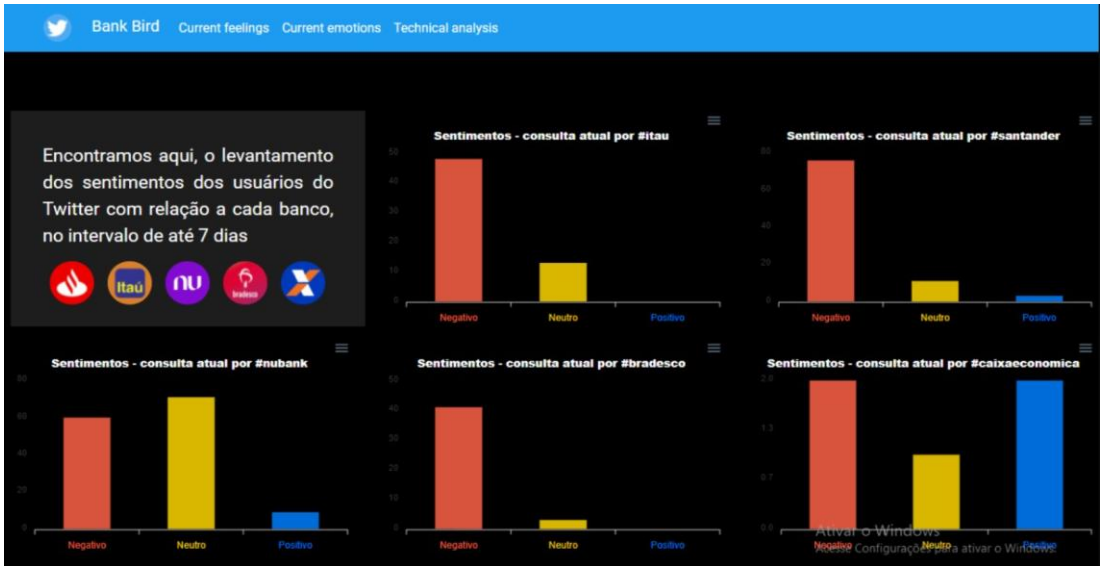
Fonte: Elaborado pelo Autor.

Para o desenvolvimento do trabalho, optou-se pela utilização da base própria para as emoções, já para os sentimentos, utilizou-se a base do governo de Minas Gerais.

5.3. Ferramenta desenvolvida

A ferramenta foi desenvolvida com 3 telas, sendo elas "*Current feelings*", "*Current Emotions*" e "*Technical Analysis*". Logo abaixo, é retratada a tela de "*Current feelings*", que, como o próprio nome ja diz, realiza uma busca dinâmica dos *tweets* com as *hashtags* dos bancos definidos dos últimos 7 dias e expressa o resultado da mineração exibindo os sentimentos de cada postagem em um gráfico para cada instituição bancária. Os gráficos expressam a quantidades de *tweets* por cada um dos sentimentos, negativo, neutro e positivo, conforme ilustra a Figura 15.

Figura 15. Tela "*Current feelings*"



Fonte: Elaborado pelo Autor.

As emoções atuais, sendo elas contidas em todos os *tweets* dos últimos 7 dias, são expressas na tela de "*Current emotions*", que possui sua divisão de barras baseadas em três emoções, sendo elas: raiva, tristeza e felicidade. Um gráfico foi separado para cada banco, permitindo assim a comparação entre eles, conforme mostra a Figura 16.

Figura 16. Tela "*Current emotions*"

Fonte: Elaborado pelo Autor.

Para uma análise mais técnica foi desenvolvida a tela de "*Technical analysis*", na qual apresenta todo o levantamento da base de dados usada para a mineração, contendo todos os *tweets* coletados relacionados aos bancos apresentados. Nela nota-se a exibição dos valores de *accuracy*, f1 e *recall* para a base de sentimentos e a de emoções, estes, obtidos através da base teste, comparada com os resultados da predição realizada pelos algoritmos de ML, expressos no gráfico de barras. O outro gráfico, ilustra a matriz de confusão gerada por cada uma delas, conforme apresenta a Figura 17.

Figura 17. Tela "*Technical analysis*"

Fonte: Elaborado pelo Autor.

6. Considerações Finais

Com o desenvolvimento deste trabalho, foi possível obter conhecimento a respeito de técnicas de pré-processamento, PLN e de ML coletados em textos. Tais técnicas aplicadas ao longo do desenvolvimento da aplicação trouxeram resultados significativos como a rotulagem dos sentimentos e emoções removendo palavras sem sentido semântico e classificando as palavras de acordo com a base de comparação utilizada.

Foi possível criar uma ferramenta que demonstre esses resultados em gráficos de fácil entendimento. Os resultados foram expressos aos usuários representando os sentimentos positivo, negativo e neutro e as emoções raiva, tristeza e felicidade. Estas classificações foram as mais assertivas obtidas.

Considera-se também que este tratamento de dados é de suma importância quando se trata de uma análise instantânea e atual de como o público avalia as instituições bancárias estabelecidas. Estratégias de negócio podem ser tomadas a partir desta análise, visto que o conteúdo utilizado é a própria opinião do usuário.

Visando a continuidade e o aperfeiçoamento deste projeto, é proposto:

- Realizar novas coletas para complementar a base de dados e consequentemente os resultados dos algoritmos melhorem.

- Encontrar novas estratégias e algoritmos para o balanceamento da base de dados como o *Cost-Sensitive Algorithms.*
- Trabalhar a detecção de textos criados por *bots* e conseguir lidar com estes para evitar a poluição dos resultados.

7. Referências

ARANHA, Christian; PASSOS, Emmanuel. A tecnologia de mineração de textos. Revista Eletrônica de Sistemas de Informação, v. 5, n. 2, 2006.

BARBOSA, Luciano; FENG, Junlan. Robust sentiment detection on Twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics, USA, p: 36-44

BARBOSA, Gian; MIRANDA, Péricles; MELO, Rafael; SILVA, Ricardo. Sequencing Sampling Algorithms to Boost Performance of Classifiers on Imbalanced Data Sets. Universidade Do Pernambuco, 2019. Disponível em: https://sol.sbc.org.br/index.php/eniac/article/view/9302. Acesso em: 04 nov. 2021.

BECKER, Karin; TUMITAN, Diego. Introdução à mineração de opiniões: Conceitos, aplicações e desafios. Simpósio brasileiro de banco de dados, v. 75, 2013.

CABRAL, Mayara Kaynne Fragoso; BRUNO, João Carlos; CORADO, Vanessa Aires. Mineração de dados do Twitter através do R. In: 6ª JICE-Jornada de iniciação científica e extenção, 2015.

CAPOBIANCO, KELVIN RAMIRES. Avaliação da etapa de pré-processamento na Mineração de Texto em Redes Sociais Digitais. Universidade Estadual de Londrina, Londrina–PR, 2016.

CARVALHO FILHO, José Adail. Mineração de textos: análise de sentimentos utilizando Tweets referentes à Copa do Mundo 2014. 2014.

IBM. In: Entendendo a Linguagem Natural do Watson. 1.0. [S. l.], 1 jun. 2021. Disponível em: https://www.ibm.com/br-pt/cloud/watson-natural-language-understanding/details. Acesso em: 13 jun. 2021.

GUEDES, Gustavo Paiva et al. MAM: Método para Agrupamentos Múltiplos em Redes Sociais Online Baseado em Emoções, Personalidades e Textos. iSys-Revista Brasileira de Sistemas de Informação, v. 7, n. 3, p. 38-55, 2014.

LIU, Bing. Sentiment Analysis and Subjectivity. In: INDURKHYA, Nitin; DAMERAU, Fred J.. Handbook of natural language processing. 2.ed. Boca Raton: Crc Press, 2010. Cap. 26. p. 627-666.

LIU, Qian; ZHOU, Bing; LIU, Qingzhong. Can twitter posts predict stock behavior?: A study of stock market with twitter social emotion. In: 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA). IEEE, 2016. p. 359-364.

PANG, B.; LEE, L. Opinion Mining and sentiment analysis. Foundations and Trends in Information Retrieval, v. 2, n. 1-2, 2008, p. 1-135.

PARIKH, Ravi; MOVASSATE, Matin. Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques, Stanford University.

RANI, Sujata; KUMAR, Parteek. A sentiment analysis system to improve teaching and learning. Computer, v. 50, n. 5, p. 36-43, 2017.

CABREIRA, Jhonatan Moura. In: Repositório para armazenamento de código e notebooks de postagens do blog e cursos. Minas Gerais: Jhonatan Moura Cabreira, 23 out. 2018. Disponível em: https://github.com/stacktecnologias/stack-repo). Acesso em: 27 set. 2021.

SANTOS, Gustavo Campos. Mineração de texto baseado em grafos para identificação de conteúdos, 2018.

SERRANO-GUERRERO, Jesus et al. Sentiment analysis: A review and comparative analysis of web services. Information Sciences, [s.l.], v. 311, p.18-38, Aug. 2015. Elsevier BV. Disponível em: http://dx.doi.org/10.1016/j.ins.2015.03.040. Acesso em 27 set. 2021.

THAYNÁ, O.; ROSSI, Rafael G.; DO SUL–MS–BRAZIL, Mato Grosso. Desenvolvimento de uma Ferramenta para Análise de Sentimentos de Textos Publicados no Twitter, 2017.

TURNEY, Peter D. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, USA, p: 417-424

VILLARROEL, Rosivaldo Gabriel. O uso da análise de sentimentos como ferramenta de apoio à gestão acadêmica. 2020. Dissertação (Mestrado em Engenharia de Computação) - Instituto de Pesquisas Tecnológicas do Estado de São Paulo, [S. l.], 2020

WORDNET. Apresenta informações a respeito de uma base de dados léxica disponível em: https://wordnet.princeton.edu/. Acesso em: 21 de abr de 2021.