Negatividade no Twitter

Uma análise de sentimentos no twitter americano

Guilherme Jácome Cavalcante

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa, 2021

Guilherme Jácome Cavalcante

Negatividade no Twitter

Relatório apresentado ao curso de engenharia da computação do Centro de Informática, da Universidade Federal da Paraíba, como avaliação para a disciplina de Metodologia do Trabalho Científico

Orientador: Thaís Gaudencio do Rêgo

Dezembro de 2021

RESUMO

O presente trabalho tem como objetivo analisar *tweets* em inglês, utilizando um modelo de análise de sentimentos, para observar se, de fato, como o senso comum nos diz, o twitter é uma rede social com tendências negativas, que se torna um fato preocupante, visto que se expor a ambientes com essa carga pode afetar o humor dos usuários. Para treinar o modelo de classificação *CatBoost*, utilizado na tarefa de análise de sentimentos, utilizamos dados obtidos através do dataset *Emotions dataset for NLP* do website *kaggle* e, com o modelo treinado, obtivemos dados do twitter através da ferramenta *Twarc*. Tivemos como conclusão que, de fato, o twitter, no contexto americano, é uma rede social mais negativa, com indíces de 44,62% superiores à positividade.

Palavras-chave: Negatividade, Twitter, Análise de sentimentos, NLP.

LISTA DE FIGURAS

LISTA DE TABELAS

Sumário

1 INTRODUÇÃO

Criada na década de 90, como um meio de comunicação entre universitários, as redes sociais são sites e aplicativos que permitem a interação e troca de informações entre usuários, muitas vezes baseadas em interesses em comum. Essas plataformas ganharam muita força desde o seu surgimento, ganhando bilhões de usuários e se alastrando por todas as partes do mundo.

Uma das redes sociais com mais usuários atualmente é o twitter, com 1.3 bilhão de contas criadas até 2020, segundo o website *websiterating*. Nessa rede, as pessoas costumam compartilhar fatos da vida, publicar opiniões, acessar *memes* e entre outras coisas, sempre se expondo ou sendo expostas a diversos conteúdos.

1.1 Definição do Problema

Dado esse grande número de usuários, onde cerca de 22% são americanos, segundo o website *websiterating*, e essa grande exposição a conteúdos de terceiros, faz-se necessário um estudo sobre as tendências do ambiente, visando identificar se ele é ou não predominantemente negativo, visto que as emoções são transmitidas como uma doença contagiosa (*HILL, Alisson, L. et al, 2010*) (*FOULK, Trevor, et al, 2016*), podendo intensificar sentimentos de tristeza, medo e ódio na sociedade.

1.2 Premissas e Hipóteses

Dessa forma, através do treinamento de um modelo de análise de sentimentos e da sua utilização em uma vasta base de dados, contendo tweets e outros metadados, é possível verificar a hipótese de que o twitter, no contexto americano, é uma rede social com tendência negativa, podendo contagiar outros usuários com tal sentimento.

1.3 Objetivo geral

O trabalho tem como objetivo avaliar o comportamento dos tweets no contexto americano, afim de descobrir se existe alguma tendência negativa na rede social.

1.4 Objetivos específicos

Criar modelo de análise de sentimentos utilizando o CatBoostClassifier, para prever tweets positivos e negativos.

Análisar tweets atuais, prever seu sentimento com o modelo treinado, e então, verificar possível tendência negativa na rede social.

2 CONCEITOS GERAIS E REVISÃO DA LITERATURA

2.1 Modelo de classificação

Classificação é uma subárea da aprendizagem de maquina; esse ramo visa prever a categoria de determinado exemplar a partir de suas carecterísticas e observações passadas, podendo ser binária (duas classes), ou multiclasse (mais de duas classes). Existem diversas técnicas diferentes para problemas de classificação, tais como árvores de decisão, Bayes ingênuo e K-vizinhos mais próximos. Alguns exemplos de tarefas que podem ser realizados com eles é classificar se uma imagem corresponde a um cachorro ou um gato, ou se uma frase é positiva ou negativa.

2.2 CatBoost

CatBoost é um algoritmo de aprendizagem de máquina de código aberto, criado pela empresa *Yandex*, que utiliza aumento de gradiente em árvores de decisão.

Aumento de gradiente é uma técnica na aprendizagem de máquina para problemas de regressão e classificação, geralmente utilizando árvores de decisão. Essa técnica combina vários preditores que sozinhos possuem uma baixa acurácia, para produzir um modelo preditivo com melhores resultados. Essa combinação se faz de modo que cada preditor possui o objetivo de minimizar o erro do modelo anterior, utilizando a função de perda.

Esse algoritmo se mostra muito eficiente e vantajoso por diversos motivos, entre eles: não necessita de ajustes de parâmetros para funcionar bem, diminuindo o tempo de procura dos melhores valores; não necessita de um grande tratamento dos dados, funcionando inclusive com dados categóricos como textos; velocidade de treinamento, superando modelos de *boosting* muito conhecidos como o *XGBoost*, *H20* e até *LightGBM*, que já era previamente conhecida pela sua velocidade; e ótimos resultados, mesmo quando treinado com uma pequena quantidade de dados.

3 METODOLOGIA

A metodologia adotada no trabalho pode ser separada em quatro partes principais, são elas: Obtenção e tratamento da base de dados de treino, treinamento do modelo de análise de sentimentos, obtenção e tratamento da base de dados de análise e análise dos dados. A *pipeline* de organização do processo pode ser visto na Figura 1.

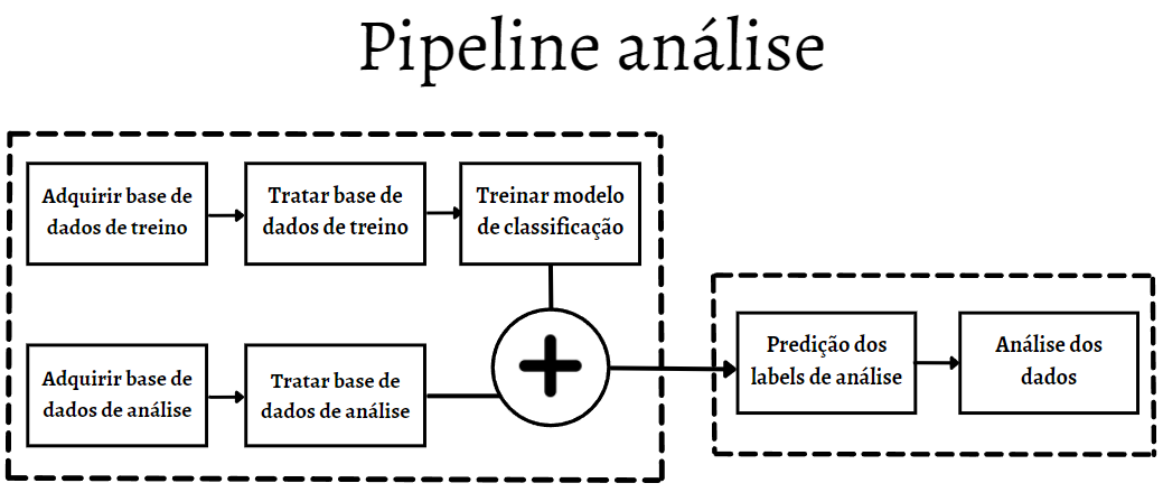

Figura 1: Pipeline de análise

3.1 Dados de treino

Para treimar o modelo, foi utilizado o *dataset Emotions dataset for NLP*, do website *Kaggle*. O *dataset* contêm duas colunas, uma contendo o texto, e outro contendo o *label*, que pode assumir os valores: *sadness*, *anger*, *fear*, *joy*, *surprise* e *love*. Para se adequar melhor ao propósito do trabalho, os *labels* foram ajustados, de forma que *sadness*, *anger* e *fear* passaram a ter o valor 0 e *joy*, *surprise* e *love* foram trocados por 1. Um exemplo do dataset pode ser visto na Figura 2.

	text	label
0	i can go from feeling so hopeless to so damned...	0
1	im grabbing a minute to post i feel greedy wrong	0
2	i am ever feeling nostalgic about the fireplac...	1
3	i am feeling grouchy	0
4	ive been feeling a little burdened lately wasn...	0

Figura 2: Exemplo do dataset de treino

Para que fosse possível o treinamento do modelo, os dados de texto sofreram bastante tratamento. Todos os caracteres especiais foram retirados e as letras foram transformadas em minúsculas. Além disso, foram retiradas as *stopwords* (palavras consideradas irrelevantes para a construção da frase), visto que não adicionam nenhum sentido a mais aos textos. Por fim, as frases passaram pelo processo de stemização, flexionando-as ao seu tronco, diminuindo assim, o tamanho do vocabulário e consequentemente, facilitando o treinamento do modelo.

3.2 Treinamento do modelo

Para possibilitar a utilização do texto no treinamento, utilizamos o processo de vetorização *CountVectorizer*, que transforma o texto em grandes vetores, onde cada elemento corresponde a uma palavra e seu valor corresponde à quantidade de vezes que ela apareceu no texto. Feito isso, separamos o grupo de treino e o grupo de teste, nas proporções 0.8 e 0.2. Por fim, como as classes não estavam balanceadas, utilizamos o processo de *undersampling*, retirando exemplares aleatórios da classe dominante, para garantir que os *labels* possuiam a mesma quantidade de representantes, e que não haveriam tendências na predição.

Para treinar o modelo de análise de sentimentos, utilizamos o algoritmo *CatBoost* e seu modelo de classificação *CatBoostClassifier*, que foi escolhido por todas suas vantagens apresentadas na seção 2.2. O treinamento foi realizado com todos os parâmetros padrões do modelo.

Como resultado, após 10.2 segundos de treino no *google colaboratory* (sem utilizar aceleradores de hardware), obtivemos uma acurácia de 80%, e geramos a matriz de confusão a seguir:

$$\mathbf{MC} = \begin{pmatrix} 1571 & 444 \\ 274 & 1316 \end{pmatrix}$$

3.3 Dados de análise

Para adquirir os dados de análise, foi utilizado a ferramenta *twarc*, que possibilita fazer requisições na API do twitter, para então adquirir postagens dos ultimos dias, atrelados aos seus respectivos metadados. As requisições foram feitas de forma aleatória, garantindo uma representatividade real do ambiente.

O *dataset* adquirido possuia 42.000 linhas e 74 colunas, que após a eliminação das colunas desnecessárias e a retirada de todos os textos que não estavam em inglês, foi reduzida para 16.529 linhas e 10 colunas. As colunas restantes com suas respectivas definições podem ser encontradas na tabela 1.

Coluna	Definição
text	Conteúdo do tweet
lang	Idioma do tweet
public_metrics.like_count	Número de likes no tweet
public_metrics.reply_count	Número de respostas no tweet
public_metrics.retweet_count	Número de compartilhamentos do tweet
author.created_at	Data da criação do usuário
author.public_metrics.followers_count	Número de seguidores
author.public_metrics.following_count	Número de usuários seguidos
author.public_metrics.tweet_count	Número de tweets do usuário
author.verified	Conta verificada no twitter

Tabela 1: Colunas do dataset de avaliação

Com todos os processos anteriores completos, foi criado uma nova coluna no *dataset* de análise, correspondendo ao *label* previsto pelo modelo de acordo com o texto de sua respectiva linha (podendo assumir os valores de 0, quando negativa, e 1, quando positiva), possibilitando então, a análise final dos resultados.

3.4 Ferramentas

Para o projeto, foi utilizado a linguagem de programação *Python*, por possuir diversas bibliotecas que facilitam na hora de trabalhar com *datasets*, análise de dados e treinamento de modelos de inteligência artificial.

Com relação às bibliotecas, foi utilizado o *pandas*, para trabalhar com *datatasets*; *nltk*, para pré-processamento do texto; *sklearn*, para vetorização do texto e separação em treino e teste; *catboost*, para treinamento do modelo; e *matplotlib* para visualização dos resultados.

Além disso, foi utilizado o ambiente do *google colaboratory* para executar todo o código, visto que é mais rápido, disponibilzando GPUs gratuitas e também por já possui uma grande variedade de bibliotecas pré-instaladas.

4 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

4.1 Análise geral dos dados

Para que seja feita a análise final da negatividade no ambiente do twitter, devemos antes entender um pouco mais sobre as particularidades que regem os textos com traços positivos e negativos, afim de verificar quais características os diferem.

Dessa forma, foram selecionados exemplos aleatórios de textos negativos e positivos presentes na base de dados, que podem ser observados na Tabela 2.

Label	**Texto**
Negativo	Y'all taking away the best rappers man fuck y'all
Negativo	Cranberry sauce belong in the trash.
Negativo	R. I.P. to my friend Dolph this broke my heart
Negativo	This is so disappointing!
Positivo	Taehyung skipping through the airport, pls he's so excited
Positivo	okay this isnt another weird vague tweet its just my birthday
Positivo	@MangaMoguraRE I want to read the first two, the covers look interesting.
Positivo	Why people brag about taking care of their kids? You're literally a parent

Tabela 2: Exemplos do dataset com seus labels

Quando separamos os *datasets* com base nos *labels*, conseguimos observar algumas das palavras mais frequentes dos *tweets* positivos e negativos, como pode ser visualizado nas Figuras 3 e 4. Como era esperado, é possível visualizar uma maior quantidade de palavras que remetem a coisas boas nos textos positivos, tais como: *love*, *need*, *better*, *enjoy*, entre outras. Já nos textos negativos, também conseguimos verificar o surgimento de mais frases que remetem a xingamentos, tais como: *fuck*, *shit*, *bitch*, *die*, entre outras.

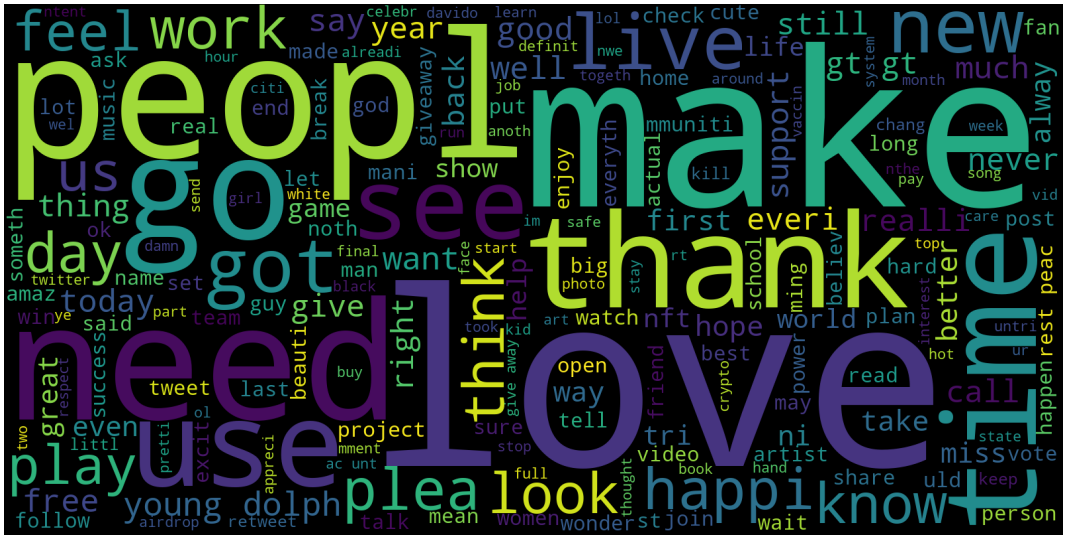

Figura 3: Palavras mais usadas em frases positivas

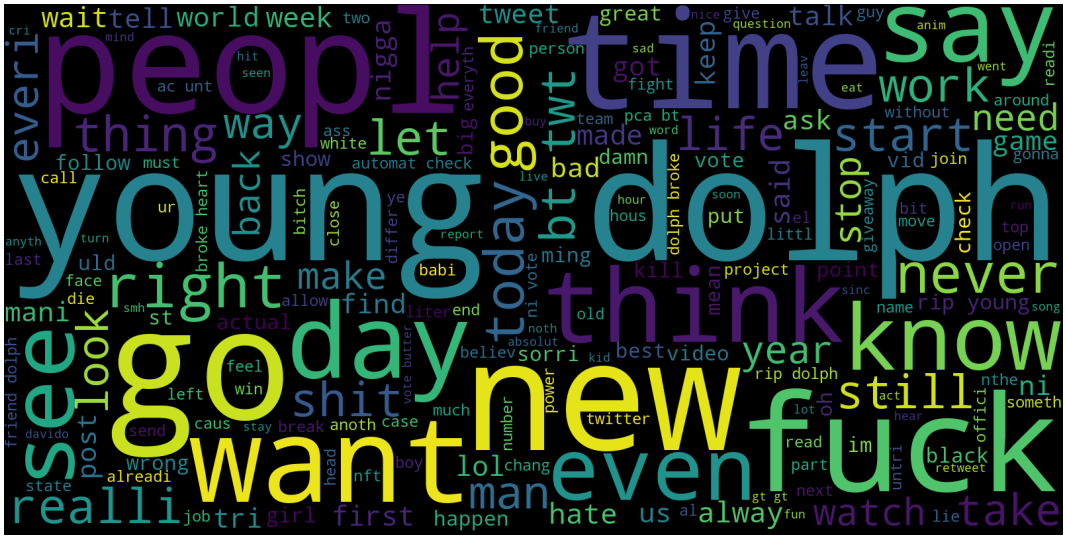

Figura 4: Palavras mais usadas em frases negativas

4.2 Negatividade no Twitter

Para que seja possível averiguar a existência de tendências negativas no *twitter*, é necessário um estudo sobre a distribuição dos *labels* positivos e negativos no *dataset* de análise, afim de observar se, de fato, existem mais textos com configurações negativas na rede.

Ao observar a Figura 5, conseguimos verificar que, de fato, existe uma predominância negativa na rede, com indíces negativos 44,62% superiores aos positivos.

Figura 5: Distribuição dos labels no dataset

Além disso, ao verificar as médias dos *labels* baseadas no ano da criação do usuário, embora não seja possível verificar uma relação entre a idade da conta com os índices de negatividade, podemos verificar que as médias são sempre mais próximas de 0 do que de 1, mostrando que em geral, independente de se o usuário adentrou recentemente na rede social ou não, ela possui uma tendência a ter comportamentos negativos dentro do ambiente.

Figura 6: Média do label por ano de criação do usuário

5 CONCLUSÕES E TRABALHOS FUTUROS

A partir dos resultados obtidos, pode se observar que, de fato, como proposto na hipótese do trabalho, o *twitter*, no contexto americano, é uma rede social com tendência à negatividade, se tornando um fator preocupante para a sociedade, visto que pode contaminá-la, trazendo diversos outros problemas.

Por fim, vale salientar algumas limitações percebidas ao longo do projeto. Primeiramente, pela escassêz de tempo, o modelo criado para a tarefa de análise de sentimentos ainda é rudimentar, podendo ser melhorado com refinamento dos parâmetros de treino e com o aumento da base de dados utilizada no processo. Além disso, pela ausência de uma conta acadêmica na API oficial do *twitter*, se tornou inviável a análise da negatividade com base nos anos, abrindo margem para trabalhos futuros que visam o estudo da progressão das emoções na rede social.

REFERÊNCIAS

[1] HILL, Alison L et al. **"Emotions as infectious diseases in a large social network: the SISa model."** Proceedings. Biological sciences vol. 277,1701 Dec. 2010.

[2] FOULK, Trevor et al. **"Catching rudeness is like catching a cold: The contagion effects of low-intensity negative behaviors."** The Journal of applied psychology vol. 101,1 Jan. 2016.

[3] Adami, Anna. **Redes Sociais**. InfoEscola, 2008. Disponível em: <https://www.infoescola.com/sociedade/redes-sociais-2/>. Acesso em: 23. Nov. 2021.

[4] AHLGREN, Matt. **Mais de 50 estatísticas e fatos do Twitter**. Websiterating, 2021. Disponível em: <https://www.websiterating.com/pt/research/twitter-statistics>. Acesso em: 23. Nov. 2021.

[5] SILVA, Jonhy **Uma breve introdução ao algoritmo de Machine Learning Gradient Boosting utilizando a biblioteca Scikit-Learn**. Medium, 2020. Disponível em: <https://medium.com/equals-lab/uma-breve-introdu%C3%A7%C3%A3o-ao-algoritmo-de-machine-learning-gradient-boosting-utilizando-a-biblioteca-311285783099>. Acesso em: 30. Nov. 2021.

[6] Praven, **Emotions dataset for NLP**. Kaggle, 2021. Disponível em: <https://www.kaggle.com/praveengovi/emotions-dataset-for-nlp>. Acesso em: 15. Nov. 2021.

[7] Edsu, **Twarc**. Twarc, 2021. Disponível em: <https://twarc-project.readthedocs.io/en/latest/>. Acesso em: 20. Nov. 2021.