Análise de sentimentos sobre a Copa do Mundo 2022

Ezequiel S. D. Santos

22 de setembro de 2022

Abstract

Taking into account that social networks offer a quick and easy-to-use means of communication, the world was driven to turn to these social structures. Thus, we have an increasing amount of data that is generated every second. This information, given its importance, is used in the economy, security, health, education, etc. This study focuses on applying Text Mining and NLP techniques to observe the feelings shown by Tweeter users about the World Cup 2022. A collection of 21,866 tweets was made. This collection took place between September 01 and 13, 2022. According to the analysis, it was found that the highlighted classification was positive.

Resumo

Levando em consideração que as redes sociais oferecem um meio de comunicação rápido e de fácil usabilidade, o mundo foi impulsionado a se voltar para estas estruturas sociais. Assim, temos uma crescente quantidade de dados que são gerados a cada segundo. Essas informações, dada a sua importância, são usadas na economia, segurança, saúde, educação, etc. Este estudo foca em aplicar técnicas de Text Mining e NLP para observar os sentimentos evidenciado pelos usuários do Tweeter sobre a Copa do Mundo 2022. Foi feita uma coleta de 21.866 tweets. Essa coleta foi entre os dias 01 a 13 de Setembro de 2022. Conforme a análise, constatou-se que a classificação destacada foi positiva.

1 Introdução

O compartilhamento de dados foi revolucionado com a tecnologia Web 2.0. Essa mudança deu mais autonomia aos usuários para criarem conteúdo. A era da informação trouxe grandes benefícios para o nosso cotidiano. Como consequência, estamos vendo um número crescente de informações compartilhadas (como textos, áudios e vídeos) [3, 5]. Sendo que cerca de 80% dessa informação é não estruturada. Com isso, as grandes empresas e governos começaram a usar as redes sociais (fruto da web 2.0) para receber feedback das pessoas [2]. As técnicas mais comuns usadas são: Text Mining (do português: Mineração de Texto) e NLP - Natural Language Processing (do português: Processamento de Linguagem Natural).

A NLP é responsável por facilitar as interações entre os computadores e a linguagem humana natural. Isso significa que ela executa a tarefa de facilitar que os computadores entendam a linguagem humana natural para desenvolver diferentes tipos de tarefas e aplicações. A NLP está em áreas como ciência de dados, inteligência artificial, linguística, filosofia, entre várias outras áreas. Alguns exemplos de funcionalidade da NLP são: tradução de idiomas e sumarização automática. Além disso, suas principais ferramentas são: normalização, remoção de stopwords e tokenização [9].

Text Mining é o processo para extrair informações relevantes de um texto (ou seja, ela não ajuda a interpretar as informações da estrutura gramatical ou semântica de um texto). Suas aplicações são em análise de dados de rede social, prevenção de crimes cibernéticos, entre outras

[1, 6].

A Copa do Mundo, criada em 1928 na França, é o evento internacional (esportivo) de futebol masculino mais conhecido do mundo, que é organizado pela Federação Internacional de Futebol (FIFA - Federação Internacional de Futebol Associado). Ocorrendo a cada quatro anos, ela reúne 32 seleções do mundo. De 20 de novembro a 18 de dezembro de 2022, ocorrerá a vigésima segunda edição, que se passará no Catar. Por ser um evento muito lucrativo, as grandes empresas investem bastante dinheiro para entender as opiniões e gostos do público em geral para desenvolver produtos mais atraentes [4, 8, 7].

Este artigo tem como objetivo aplicar técnicas de Text Mining e NLP para analisar os sentimentos dos brasileiros em relação à Copa do Mundo 2022.

2 Metodologia

Veremos a seguir técnicas de Text Mining sobre os brasileiros em relação à Copa do Mundo 2022, onde os dados coletados foram a partir do Twitter. Sendo assim, através da classificação de sentimentos, entenderemos os tipos de sentimentos mais frequentes.

2.1 Coleta de Dados

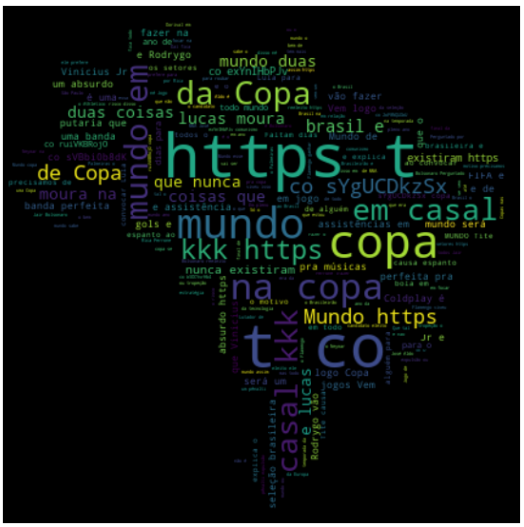
Este artigo foi possível através da coleta de dados do Twitter, pois é uma rede em que a maior parte da população usa para expor suas opiniões. Como de interesse, utilizamos a API (Application Programming Interface) pública do Twitter, que foi disponibilizado para fazer a coleta de dados textuais. A coleta desses dados foi feita a partir do dia 01 de Setembro a 13 de Setembro de 2022. Isso resultou em 21.866 tweets coletados, totalizando 13 dias de coleta. A Tabela 1 mostra as palavras-chave usadas neste contexto.

Tabela 1: Palavras usadas para pesquisa.

Palavras-chave
#COPACATAR202
COPADOMUNDO2022
copadomundo2022
COPA DO MUNDO
CATAR2022
copadomundocatar2022

2.2 Pré-Processamento de Dados

Esta etapa de pré-processamento possui o objetivo de limpar qualquer informação irrelevante para os dados e, após isso, preparar esses dados para a etapa de classificação. As técnicas utilizadas aqui foram: remoção de letras repetidas, remoção de stopwords, tokenização, lematização, word embeddings, remoção de links, remoção de caracteres irrelevantes, padronização de letras maiúsculas em minúsculas.



(a) Antes do pré-processamento.

(b) Depois do pré-processamento.

Figura 1: Nuvem de palavras.

2.3 Modelo

O modelo adotado foi através da biblioteca MLlib do Spark, e para a validação do modelo foi utilizado a interface pyspark, que contém um avaliador para classificação multi-classe, do Python [9]. Para o conjunto de dados para treinamento e testes, os tweets foram classificados por duas polaridades. Estas polaridades são: Positivo e Negativo.

2.4 Classificação e validação do modelo

Foi tomado como método de análise a Regressão Logística, sendo este um modelo de aprendizagem probabilística para classificação. Foram usados 4.600 registros para o modelo, sendo 20% desses registros usados para teste. Podemos ver essa divisão na Tabela 2.

Tabela 2: Divisão dos dados.

Sentimento	**Total**	**Porcentagem %**
Positivo	2.300	50%
Negativo	2.300	50%
Total	4.600	100%

Como avaliação do modelo, tomamos como métrica:

- A matriz de confusão para identificar a qualidade do modelo.

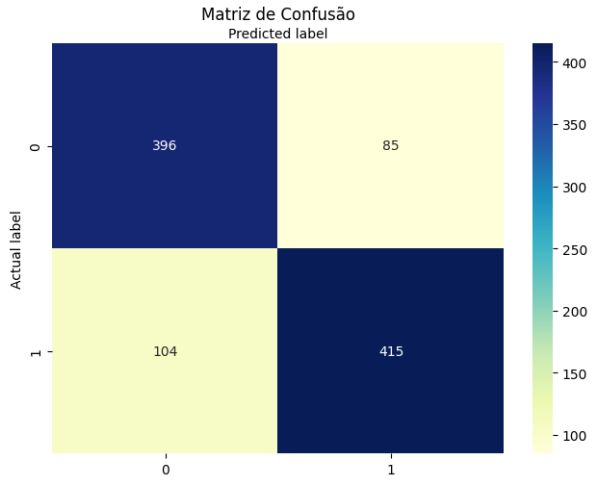


Figura 2: Matriz de Confusão

- A acurácia que identifica o percentual de acertos;
- A precisão que possui o objetivo verificar a quantidade de acertos dentre aqueles que foram previstos como positivo;
- O recall que tem a função de calcular a quantidade de acertos dentre os positivos reais;
- O f1-score que é um balanceamento entre a precisão e o recall.

Tabela 3: Desempenho do modelo.

Acurácia	**Precisão**	**Recall**	**F1-score**
81%	82,29%	79%	80,99%

Pode-se notar, neste contexto, que o modelo obteve um bom desempenho, dada a sua alta acurácia, precisão e recall.

3 Resultados

A figura (3) mostra os sentimentos sobre a copa do mundo 2022, onde foram analisados 21.866 tweets. É notável que a maior parte dos torcedores brasileiros estão otimistas para a Copa do Mundo 2022, porém existe uma parcela, não desprezível, sem otimismo.

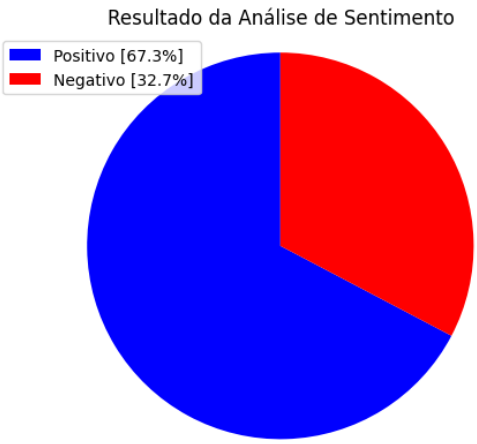


Figura 3: Análise de sentimento.

Essas polaridades foram influenciadas pelo ano eleitoral em que o brasil se encontra.

4 Conclusão

Foram abordados neste trabalho métodos de Text Mining e NLP. Para efeito, aplicamos estes métodos com o intuito de analisar tweets de brasileiros em relação à Copa do Mundo.

A primeira etapa foi realizada a coleta e o pré-processamento dos tweets. Na segunda etapa realizamos a classificação e análise dos tweets, cujo modelo usado foi a Regressão Logística. Esse classificador obteve um resultado satisfatório para a classificação de texto, apesar de algumas limitações encontradas.

Como consequência, este estudo pode ser aplicado em vários outros contextos para analisar dados não estruturados. Para trabalhos futuros, tem-se como objetivo incluir a classificação da polaridade neutra, visando melhorar o desempenho do modelo. E aplicar o estudo para prever os sentimentos dos investidores no mercado de ações.

Referências

[1] Jonathan Benchimol, Sophia Kazinnik, and Yossi Saadon. Text mining methodologies with r: An application to central bank texts. *Machine Learning with Applications*, 8:100286, 2022.

[2] João Paulo Ciribeli and Victor Hugo Pereira Paiva. Redes e mídias sociais na internet: realidades e perspectivas de um mundo conectado. *Revista Mediação*, 2011.

[3] Sanjiv Ranjan Das. Data science: theories, models, algorithms, and analytics. *Learning*, 143:145, 2016.

[4] futeboleiro.com. Copa do Mundo Catar 2022: Calendario, equipes e jogadores. `https://www.futeboleiro.com/copa-do-mundo-2022/#:~:text=O%20jogo%20de%20abertura%20da,9%20e%2010%20de%20dezembro.`

[5] Karwan Jacksi and Shakir M Abass. Development history of the world wide web. *Int. J. Sci. Technol. Res*, 8(9):75–79, 2019.

[6] Jão Junior. Etapas da metodologia de mineração de textos. 2008.

[7] Murat Kucukvar, Adeeb A Kutty, Abathar Al-Hamrani, Doyoon Kim, Nadejhda Nofal, Nuri Cihat Onat, Polina Ermolaeva, Tareq Al-Ansari, Soud Khalifa Al-Thani, Nasser Mohammed Al-Jurf, et al. How circular design can contribute to social sustainability and legacy of the fifa world cup qatar 2022™? the case of innovative shipping container stadium. *Environmental Impact Assessment Review*, 91:106665, 2021.

[8] Stefan Szymanski. The economic impact of the world cup. In *Football Economics and Policy*, pages 226–235. Springer, 2010.

[9] Alex Thomas. *Natural Language Processing with Spark NLP: Learning to Understand Text at Scale*. O'Reilly Media, 2020.