UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

IGOR DE SOUSA PEREIRA

ANÁLISE DE TWEETS EM GRANDES EVENTOS: UM ESTUDO DE CASO DA COPA DO MUNDO NO TWITTER

CAMPINA GRANDE - PB

2023

IGOR DE SOUSA PEREIRA

ANÁLISE DE TWEETS EM GRANDES EVENTOS: UM ESTUDO DE CASO DA COPA DO MUNDO NO TWITTER

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador : Professor Dr. Claudio Elízio Calazans Campelo.

CAMPINA GRANDE - PB

2023

IGOR DE SOUSA PEREIRA

ANÁLISE DE TWEETS EM GRANDES EVENTOS: UM ESTUDO DE CASO DA COPA DO MUNDO NO TWITTER

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professor Dr. Claudio Elízio Calazans Campelo
Orientador – UASC/CEEI/UFCG

Professor Dr. Carlos Eduardo Santos Pires
Examinador – UASC/CEEI/UFCG

Professor Dr. Francisco Vilar Brasileiro
Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 28 de JUNHO de 2023.

CAMPINA GRANDE - PB

RESUMO

A internet tem possibilitado uma maior interação entre as pessoas ao redor do mundo, e o Twitter tem desempenhado um papel significativo nisso. Como uma das redes sociais mais utilizadas do planeta, o Twitter permite que os usuários façam postagens expressando diversos aspectos do seu dia a dia. Com a aproximação de grandes eventos de reconhecimento mundial(e.g., eventos esportivos), é natural que os usuários também expressem seus comentários sobre os diversos aspectos dessas competições. No entanto, devido à enorme quantidade de comentários gerados diariamente, é possível observar uma variedade de interpretações que refletem diferentes aspectos do evento, como discussões sobre escolha dos técnicos, desempenho das equipes e até mesmo comentários pessoais sobre os participantes. Este artigo apresenta uma abordagem para analisar as discussões em torno de um evento, utilizando a Copa do Mundo como exemplo para validar o método. Para isso, foram empregados dados coletados do Twitter, os quais foram usados como entrada em técnicas de agrupamento, a fim de identificar potenciais conjuntos de comentários relacionados à competição. Como resultado, foram obtidos 13 grupos diversos, cada um com características únicas, abrangendo desde avaliações individuais, até críticas ao país-sede Qatar, rivalidades entre seleções, entre outros aspectos. Esses resultados indicam a existência de padrões nos comentários sobre um tema específico, sugerindo que os usuários buscam comentar temas do momento e com grande engajamento.

ANALYSIS OF TWEETS ON MAJOR EVENTS: A CASE STUDY OF THE WORLD CUP ON TWITTER

ABSTRACT

The Internet has enabled greater interaction between people around the world, and Twitter has played a significant role in this. As one of the most widely used social networks on the planet, Twitter allows users to make posts expressing various aspects of their daily lives. With the approach of major events of worldwide recognition (e.g. sporting events), it is natural that users also express their comments on the various aspects of these competitions. However, due to the huge amount of comments generated daily, it is possible to observe a variety of interpretations that reflect different aspects of the event, such as discussions about coaches' choices, teams' performance, and even personal comments about the participants. This paper presents an approach to analyze the discussions surrounding an event, using the World Cup as an example to validate the method. For this purpose, data collected from Twitter was employed, which was used as input in clustering techniques in order to identify potential sets of comments related to the competition. As a result, 13 diverse clusters were obtained, each with unique characteristics, ranging from individual assessments, to criticism of the host country Qatar, to team rivalries, among other aspects. These results indicate the existence of patterns in comments on a specific theme, suggesting that users seek to comment on topics of the moment and with high engagement.

Analise de *Tweets* em grandes eventos: Um Estudo de Caso da Copa do Mundo no Twitter

Trabalho de Conclusão de Curso

Igor de Sousa Pereira (Aluno), Cláudio Campelo (Orientador)
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba - Brasil

RESUMO

A internet tem possibilitado uma maior interação entre as pessoas ao redor do mundo, e o *Twitter* tem desempenhado um papel significativo nisso. Como uma das redes sociais mais utilizadas do planeta, o Twitter permite que os usuários façam postagens expressando diversos aspectos do seu dia a dia. Com a aproximação de grandes eventos de reconhecimento mundial(e.g., eventos esportivos), é natural que os usuários também expressem seus comentários sobre os diversos aspectos dessas competições. No entanto, devido à enorme quantidade de comentários gerados diariamente, é possível observar uma variedade de interpretações que refletem diferentes aspectos do evento, como discussões sobre escolha dos técnicos, desempenho das equipes e até mesmo comentários pessoais sobre os participantes. Este artigo apresenta uma abordagem para analisar as discussões em torno de um evento, utilizando a *Copa do Mundo* como exemplo para validar o método. Para isso, foram empregados dados coletados do *Twitter*, os quais foram usados como entrada em técnicas de agrupamento, a fim de identificar potenciais conjuntos de comentários relacionados a competição. Como resultado, foram obtidos 13 grupos diversos, cada um com características únicas, abrangendo desde avaliações individuais, até críticas ao país-sede Qatar, rivalidades entre seleções, entre outros aspectos. Esses resultados indicam a existência de padrões nos comentários sobre um tema específico, sugerindo que os usuários buscam comentar temas do momento e com grande engajamento.

PALAVRAS-CHAVE

Coleta de Dados, Clusterização, Analise de Dados, Processamento de Linguage Natural

1 INTRODUÇÃO

A internet é um sistema global de transferência de dados e que vem se tornando uma ferramenta que participa cada vez mais de vários ramos da sociedade, como negócios, comunicação e cultura popular através do mundo Brignall et al, 2007 [9]. No entanto, é possível afirmar que seu imenso crescimento está trazendo muita preocupação para os especialistas. Diante da popularização da internet logo após 1988[1], inúmeros serviços de transmissão de mensagens foram criados, como Yahoo e serviços de e-mail, e mais recentemente as redes sociais se destacaram por levar a informação aos clientes de maneira segura, rápida e fácil.

As redes sociais são plataformas online onde as pessoas trocam mensagens, retratando opiniões e ideias acerca de acontecimentos do dia a dia. Dentre as redes sociais mais usadas no Mundo, o *Twitter*[4] se destaca por ser um serviço de microblogs, sendo caracterizada por pequenas postagens, chamadas *Tweets*, limitadas até 280 caracteres, forçando assim que os usuários façam postagens curtas e diretas.

Considerando o vasto contingente populacional do planeta, foram criados eventos esportivos com o objetivo de promover uma competição saudável entre as nações ao redor do mundo. Alguns eventos de destaque incluem a Copa do Mundo, a Fórmula 1, as Olimpíadas e a NBA, que abrangem uma variedade de esportes diferentes. Além de proporcionarem entretenimento ao público, esses eventos despertam o interesse das pessoas em discutir o que está acontecendo nesse espetáculo. Consequentemente, o público busca plataformas onde se possa debater e comentar sobre os acontecimentos, destacando-se assim as redes sociais como o Twitter, o Facebook e o Reddit.

A partir disso, considerando a quantidade massiva de tweets publicados diariamente pela população (mais de 800 milhões de tweets por dia[5]), é possível afirmar que uma imensidão de comentários surge em relação a um determinado tema. Essa enorme quantidade de *tweets* proporciona uma variedade de características sobre o evento, desde comentários sobre a performance dos atletas dentro e fora do campeonato, até desempenhos individuais e em equipe, entre outros aspectos.

Este artigo apresenta uma abordagem para analisar o comportamento dos *Tweets* durante grandes eventos. Para isso, foram utilizadas técnicas de *coleta e limpeza de dados*, a fim de reunir e tratar os dados para torná-los melhores para o processo de *Clusterização*, onde esses dados serão organizados em grupos, potencialmente revelando características em comum. Para validação da abordagem proposta, foi conduzido um estudo de caso com dados obtidos do *Twitter*, onde esses dados possuem alguma conexão com o evento Copa do Mundo de 2022, no Qatar.

Dois experimentos foram conduzidos para compreender esses dados. No primeiro experimento, os dados foram agrupados utilizando redutores de dimensionalidade com a ajuda de dois algoritmos de clusterização diferentes. No entanto, os resultados não foram satisfatórios, pois o redutor acabou removendo muitas informações e simplificando demais o texto. Com o objetivo de melhorar os resultados, um segundo experimento foi realizado no qual o processo de redução de dimensionalidade não foi utilizado. Isso resultou em clusters de maior qualidade e melhor divisão entre si, demonstrando a existência de grupos de dados.

O restante desse trabalho está estruturado da seguinte maneira. A Seção 2 fornece detalhes dos fundamentos teóricos para o entendimento da pesquisa. Logo após, a Seção 3 discorre sobre os procedimentos executados para a realização dos experimentos. Posteriormente, a Seção 4 descreve os experimentos realizados e os resultados alcançados, assim como, as possíveis causas dos resultados obtidos. Por fim, a Seção 5 apresenta conclusões para esse trabalho e aponta para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta os fundamentos teóricos essenciais para uma compreensão abrangente da pesquisa. Serão explicados os algoritmos utilizados para gerar embeddings a partir de sentenças, bem como a redução de dimensionalidade e a clusterização. Além disso, serão abordados os dois tipos específicos de algoritmos de clusterização empregados neste artigo.

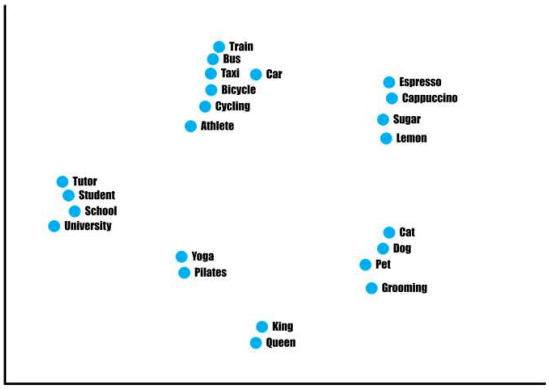
2.1 Embeddings

O Processamento de Linguagem Natural (PLN) desempenha um papel fundamental na compreensão e interpretação da linguagem humana, visando capturar suas relações sintáticas e semânticas. Essa é uma tarefa complexa, uma vez que essas relações não são facilmente discerníveis. Para lidar com esse desafio, é necessário criar representações da linguagem humana que possam ser compreendidas pelas máquinas. Um paradigma amplamente utilizado para adquirir essas representações é a hipótese distribucional de Harris (1954) [3], que postula que palavras em contextos semelhantes possuem significados semelhantes. Por exemplo, no contexto de "comer", as palavras "hambúrguer" e "pizza" compartilham significados semelhantes. Outro ponto importante descrito no artigo é que, em diversas linguas, palavras possuem ligações entre si, onde dada uma certa palavra, apenas outras palavras virão depois, para completar o sentido da frase.

Existem diversas técnicas para representar essas palavras, e este artigo utilizou um método que atualmente está em destaque no campo do PLN devido à sua eficácia e poder na captura dessas relações, chamado de Embeddings. Esse método cria uma representação numérica de palavras, frases ou documentos, que foi compreendida por modelos de aprendizado de máquina, transformando o texto em vetores numéricos densos em vez de representações esparsas individuais, como exemplo o Bag of Words. Como ele é treinado com bases de dados enormes, esses algoritimos conseguem capturar as semelhanças entre as palavras, percebendo as relações indiduais e gerais das palavras em um texto, como foi descrito anteriormente. Como podemos ver na Figura 1, temos uma distribuição de palavras em um espaço 2D, geradas a partir de embeddings, que compreendem as semânticas das palavras. A imagem foi obtida em [7].

2.2 Redução de dimensionalidade

Por mais que esses embeddings gerados sejam bastante representativos, eles possuem um vetor denso contendo muitas informações que podem não ser vantajosos para análises, além de gerar um custo a mais para o algoritimo de agrupamento para compreender o grupo. Como descrito em [2], a redução de dimensionalidade é



Fonte: George Pipis

Figura 1: Representação dos embeddings obtidos em um espaço 2d

usada em diversas areas de pesquisa, incluindo processamento de imagem, analise de series temporais, analise automatica de texto.

Na tentativa de melhorar a eficiência e preservar os principais pontos da semânticas das frases foi utilizado a técnica de UMAP[1]. Ele é um algoritmo redutor de dimensionalidade não-linear que procura apreender a estrutura dos dados e achar embeddings de baixa dimensão que conseguem preservar a estrutura essencial da entrada. Seu comportamento busca balancear a estrutura local ou estrutura global dos dados, ou seja, com a definição de alguns parâmetros para valores baixos ele concentra valores mais especificos, enquanto valores altos aprendem de maneira mais geral, em detrimento dos detalhes.

2.3 Clusterização

A fim de agrupar todos esses embeddings, e consequentemente obter insights acerca dos dados, foi utilizado métodos de Agrupamento, ou Clustering, que tentam achar estruturas em conjuntos que não foram categorizadas. Clustering é um método de aprendizagem não-supervisionada que acaba por conseguir identificar grupos nos dados recebidos.[6] Existem diversas técnicas para amontoar esses grupos, mas foi usado 2 tipos: os particionais e os baseados em Densidade.

Os algoritmos particionais são baseados em centróides, onde existem uma quantidade pré-definida de centros e os pontos pertencentes ao cluster são calculados com base na sua distância para o centróide [6]. A técnica usada aqui foi o K-means que, além do citado antes, busca sempre reduzir a variância interna dentro do cluster, e maximizar a variância entre os clusters de fora. Isso é necessário para que dados sejam colocados em seus respectivos grupos.

Outra técnica que foi usada, foi o baseado em densidade, onde ele se caracteriza por identificar nuvens de dados, seguindo a densidade dos pontos [8]. O algoritmo usado foi o DBScan(Density-Based Spatial Clustering of Applications with Noise), onde para encontrar

[1] https://umap-learn.readthedocs.io/en/latest/index.html

essas áreas de alta densidade, ele vai se expandindo com o intuito de aglutinar pontos que estão dentro de determinado raio. Uma diferença dele é que ele é capaz de detectar outliers, que podem surgir nos dados obtidos.

3 METODOLOGIA

Esta Seção descreve a metodologia utilizada para analisar os conteúdos dos *tweets*. Para explicar melhor esse material serão apresentados tópicos sobre coleta de dados, tratamento dos mesmos e uso de algoritmos de clusterização para agrupar os dados. Foi sintetizado um notebook, contendo todo o pipeline de estudo, disponibilizado no *Google Colaboratory* [2]

3.1 Recolhimento dos dados

3.1.1 Escolhas de ferramentas. Para que fosse possível analisar os dados, uma etapa principal é recolher o material de alguma fonte de informação. Para isso, foi escolhido o *Twitter*[3], devido a grande quantidade de Tweets gerados todos os dias. Para realização da consulta, foi utilizado a biblioteca *Tweepy*[4], onde provê uma interface de comunicação mais fácil com o *Twitter*, formatando as entradas de sua função para um objeto *json*, que é enviada diretamente a API do *Twitter*.

Para poder recuperar as mensagens, foi necessário possuir credenciais de acesso do próprio *Twitter* através do *Twitter Developer*[5]. Existe uma restrição da própria api, que para retornar os dados anteriores à 7 dias da consulta, deve-se adquirir o nivel de *Acesso Acadêmico*, que obtém acesso a todos os dados desde Março de 2006. A forma de obtenção desse acesso é requisitando diretamente ao próprio *Twitter*.

3.1.2 Definição da consulta. Com base nisso, destacamos o período da Copa do Mundo de 2022 no Qatar e coletamos os dados correspondentes ao período do evento: de 20 de novembro de 2022 a 18 de dezembro de 2022. Em seguida, é necessário criar consultas relacionadas ao tema em questão. Essas consultas podem ser qualquer sequência de caracteres, limitadas a um máximo de 1024 caracteres. Levando isso em consideração, foram criadas 8 consultas contendo hashtags, que são frases ou palavras associadas a temas ou discussões que se deseja categorizar, facilitando assim a busca, ao inserir o caractere de cerquilha antes dessas sentenças. (#). As consultas escolhidas foram: *(#worldcup), (#worldcup2022), (#WC2022), (#fifaworldcup), (#qatar22), (#qatarworldcup), (#qatar22 #qatar), (#fifa #fifaworldcup #qatar2022).* Em consultas que possuem mais de 1 token, a API do *Twitter* entende como uma disjunção, buscando assim uma representação OU a outra.

Outras duas informações também foram colocadas ao final de todas as consulta: os parâmetros "lang:en" e "-is:retweet". Essas restrições foram bastante importantes por dois motivos. Em primeiro lugar, elas restringem a busca apenas a tweets escritos em inglês, o que ajuda a focalizar o escopo da pesquisa. Em segundo lugar, a restrição de retornar apenas tweets que não são retweets é relevante para evitar a inclusão de sentenças que são simplesmente compartilhamentos de tweets existentes. Isso ajuda a reduzir a quantidade de tweets repetidos na análise, tornando os dados mais concisos e relevantes para o estudo.

3.1.3 Definição do periodo das consultas. Um dos principais desafios na análise de dados são os próprios dados. Se as informações não forem significativas ou abordarem apenas um aspecto do tema, a análise ficará comprometida, não refletindo a realidade. Com o objetivo de obter amostras representativas para todo o período da Copa, foi estabelecido um período de 4 semanas que engloba todos os dias do evento. Na Tabela 1 são apresentados as semanas definidas.

Tabela 1: Definição de semanas

Semana	*Data Inicio*	*Data Fim*
1	20/11/2022	26/11/2022
2	27/11/2022	03/12/2022
3	04/12/2022	10/12/2022
4	11/12/2022	18/12/2022

Outro aspecto levado em consideração é que a *API do Twitter* retorna os tweets ordenados de forma temporal ao realizar consultas. Isso significa que os dados mais recentes dentro do intervalo de datas especificado na consulta são retornados primeiro. Como consequência, se tivéssemos realizado apenas uma consulta abrangendo os dias da Copa (20/11 a 18/12), a maioria esmagadora dos resultados seriam sentenças relacionadas ao desenrolar do evento, pois o mundo estava focado no ápice da competição. Essa constatação foi confirmada por uma consulta prévia.

Com essas informações, estabelecemos um limiar de 5000 tweets por semana para gerar uma amostra eficiente e sem viés, dentre as 4 semanas definidas, para cada uma das 8 consultas citadas anteriormente, totalizando aproximadamente $5000 \times 4 \times 8 = 160.000$ tweets.

3.1.4 Realização da consulta. Com base nas definições anteriores, o processo de busca por tweets foi realizado entre os dias 18 e 19 de abril de 2022. A consulta retornava dois campos: o primeiro correspondente ao ID do tweet e o segundo à própria frase. No entanto, devido a algumas hashtags não serem populares e retornarem apenas algumas dezenas de tweets, o número total de mensagens coletadas foi de 118.991. Todos esses dados foram importados para um DataFrame da biblioteca Pandas[6], amplamente utilizado em Ciência de Dados devido à sua eficiência no processamento de grandes volumes de dados. Após a criação de duas colunas para armazenar as informações pesquisadas, o DataFrame foi exportado para um arquivo JSON para análises posteriores. É importante ressaltar que o Twitter não permite o compartilhamento de dados obtidos por meio dessas pesquisas.

3.2 Pré-processamento dos Dados

Para analisar os dados e utilizá-los nos métodos de clusterização, foi necessário realizar o tratamento dos mesmos e aplicar diversos

[2] 1A8_ZxtSXbQ9EfbScapB_lD_hJLwZ08rS?usp=sharing
[3] https://twitter.com/home
[4] https://www.tweepy.org/
[5] https://developer.twitter.com/en
[6] https://pandas.pydata.org/

procedimentos com o objetivo de melhorar sua qualidade. Para isso, foram seguidas várias etapas a fim de alcançar essa condição.

3.2.1 Limpeza dos dados. A primeira etapa envolveu a remoção de tweets duplicados com base no ID, pois uma mesma sentença poderia conter várias hashtags e ser retornada por consultas diferentes. Isso resultou em um total de 82.549 tweets para análise. Em seguida, foram removidas informações desnecessárias para a análise, como links externos, hashtags, menções a usuários e caracteres não alfanuméricos, como sinais de pontuação e parênteses. Todas essas informações foram substituídas por espaços vazios, utilizando a biblioteca*Regex* [7].

Por fim, foram removidas as stopwords do idioma inglês utilizando a biblioteca *NLTK*[8]. Stopwords são palavras comuns que geralmente não contribuem para uma melhor compreensão do texto pela máquina. Além disso, foi aplicado o processo de *Stemming*, também utilizando o *NLTK*, a fim de reduzir a complexidade da clusterização.

3.2.2 Transformação dos Dados. Para a geração dos embeddings, conforme explicado anteriormente, optou-se por utilizar um modelo pré-treinado existente do Sentence Transformers[9]. Entre os modelos pré-treinados disponibilizados, foi escolhido o *"all-MiniLM-L6-v2"*[10] que converte sentenças ou parágrafos em vetores de tamanho 384. Esse modelo foi selecionado devido ao seu equilíbrio entre desempenho, tamanho do modelo e velocidade de geração de embeddings.

3.3 Definição dos algoritmos de clusterização

Ao final da etapa de transformação dos dados, foram aplicados dois algoritmos de agrupamento não supervisionado. Conforme explicado na Seção 2, o primeiro algoritmo utilizado foi o K-Means, devido à sua simplicidade e capacidade de lidar com diferentes tipos de dados. A definição do número de clusters precisou ser testada manualmente por meio de métricas de avaliação. O segundo algoritmo utilizado foi o DBSCAN, que é capaz de encontrar agrupamentos de dados sem a necessidade de escolher previamente a quantidade de grupos. Essa abordagem foi adotada para obter uma outra perspectiva na análise dos dados, resultando em uma possível visão diferente das informações obtidas.

3.4 Redutor de Dimensionalidade

Para complementar a execução dos algoritmos de clusterização, foi decidido utilizar a técnica de redução de dimensionalidade. O algoritmo escolhido para essa finalidade foi o *UMAP*, que tem como objetivo reduzir o vetor de alta dimensão gerado pelos embeddings, preservando os agrupamentos e separações em uma projeção de menor tamanho. A escolha por essa técnica foi motivada pelo fato de que o vetor de 384 dimensões gerado pelo embedding demandava um alto poder computacional para o cálculo dos clusters. Além disso, é importante destacar que a quantidade de dimensões é bastante elevada e provavelmente nem todas são necessárias, o que torna possível reduzi-las sem comprometer a qualidade dos grupos obtidos.

[7]https://docs.python.org/3/library/re.html
[8]https://www.nltk.org/
[9]https://www.sbert.net/
[10]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

3.5 Metricas de Avaliação e Melhoria de hiperparametros

Durante o periodo de execução dos algoritimos de agrupamento, os resultados obtidos foram submetidos a tecnicas que tentam melhorar a qualidade da solução e também para achar a quantidade de clusters, no caso do k-means. A metricas de avaliação usada foi o *Coeficiente de Silhueta*. O *Coeficiente de Silhueta* é uma metrica usada para avaliar a qualidade do algoritimo, medindo quão bem os pontos de um Cluster estão agrupados em relação ao outros clusters. Com o propósito de retornar o melhor conjunto de hiperparâmetros que, quando usado com os dados, retornem clusters bem significativos, também foi utilizado a técnica de Grid Search, que é uma técnica de testar todas as combinações de hiperparâmetro pré-escolhidos. A escolha desses hiperparâmetros se deu a testes anteriores que sugeriram uma pontapé inicial. Os parâmetros testados foram:

- para o UMAP: "n_neighbors" e "n_components"
- para o K-means: max_iter" e "num_cluster
- para o DBscan: "eps" e "min_samples":

4 RESULTADOS E DISCUSSÃO

Nesta seção, serão apresentados os experimentos realizados, os resultados obtidos e as discussões relacionadas a eles. Para isso, foi utilizado o pré-processamento dos dados descrito na Seção 3.2 na criação de dois experimentos com o mesmo conjunto de dados. O primeiro experimento consistiu na geração de agrupamentos e análise dos mesmos, utilizando os dois algoritmos de clusterização mencionados na Seção 3.3, com a inclusão da etapa de redução de dimensionalidade no pipeline. O objetivo desse experimento foi obter grupos representativos, com o uso da redução de dimensionalidade, a fim de observar o quanto essa técnica influencia nos resutados. Já o segundo experimento também utilizou os mesmos algoritmos de clusterização, porém sem a etapa de redução de dimensionalidade, buscando melhorar os resultados.

4.1 Experimento 1

Para este experimento, os dados foram pré-processados e um redutor de dimensionalidade foi implementado com o objetivo de reduzir a complexidade dos dados e facilitar a clusterização pelos algoritmos definidos na Seção 3.3. Além disso, foi utilizada a técnica de Grid Search para buscar o conjunto de parâmetros que resultem em clusters bem divididos e explicativos, variando os parâmetros definidos na Seção 3.5.

O processo ocorreu da seguinte maneira: foram criadas duas abordagens para analisar os dois algoritmos de clusterização, e os parâmetros foram configurados. Os parâmetros referentes ao UMAP foram utilizados para ambas as execuções do K-means e DBSCAN. A partir disso, as respectivas variáveis foram definidas e as combinações 2 a 2 foram geradas pela técnica de Grid Search, resultando em uma média de 50 combinações para testar em cada algoritmo.

Dentro do processo de Grid Search, também foram coletados os scores calculados pelo Coeficiente de Silhueta. Ao final da execução, o melhor par de parâmetros foi escolhido com base no melhor coeficiente obtido, e esses parâmetros foram utilizados. Para o primeiro passo da execução, selecionamos uma amostra de tweets, obtendo os primeiros 40.000 registros.

Os resultados desse experimento não foram satisfatórios pelo resultado obtido. Levando em consideração a avaliação apenas por métricas, essa tecnica evidenciou um silhuette score para o Kmeans como sendo 0.7304597, valores bons segundo a documentação devem ser maiores que 0.5. Porém retornando os melhores parametros como sendo 3 clusters e 500 interações maximas (max_iter) para análise. Analisando os resultados, vimos que 1 cluster possui quase a totalidade dos resultados, o que acaba por não definir nada. Os valores exatos calculados, estão na Tabela 2.

Tabela 2: Clusters calculados

Cluster	*Quantidade de dados*
1	36191
0	3597
2	212

Já o Dbscan, foram os resultados retornaram uma imensa quantidade de clusters. Seguindo esse mesmo pipeline citado, temos os resultados validando um silhuette score de 0.008715929, mas retornando assim 39 grupos, com os parametros 'eps': 1, 'min_samples': 9. Os resultados também não foram satisfatorios pois os agrupamentos não estavam bem definidos

Neste experimento realizado, observamos que o UMAP provavelmente está eliminando informações importantes das sentenças, resultando em uma simplificação que perde os aspectos que caracterizam as mesmas. Diante disso, foi necessário realizar outro experimento para abordar essa questão.

4.2 Experimento 2

Neste experimento, foi empregado um pipeline de execução que não incluiu a etapa de redução de dimensionalidade, pois essa etapa não se mostrou eficaz na geração de clusters satisfatórios. Além disso, utilizou-se o conjunto de dados completo, que consiste em 82.549 sentenças.

Os resultados do *DBSCAN*, no entanto, não se mostraram satisfatórios, pois o algoritmo não se comportou de maneira representativa. Os parâmetros que retornaram o melhor valor de silhueta foram -0.0042051333, com 21 clusters. No entanto, apenas um desses clusters continha 92% dos dados, enquanto 7% foram classificados como outliers.

Diante disso, partimos para o K-Means e, após a execução do pipeline, obtivemos os seguintes resultados da Tabela 3 e também plotamos em um grafico 2d a sua estrutura, de acordo como mostra a Figura 2. Como podemos perceber, existem grupos bem definidos ao centro e que possuem uma boa quantidade de integrantes.

O cálculo de silhueta para o K-Means retornou 0.025760774, com 13 clusters e 1500 iterações máximas. Embora o valor esteja próximo de 0 e não seja considerado muito bom, os clusters foram identificados de maneira representativa.

A caracterização desses agrupamentos foi realizada por meio da análise de uma amostra das 400 primeiras sentenças de cada grupo, permitindo a observação de suas distintas características. Com base nisso, iremos analisar os grupos e evidenciar suas principais particularidades.

Tabela 3: Clusters calculados

Cluster	*Quantidade de dados*
1	16149
9	9511
0	8655
3	8612
11	7915
4	5908
8	5103
12	4325
7	4292
2	4192
6	3449
5	2934
10	1504

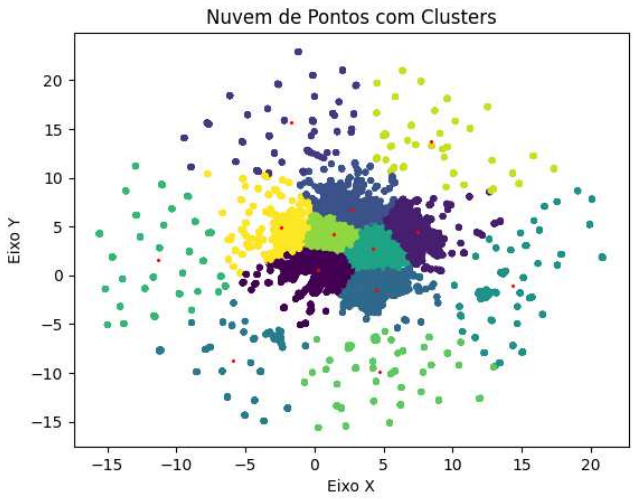


Figura 2: Representação da nuvem de pontos

Grupo 1: Representa tweets rápidos relacionados à Copa, com predominância de opiniões negativas, como frustração, raiva e tédio. Exemplos de sentenças incluem "Well that was pathetic" e "Wasn't great was it", mostrando a insatisfação dos usuários.

Grupo 9: Reflete opiniões sobre aspectos técnicos do jogo, como posicionamento em campo, críticas ao técnico e baixo desempenho dos jogadores. As sentenças expressam insatisfação com o desempenho da equipe e às decisões tomadas, como "Playing mount a whole 90 minutes while foden is on the bench is mental to me" e "What a joke of a game".

Grupos 0 e 3: Esses clusters estão relacionados a previsões sobre as partidas, ou seja, qual time específico vencerá um determinado jogo. O Cluster 3 inclui discussões provenientes de sites especializados, enquanto o Cluster 0 consiste em opiniões informais de pessoas comuns. Algumas frases que auxiliam na compreensão são: "Incredible performance by the US team! They outplayed and now gotta beat Iran. Wow, is Tuesday going to be insane!" para o Cluster 3 e "I hope we play like this against Iran" e "Can South Korea beat Brazil?" para o Cluster 0.

Grupo 8: Este cluster aborda o Catar, país-sede, e envolve preocupações com direitos humanos, tratamento de trabalhadores migrantes, restrições e direitos LGBTQ+. É importante ressaltar que algumas sentenças podem conter conteúdo ofensivo e sensível. Exemplos de sentenças que expressam insatisfação ou repúdio ao país são: 'Did Qatar actually qualify or are they only there because they're the hosts', 'Have you guys sponsored the murder of any migrant or LGBTQ people recently?', 'That's for the migrants that died building your corrupt'.

Grupos 2, 4, 5 e 11: Esses clusters são considerados spam e não estão diretamente relacionados ao evento em si. O Cluster 11 pode envolver discussões sobre o valor de mercado dos times, o Cluster 4 pode conter mensagens promocionais e apostas esportivas, o Cluster 2 pode abordar a participação em torneios para ganhar prêmios, e o Cluster 5 pode estar relacionado a serviços de streaming. Alguns exemplos incluem: 'I just earned the World Pint 2022 badge on', 'is LIVE Talking World Cup'.

Grupo 6: Este cluster destaca os papéis individuais dos jogadores, incluindo suas conquistas, atuações e comparações com outros jogadores. Há diversas opiniões, principalmente sobre Messi e sua atuação na Copa, chegando até o nível de idolatria. Exemplos de frases que representam esse grupo são: 'Ronaldo has to face his own creation', 'Ronaldo The best player ever', 'There are only two games left for Lionel to win his first', 'Mbappe or Messi captain tomorrow'.

Grupos 7 e 12: Esses clusters refletem rivalidades entre equipes e suas reações aos jogos. O Cluster 7 se refere principalmente às disputas da Argentina contra Holanda e México, enquanto o Cluster 12 aborda o confronto entre Inglaterra e Estados Unidos, incluindo aspectos do desempenho das equipes e opiniões sobre o resultado do jogo. Alguns exemplos de sentenças desse grupo são: 'ARGENTINA VS MEXICO TOMORROW IS CRAZY', 'What a game! Massive win for Argentina', 'USA wins 2-0 against the Brits'.

Grupo 10: Esse cluster contém menções específicas à música tema da Copa do Mundo. Os usuários expressam admiração, orgulho e apoio à música, além de mencionarem seu vício em ouvi-la nas plataformas de streaming. Também há gratidão pelo trabalho do artista Jungkook, responsável pela música. Alguns exemplos de sentenças que representam esse grupo são: 'WE ARE THE DREAMERS', 'One of my favorite songs ever', 'I am so amazed by Jungkook's new single "Dreamers"'.

5 CONCLUSÕES E TRABALHOS FUTUROS

Para a realização das investigações e experimentos deste trabalho, foram utilizados algoritmos de clusterização, técnicas de redução de dimensionalidade e geração de embeddings. Procurou-se obter um conjunto de dados em inglês, com a presença de hashtags que representassem aspectos relevantes da Copa do Mundo.

Após a obtenção do conjunto de dados, foram realizadas uma série de melhorias para prepará-lo adequadamente para os algoritmos de clusterização. Os dados foram submetidos a um gerador de embeddings e, em seguida, foram exploradas duas abordagens distintas. A primeira envolveu a utilização de um redutor de dimensionalidade na tentativa de aprimorar os resultados, enquanto a segunda descartou essa etapa. Em seguida, os dados foram submetidos aos algoritmos de clusterização. A avaliação inicial foi realizada por meio de métricas de silhueta, seguida de uma análise textual dos dados.

A partir dos experimentos realizados neste trabalho, foi possível observar a existência de grupos com características únicas entre si. Esses grupos revelam que, mesmo com a diversidade de opiniões entre os usuários, há uma ligação entre essas opiniões. Essa ligação pode ser atribuída ao fato de que, mesmo sendo tweets rápidos, os usuários estão expostos a conteúdos semelhantes em outras fontes de informação(e.g. Televisão, Radio, Transmissões na internet), o que influencia suas percepções e opiniões.

Para trabalhos futuros, pretende-se avaliar os dados utilizando novos algoritmos de clusterização, como MeanShift ou Affinity Propagation, a fim de compreender como esses dados podem se comportar em abordagens diferentes. Além disso, serão exploradas outras técnicas de redução de dimensionalidade, como PCA ou t-SNE, em conjunto com algoritmos de clusterização, para entender melhor o motivo da baixa qualidade dos grupos gerados pelo UMAP.

Por fim, planeja-se recuperar mais tweets e seus metadados, identificando assim sua origem e o total de compartilhamentos. Essa abordagem permitirá obter mais insights sobre os dados, incluindo possíveis correlações entre o conteúdo dos tweets, sua disseminação e a formação dos grupos. Essas informações adicionais podem enriquecer ainda mais a compreensão dos resultados e contribuir para futuras pesquisas.

AGRADECIMENTOS

Primeiramente, gostaria de expressar minha gratidão a Deus por permitir que eu vivenciasse toda essa experiência de graduação. Agradeço de coração aos meus pais, José Ivo e Jeane Ramos, por todo o esforço e sacrifício que fizeram para possibilitar minha permanência no curso. Através de seus conselhos e apoio incondicional, eles sempre estiveram ao meu lado nos momentos em que mais precisei. Também gostaria de agradecer ao meu irmão, Ian, que, mesmo com apenas um ano de existência, me trouxe uma imensa força e motivação para superar desafios incrivelmente difíceis.

Gostaria de expressar minha profunda gratidão ao meu orientador, Cláudio Campelo, por sua dedicação, orientação e compartilhamento de conhecimento ao longo deste projeto. Seu apoio e feedback foram indispensáveis para o desenvolvimento e aprimoramento deste trabalho.

Não posso deixar de mencionar meus amigos e companheiros de apartamento, Carlos Roberto e Matheus Lisboa, cuja generosidade e apoio me permitiram estabelecer residência na cidade e, assim, me dedicar mais aos estudos. A presença deles em minha vida foi fundamental para tornar minha jornada acadêmica mais tranquila e produtiva.

Também gostaria de agradecer aos amigos que fiz durante o curso, os quais contribuíram significativamente para tornar minha graduação uma experiência enriquecedora e memorável. Agradeço ainda à Universidade Federal de Campina Grande e a todos os professores do curso de Ciência da Computação, que desempenham um trabalho excepcional, sempre dispostos a compartilhar seus conhecimentos e proporcionar um ambiente acadêmico inspirador.

REFERÊNCIAS

[1] Thiago Barros. 2013. Internet completa 44 anos; relembre a história da web. https://www.techtudo.com.br/noticias/2013/04/internet-completa-44-anos-relembre-historia-da-web.ghtml
[2] A. Pascual-Montano C.O.S. Sorzano, J. Vargas. 2014. A survey of dimensionality reduction techniques. *arXiv:1403.2877* (Mar 2014).
[3] Zellig S. Harris. 1954. Distributional Structure. *Word* 10, 2-3 (1954), 146–162.
[4] Alfred Lua. 2023. 21 Top Social Media Sites to Consider for Your Brand in 2023. https://buffer.com/library/social-media-sites/#:~:text=1.,billion%20people%20using%20it%20monthly.
[5] Yaqub M. 2023. How Many Tweets per Day 2022 (New Data). https://www.businessdit.com/number-of-tweets-per-day/#:~:text=How%20Many%20Tweets%20per%20Day%20in%20the%20World,of%2010%2C033%20tweets%20every%20second!
[6] T. Soni Madhulatha. 2012. An Overview on Clustering Methods. *arXiv preprint arXiv:1205.1117* (2012).
[7] George Pipis. 2020. A High-Level Introduction To Word Embeddings. https://predictivehacks.com/a-high-level-introduction-to-word-embeddings/
[8] Jörg Sander Arthur Zimek Ricardo J. G. B. Campello, Peer Kröger. 2019. Density-based clustering. *WIREs Data Mining Knowl Discovery* 10 (Oct 2019). https://doi.org/10.1002/widm.1343
[9] Thomas Van Valey Thomas Wells Brignall III. 2005. THE IMPACT OF INTERNET COMMUNICATIONS ON SOCIAL INTERACTION. *Sociological Spectrum* 25 (2005), 335–248.