ISSN: 2317-0840

Popularidade de Donald Trump e Barack Obama no Twitter: um estudo de caso via análise de cluster

Gabriel Baldasso[1†], Ricardo M. Salgado[2], Eric B. Ferreira[3]

[1] *Mestre em Estatística Aplicada e Biometria, Universidade Federal de Alfenas.*
[2] *Professor do Departamento de Ciência da Computação, Universidade Federal de Alfenas.*
[3] *Professor do Departamento de Estatística, Universidade Federal de Alfenas.*

Resumo: *O avanço da tecnologia e da internet nos últimos anos levou a um aumento massivo no uso das redes sociais. Somente o Twitter de 2019 a 2020 viu seu número de usuários crescer mais de 20%, o que trouxe um aumento significativo no número de publicações. Mensagens postadas no Twitter são chamadas de tuítes e têm uma métrica de engajamento medidas por curtidas, retuítes e respostas. Logo, este artigo busca realizar uma análise de agrupamento dos tuítes dos ex-presidentes dos Estados Unidos (EUA): Barack Obama e Donald Trump, com base nas "proporções"entre as métricas: curtidas, respostas e retuítes. Os tuítes dos ex-presidentes foram agrupados em dois e três grupos, usando os algoritmos K-Means e Fuzzy C-Means e comparados com as médias das proporções. De acordo com a análise do K-Means, os resultados dos tuítes de Barack Obama mostraram um grupo com um valor médio maior para a relação retuíte/curtida. No entanto, esses resultados não foram encontrados na análise com Fuzzy C-Means. No agrupamento com os tuítes de Donald Trump, um grupo apresentou uma maior proporção de respostas/retuítes em ambas as análises. Os resultados do algoritmo K-Means, reuniu tuítes com natureza populares para a conta de Barack Obama e tuítes com caráter polêmicos para as publicações de Donald Trump. Os resultados do algoritmo Fuzzy C-Means diferiram dos resultados do K-Means para os tuítes de Barack Obama. No entanto, para os tuítes de Trump, os dois algoritmos mostraram maior similaridade nos resultados.*

Palavras-chave: Twitter; Análise de Cluster; K-Means; Fuzzy C-Means; Ciência dos Dados

Abstract: *The advancement of technology and the internet in recent years has led to a massive increase in the use of social networks. Only Twitter from 2019 to 2020 saw its number of users grow by more than 20%, which brought a significant increase in the number of publications. These messages posted on Twitter are called tweets and have an engagement metric measured by: likes, retweets, and responses. This paper performs a clustering analysis of the tweets from the Ex-Presidents: Barack Obama and Donald Trump, based on the "proportions"between the metrics: likes, responses, and retweets. The tweets of the former Presidents were grouped into two and three groups, using the K-Means and Fuzzy C-Means algorithms and compared with the averages of the proportions. According to the K-Means analysis, the results of Barack Obama's tweets showed a group with a high average value for the retweet/like ratio. However, these results were not found in the analysis with Fuzzy C-Means. In the grouping with Donald Trump's tweets, one group showed a high response/retweet ratio in both analyzes. The results presented by the Means algorithm are promising for being able to group tweets with general characteristics for Barack Obama's account and controversial and controversial tweets for Donald Trump's publications. The results of the Fuzzy C-Means algorithm differed from the K-Means results for Barack Obama's tweets. However, for Trump's tweets, the two algorithms showed high similarity in the results.*

Keywords: Twitter; Clustering Analysis; K-Means; Fuzzy C-Means; Data Science.

Introdução

As eleições dos Estados Unidos que ocorrem a cada 4 anos é um importante evento de interesse mundial, onde figuras públicas concorrem ao cargo de maior poder do país. Este evento movimenta todas as redes sociais, como é o caso do *Twitter*, que desempenha um papel central nas campanhas eleitorais, desde o seu uso pioneiro por Barack Obama nas eleições de 2008 (MINOT et al., 2021).

[†]Autor correspondente: gabrielbaldassso@gmail.com.

O *Twitter* é considerado um bom indicador quando se trata da opinião pública e divulgação de grandes eventos, este é constantemente utilizado para levantar e expressar opiniões políticas. Estudos recentes mostraram que candidatos políticos dos Estados Unidos que estão bem posicionados no *Twitter* e são popularmente aceitos, possuem maior probabilidade de serem eleitos (YAQUB et al., 2017). Quanto mais os conteúdos da página destes políticos forem compartilhados, maior será a influência destes na rede (SUNGE, 2021).

Conteúdos e publicações do *Twitter*, podem ser criados em forma de texto curto de até 280 caracteres conhecido como tuíte, ou ser compartilhado de alguma conta específica, sendo chamado de retuíte. Os usuários do *Twitter* possuem métricas de engajamento para a sua conta, como o número de seguidores e para as publicações são utilizados os: retuítes, curtidas e respostas (MINOT et al., 2021).

Buscar entender qual foi a aceitação dos seus seguidores em uma publicação é de extrema importância para identificar qual o seu nível de popularidade (SUNGE, 2021). Existem diversos métodos que buscam medir a popularidade de uma conta no *Twitter*, como, analisar o número de seguidores, interpretação das métricas dos tuítes e a análise de texto (SHAOZHI; FELIZ, 2010).

O *Twitter* tem sido alvo de intensas análises textuais, por meio de algoritmos de inteligência artificial, que buscam agrupar informações de interesse sobre a popularidade dos tuítes. Oktarina, Notodiputro e Indahwati (2020), fizeram uma aplicação do algoritmo *K-Means* em agrupar tuítes positivos e negativos relacionados às eleições presidenciais da Indonésia. Já Garg e Rani (2017), realizaram um agrupamento via *K-Means*, com informações geográficas dos tuítes. Muliawati e Murfi (2017) propõem o uso do algoritmo *Fuzzy C-Means*, como forma automática para detectar tópicos de tendência no *Twitter*. Já, Zadeh e Abbasov (2015), também utilizaram o algoritmo *Fuzzy C-Means*, para analisar tendências de popularidade das *hashtags* retiradas dos tuítes.

Um método alternativo à análise textual para medir a popularidade no *Twitter* ficou conhecido como *The Ratio*, que consiste em analisar a razão entre o número de respostas e retuítes que a publicação recebeu (MINOT et al., 2021). Se esta razão em um tuíte for de 2:1 em diante, é um forte indício que sua publicação possuem um caráter polêmico e não foi bem aceita pelos seus seguidores (ROTH, 2017).

Nesse contexto, o objetivo deste trabalho é realizar uma análise de *clustering* por meio dos algoritmos *K-Means* e *Fuzzy C-Means*, para agrupar tuítes baseado em três razões: "respostas/retuítes", "respostas/curtidas" e "curtidas/retuítes". Os grupos obtidos serão comparados com as médias das proporções, com o intuito de identificar grupos de publicações com características populares e polêmicas. Para este trabalho será utilizado os dados da plataforma *github* da conta do *FiveThirtyEight*, que contém as métricas dos tuítes das contas de Donald Trump "*@RealDonaldTrump*" e do Barack Obama "*@BarackObama*.

@RealDonaldTrump vs @BarackObama

Donald Trump foi o 45° presidente dos Estados Unidos, onde anunciou a sua candidatura no partido republicano em 2015 e emergiu como o favorito para as eleições presidenciais de 2016, levando assim a vitória. Trump, durante todo o seu mandato, fazia uso constante da sua conta particular do *Twitter @RealDonaldTrump* e era famoso por sempre fazer publicações e compartilhamentos polêmicos entre seus seguidores (YAQUB et al., 2017).

Trump, segundo pesquisas de opiniões públicas, foi considerado o presidente mais impopular dos Estados Unidos, sendo mal avaliado nas pesquisas de opiniões públicas até o final do seu mandato (MINOT et al., 2021). Trump chegou a possuir mais de 50 milhões de seguidores em sua conta do *Twitter* em 2020, segundo pesquisas do *WeAreSocial*, 2020.

Por não medir as palavras quando o assunto era utilizar a sua conta particular *@RealDonaldTrump* despertou a curiosidade de muitos pesquisadores do meio digital em estudar as suas publicações, tanto em caráter político, quanto em popularidade e poder de influência sobre seus

seguidores (ANSARI, 2020). Atualmente sua conta *@RealDonaldTrump* está banida permanentemente por estimular e disseminar manifestações pró-Trump, que levou a invasão ao Capitólio dos Estados Unidos e terminou com 5 americanos mortos.

Já Barack Obama, ficou famoso por ser o 44° presidente dos Estados Unidos e o primeiro afro-americano a ocupar o cargo. Sempre foi visto como o querido do *Twitter*, por apresentar zelo e cuidado com o que publica neste meio (YAQUB et al., 2017). Segundo pesquisa realizada por *WeAreSocial*, 2020 a conta *@BarackObama* é a conta que mais possui seguidores do *Twitter* com mais de cem milhões de seguidores, ficando a frente de famosos, como Katy Perry e Justin Bieber.

Obama foi o pioneiro em utilizar o *Twitter* como forma de divulgação política em sua campanha para as eleições presidenciais dos Estados Unidos em 2008 (ANSARI, 2020). Obama, por possuir características contrárias às do Trump e ser muito polido com relação às suas publicações, possui um elevado número de retuítes quando comparados aos do Trump, mostrando um maior nível de aceitação entre seus seguidores.

A junção dos tuítes do Trump e Obama é uma bela relação de água e vinho, quando se compara a popularidade e compartilhamento de assuntos polêmicos. Por isso, utilizar os tuítes de ambos para análise de engajamento e popularidade tem sido frequente entre as análises do *Twitter* (SUNGE, 2021).

Popularidade no *Twitter*

Redes Sociais como o *Twitter* que possuem atualmente mais de 330 milhões de usuários em todo o mundo, trazem a possibilidade de interação entre seus integrantes por meio da leitura de conteúdos e compartilhamento de notícias. O *Twitter* ficou mundialmente conhecido pelo grande uso pelos políticos e seus partidos no mundo inteiro (DATAREPORTAL, 2020).

Hoje, todos os candidatos e partidos políticos possuem presença marcada no *Twitter* e fazem uso constante deste meio para manter contato com os seus seguidores. Yaqub et al (2017), colocam que o uso constante do *Twitter*, demonstra transparência pelo político, o que pode levar a maior confiança da população ao seu respeito. O *Twitter* tornou-se o maior canal de comunicação política já existente, por ser um canal de interação rápida com o eleitorado, que permite medir em tempo real e de forma contínua as reações do público (EMA; MARK, 2018).

O uso do *Twitter* como fonte de dados para pesquisa, deu-se início, com o foco de buscar entender, qual a relação entre o conteúdo postado nesta rede com o dia-a-dia do usuário. Com este intuito, existem inúmeros estudos que utilizam os dados textuais presentes nos tuítes para análise de sentimentos (ANGER; KITTL, 2011).

Já medir o nível de popularidade de uma conta do *Twitter* não é uma tarefa fácil, visto que não existe um meio central para obter estas respostas (SHAOZHI, FELIX, 2010). Alguns trabalhos utilizam o número de seguidores junto com a análise da frequência de publicações (OKTARINA; NOTODIPOTRO; INDAHWATI, 2020). Outros trabalhos, utilizam a quantidade de menções que uma conta recebe e número de retuítes de suas publicações (GARG; RANI, 2017).

Riquelme e Cartergiani (2016) realizaram uma revisão bibliográfica e listaram os principais trabalhos envolvidos na busca por entender a popularidade das publicações no *Twitter*. Em seus resultados, apresentaram que a utilização do número de seguidores, quantidade de menções e número de retuítes, são as métricas mais utilizadas entre os 70 artigos analisados entre 2010 a 2015, para estudar popularidade no *Twitter*. Realizar a análise textual dos tuítes foi a análise mais popular entre os 70 artigos analisados, aparecendo em 41 deles. Porém, somente 3 trabalhos dos 70 analisados, utilizaram as métricas: curtidas, respostas e retuítes como variáveis para estudar as medidas de engajamento e popularidade, o que traz a ideia de que não existe um consenso definido na literatura com relação à análise e medidas de engajamento no *Twitter*.

Os algoritmos *K-Means* e Fuzzy são muito utilizados para análise de agrupamento dos dados textuais no *Twitter*. Contudo, utilizar estes algoritmos para agrupar tuítes baseados nas razões das métricas de engajamento é uma proposta inovadora.

Materiais e Métodos

Algoritmo *K-Means*

O algoritmo *K-Means* pertence à família dos algoritmos não supervisionados que possui um método de aprendizagem que agrupa as amostras fornecidas em K grupos que são definidos a priori por um centróide. Cada amostra utilizada no conjunto de dados está associada a um grupo específico. O processo de funcionamento do algoritmo consiste em duas etapas: Primeiro, cada amostra presente no conjunto de dados são associadas a um dos grupos com base na menor distância entre o centro do grupo e a posição da amostra, e por último a localização dos grupos são atualizadas (SINAGA; YANG, 2020).

Essas duas etapas são repetidas até que nenhuma mudança na posição dos centróides tenha acontecido, ou a distância entre os centróides antes e depois da atualização seja menor do que um valor previsto. Uma medida de distanciamento simples que pode ser utilizada é a distância Euclidiana entre dois pontos (MULIAWATI; MURFI, 2017).

Como resultado o algoritmo *K-Means* divide e agrupa todo o conjunto de dados dos K grupos fornecidos a ele no começo do processo. O *K-Means* é um algoritmo muito utilizado no processo de análise de texto e imagens, aprendizagem de máquina, meteorologia entre outras utilizações que não tenham a necessidade de possuir um rótulo específico para o conjunto de dados (ISAAC, 2018).

Algoritmo *Fuzzy C-Means*

O algoritmo *Fuzzy C-Means* baseia-se na identificação de grupos de dados pela similaridade apresentada entre as variáveis. A metodologia *Fuzzy* ficou famosa entre os algoritmos de agrupamento, por não apresentar resultados binários como o *K-Means*, mas sim, apresentar uma estimativa de localização dos dados dentro dos grupos definidos (FERRARO; GIORDANI, 2015). Esta técnica difusa de apresentação dos resultados faz com que os dados pertençam a um grupo até certo ponto.

O método Fuzzy foi proposto por Bezdek (1974; 1981) e é caracterizado pelo fato de um objeto poder ser membro de todos os grupos com diferentes graus de associação difusa. Como resultado no agrupamento *Fuzzy* sempre vão existir pontos com um nível de adesão a cada grupo variando de 0 a 1, o que traz uma visão mais realista dos dados, visto que raramente lidamos com dados que possuem divisões exatas (ZADEH; ABBASOV, 2015).

O algoritmo *Fuzzy*, utilizado para este trabalho, baseia-se no método *Fuzzy*. Os grupos obtidos pela análise, são distribuídos no espaço e estão entre os limites do número de grupos selecionados. Alguns pontos estão internamente aos grupos e outros na periferia (FERRARO; GIORDANI, 2015).

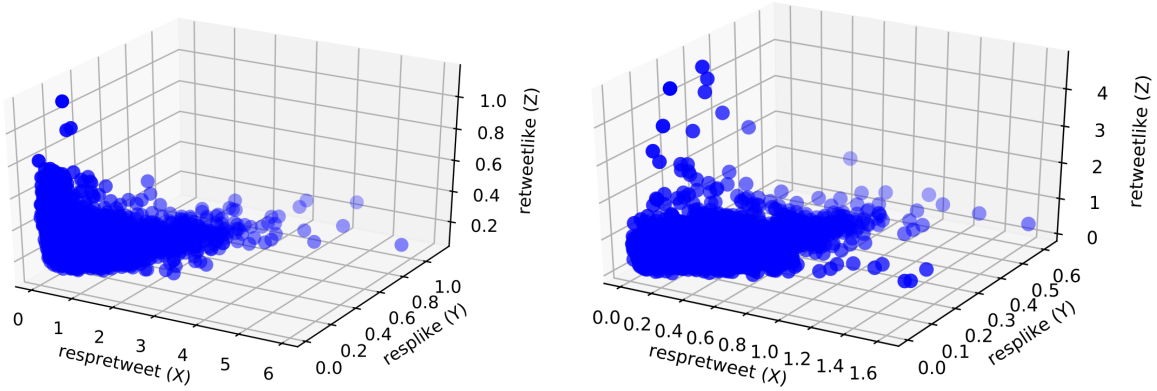
Base de Dados

Os tuítes utilizados para este trabalho são referentes ao artigo *The Worst Tweeter In Politics Isn't Trump* da página *FiveThirtyEight*, que estudou as métricas dos tuítes de Donald Trump e Barack Obama por meio de gráficos ternários (ROEDER, 2017).

Esta base de dados possui um arquivo chamado *Twitter Ratio*, que contém tuítes das contas "*@RealDonaldTrump*" e do "*@BarackObama*", em um total de 3232 tuítes para Donald Trump no período de 2016 a 2017. Já os tuítes do Barack Obama são de 3209 no período de 2014 a 2017. Este conjunto de dados possui 7 variáveis, contudo nos restringimos a utilizar somente as variáveis: respostas, retuítes e curtidas.

Para este trabalho foram criadas as razões entre as variáveis: respostas, retuítes e curtidas, que foram montadas da seguinte maneira: "respostas/retuítes", "respostas/curtidas" e "curtidas/retuítes" com suas respectivas representações: *respretweet*, *resplike* e *retweetlike*. Estas razões foram utilizadas baseadas na definição *The Ratio* que tem o objetivo de criar variáveis de

engajamento que sejam comuns para todas as publicações (ISAAC, 2018). Gráficos tridimensionais foram plotados para as duas contas e são apresentados na Figura 1. Os eixos tridimensionais são representados pelas variáveis *respretweet*, *resplike* e *retweetlike*.



(a) (b)

Figura 1: Distribuição dos tuítes da conta de Donald Trump (a) e Barack Obama (b).

A distribuição das publicações do Trump, apresentadas na Figura 1a, apresentam uma grande quantidade de tuítes no plano (XZ), quando comparado às publicações do Obama. Contudo, Obama possui publicações no plano (XZ) com maiores valores para a variável Z, comparados ao do Trump, como pode ser observado na Figura 1b. Contudo, os tuítes do Obama na média estão distribuídos no plano (XY).

Como o intuito do trabalho é definir grupos de tuítes baseados em métricas de popularidade, foi estabelecido a hipótese de encontrar de dois a três grupos de tuítes, definidos como: populares e polêmicos (K=2), levando em consideração a razão entre as métricas respostas/retuítes e curtidas/retuítes respectivamente. E grupos, populares, polêmicos e neutros (K=3). Portanto, baseado nesta hipótese, o número de grupos a ser utilizado nos algoritmos *K-Means* e *Fuzzy C-Means* será de K=2 e K=3.

Análises e Resultados

Análise agrupamento K=2

Com as variáveis calculadas, submetemos o conjunto de dados de cada conta nos algoritmo *K-Means* e *Fuzzy C-Means*. O algoritmo *K-Means* foi implementado na plataforma do *Google Colab* que disponibiliza em nuvem, um ambiente para execução de códigos *Python* sem a necessidade de configurações prévias. A função *K-Means* foi utilizada da biblioteca `Scikit-learn`. A visualização gráfica dos dados e as manipulações matriciais foram importadas das bibliotecas:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np.
```

Já a análise de agrupamento com o algoritmo *Fuzzy C-Means* foi realizada por meio do *software* R com o pacote `ppclust` e função `fcm`. Os pacotes: `factoextra`, `cluster` e `fclust` também foram utilizados para auxiliar no armazenamento dos resultados (FERRARO, GIORDANI, 2015).

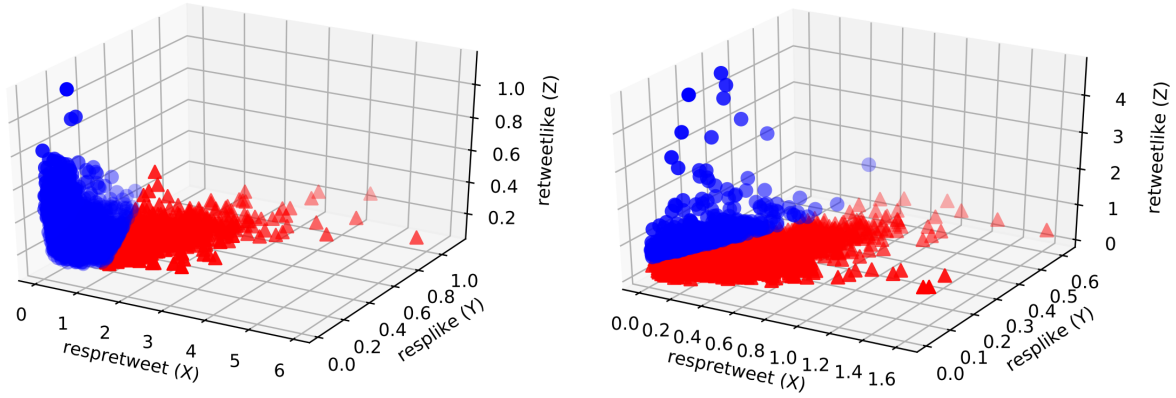
Ambos os algoritmos foram implementados para retornar dois grupos (K=2). Os resultados das análises para o algoritmo *K-Means* e *Fuzzy C-Means* estão na Tabela 1.

Tabela 1: Resultados do *K-Means* e *Fuzzy C-means* (K = 2).

K-Means	Total	0	1
Trump	3232	2319	913
Obama	3209	2792	416
Fuzzy	Total	0	1
Trump	3232	2264	968
Obama	3209	2266	942

Na Tabela 1 pode-se observar que o *K-Means* agrupou os tuítes das contas do Trump e Obama de formas diferentes. O grupo 1 para o Trump apresentou 913 tuítes, enquanto que para o Obama o mesmo grupo reuniu 416 tuítes. O grupo 0, reuniu 2319 e 2792 tuítes para o Trump e Obama respectivamente, sendo um grupo com a característica de conseguir obter mais de 2000 tuítes para ambas as contas.

Repetindo a mesma análise, porém agora com o algoritmo *Fuzzy C-Means*. O grupo 0 obtido pelo algoritmo *Fuzzy* para o Obama reuniu 942 tuítes e o 1 reuniu 2266 tuítes, enquanto que os mesmos grupos do *K-Means* obtiveram: 2792 e 416 tuítes respectivamente. Se comportando quase de forma oposta, quando comparado ao menor grupo de tuítes (416 para 942). Para os tuítes do Trump, o algoritmo *Fuzzy* aumentou a quantidade de tuítes no grupo 1 e diminuiu do grupo 0, conforme mostra a Tabela 1.



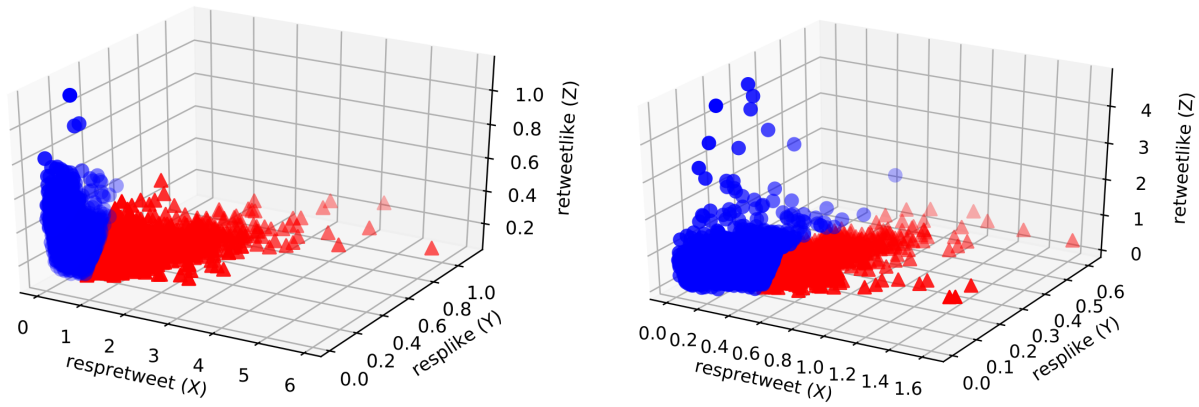
(a) (b)

Figura 2: Resultado do algoritmo K-Means para Donald Trump (a) e Barack Obama (b), com um K=2.

A Figura 2 apresenta a distribuição tridimensional dos grupos de tuítes obtidos pelo algoritmo *K-Means*. O grupo 1 para o Obama e Trump, presentes na Tabela 1, são representados pelo marcador triangular vermelho para o Trump, (Figura 2a) e circular azul para o Obama, (Figura 2b). Já o grupo 0, da Tabela 1 é representado pelo marcador circular azul para o Trump e triangular vermelho para o Obama. Esta mesma análise dos grupos foram realizadas pelo algoritmo *Fuzzy* e seus resultados podem ser observadas na Figura 3.

A divergência entre os algoritmos *K-Means* e *Fuzzy* no agrupamento dos tuítes do Obama, podem ser visualizados na Figura 2b e Figura 3b entre os marcadores circulares azuis, de forma que a Figura 3b, possui maior número de tuítes no grupo circular azul do que quando comparado a Figura 2b. Podendo ressaltar que a divergência entre os algoritmos foi em relação a quantidade em cada grupo, mas a região de agrupamento foi a mesma em ambos algoritmos para o Trump e Obama.

Vale ressaltar que os resultados dos grupos (0 e 1), presentes na Tabela 1, estão representados



(a) (b)

Figura 3: Resultado do algoritmo Fuzzy C-Means para Donald Trump (a) e Barack Obama (b), com um K=2.

de formas diferentes nas representações tridimensional dos resultados, como por exemplo o grupo 1 para o Trump é representado pelo marcador triangular vermelho e para o Obama o mesmo grupo está representado pelo marcador circular azul. Isto acontece, pelo fato dos algoritmos diferenciarem o formato de saída para cada base de dados.

Análise agrupamento K=3

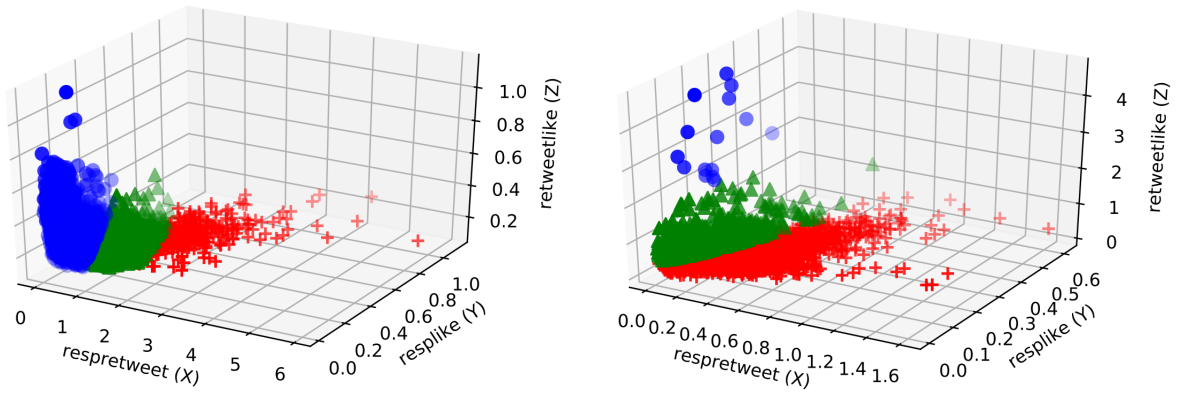
A Tabela 2 apresenta os resultados dos algoritmos *K-Means* e *Fuzzy*, com três grupos (0, 1 e 2) (K=3), para ambas as contas. Estes resultados mostraram novamente uma grande divergência entre os dois algoritmos para os tuítes do Obama, com diferenças expressivas entre os três grupos. O algoritmo *Fuzzy* aumentou o grupo 0 de 765 tuítes obtidos pelo *K-Means* para 958 tuítes. Diminui o grupo 1 de 2428 para 1703 tuítes e aumentou o grupo 2 de 15 para 547 tuítes, conforme mostra a Tabela 2.

Tabela 2: Resultados do *K-Means* e *Fuzzy-Cmeans* (K = 3).

K-Means	Total	0	1	2
Trump	3232	1790	1109	333
Obama	3209	765	2428	15
Fuzzy	Total	0	1	2
Trump	3232	1668	1164	400
Obama	3209	958	1703	547

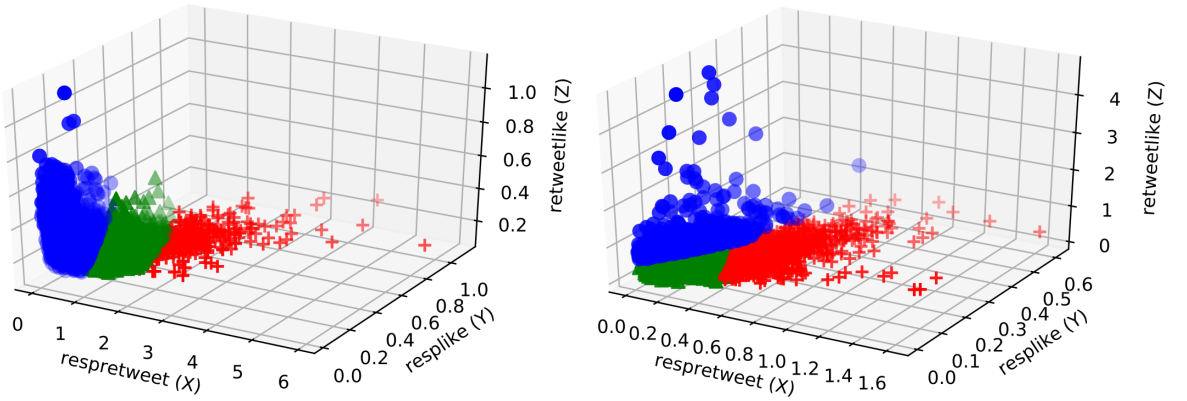
Já, para os tuítes do Trump, os dois algoritmos novamente apresentam semelhanças nas análises com os três grupos. Nos grupos 1 e 2 o *Fuzzy* apresentou 55 e 67 tuítes a mais respectivamente, do que o *K-Means*. No grupo 0 o *Fuzzy* apresentou redução de 122 tuítes quando comparados aos resultados do *K-Means*.

A representação gráfica tridimensional do agrupamento realizado com (K=3), pelos algoritmos são apresentadas nas Figuras 4 e 5. Os grupos (0, 1 e 2) presentes na Tabela 2, para o Trump, são representados pelos marcadores: circular azul, triangular verde e sinal de soma vermelho. Já para o Obama os mesmos grupos são representados pelos marcadores: triangular verde, e sinal de soma vermelho e circular azul. Na Figura 4 as 3 regiões que o algoritmo *K-Means* separa os tuítes do Trump e Obama possuem quantidades diferenciadas de tuítes, o que pode ser observado no eixo X representado por *respretweet*, onde possuem mais tuítes próximos do zero da escala do que do seu valor máximo.



(a) (b)

Figura 4: Resultado do algoritmo *K-Means* para Donald Trump (a) e Barack Obama (b), com um K=3.



(a) (b)

Figura 5: Resultado do algoritmo Fuzzy C-Means para Donald Trump (a) e Barack Obama (b), com um K=3.

Com um agrupamento com (K=3) a distribuição tridimensional para os tuítes do Trump com os dois algoritmos Figura 4a e Figura 5a são quase idênticas, mostrando a maior similaridade que os dois algoritmos tiveram com esses tuítes. Contudo, para os tuítes do Obama os algoritmos mostram uma divergência entre as suas análises, onde o *Fuzzy* aumenta o número de tuítes no grupo representado pelo marcador circular azul, quando comparado a análise do *K-Means*.

Discussão

Como o objetivo é analisar grupos de tuítes por sua popularidade baseado na teoria *The Ratio*, que leva em consideração que os tuítes são polêmicos e populares por apresentarem maior valor para as razões *respretweet* e *retweetlike* respectivamente. Este padrão pode ser observado nos eixos das distribuições dos tuítes do Trump e Obama, como mostra as Figuras 4 e 5, com a escala do eixo X sendo superior a do eixo Z para o Trump, enquanto que para o Obama a escala do eixo Z mostra-se superior a do eixo X.

Estes resultados são características que corroboram a hipótese de que Trump possui um discurso mais polêmico do que popular no Twitter, enquanto que o Obama possui um tom mais

Tabela 3: Comparação entre os resultados do *K-Means* e *Fuzzy-Cmeans* e as médias das variáveis dos tuítes do Barack Obama.

K-Means	Tuítes	respretweet	resplike	retweetlike
0	2792	0,304	0,119	0,417
1	416	0,123	0,118	0,822
0	765	0,154	0,106	0,687
1	2428	0,326	0,122	0,396
2	15	0,069	0,189	3,030
Fuzzy	Tuítes	respretweet	resplike	retweetlike
0	2266	0,357	0,136	0,387
1	942	0,209	0,133	0,686
0	958	0,505	0,210	0,428
1	1703	0,240	0,093	0,399
2	547	0,168	0,129	0,816

Tabela 4: Comparação entre os resultados do *K-Means* e *Fuzzy-Cmeans* e as médias das variáveis dos tuítes do Donald Trump.

K-Means	Tuítes	respretweet	resplike	retweetlike
0	2319	0,440	0,118	0,280
1	913	1,395	0,319	0,222
0	1790	0,344	0,093	0,322
1	1109	1,023	0,234	0,222
2	333	1,920	0,434	0,224
Fuzzy	Tuítes	respretweet	resplike	retweetlike
0	2264	0,430	0,119	0,322
1	968	1,508	0,337	0,225
0	1668	0,314	0,099	0,353
1	1164	0,985	0,225	0,231
2	400	1,957	0,436	0,224

popular em seu discurso.

Para facilitar a comparação entre os resultados dos dois algoritmos e sua interpretação, foram criados as Tabelas 3 e 4, com os resultados do agrupamento do Obama e Trump. O valor médio das razões para cada grupo são apresentados nestas tabelas, o que irá facilitar a interpretação prática dos valores que as razões possuem.

Para a primeira análise realizada definida com 2 grupos, o algoritmo *Fuzzy* apresentou respostas diferentes do *K-Means* para os tuítes do Obama, o que pode ser observado nos valores dos tuítes e nas médias das razões, presentes em cada grupo na Tabela 3. O *K-Means* com um (K=2) reuniu 416 tuítes com um valor médio de 0,822 para *retweetlike*, o que demonstra uma característica positiva para os tuítes deste grupo. Ao rodar o *Fuzzy* este valor para a variável caiu para 0,686 e o total de tuítes do grupo subiu para 942, demonstrando assim uma divergência entre as análises dos dois algoritmos. Esta divergência pode ser explicada pelo fato de que os tuítes do Obama, possuem semelhança no número de curtidas, respostas e retuítes, o que acaba dificultando os algoritmos em reuni-los em grupos diferentes.

Já o agrupamento realizado para os tuítes do Trump, apresentaram maior semelhanças entre os dois algoritmos, como mostra a Tabela 3, mostrando a primeira similaridade entre os algoritmos em identificar tuítes, baseado nas razões das métricas de engajamento.

Segundo Neil (2017), valores para a razão respostas/retuítes acima de 2, trazem um grande indício de que suas publicações possuem um caráter: discutível, duvidoso, problemático etc. Na Tabela 4 é possível notar que os grupos 1 possuem valores para *respretweet* maiores que

um. Enquanto os grupos 0 possuem para esta mesma variável valores menores que 0,5. Como os valores ainda são baixos, não é possível definir se este grupo de tuítes são controversos ou populares. Contudo é plausível a hipótese de que existe diferença entre a popularidade destes grupos, encontrados em ambos os algoritmos.

Comparando agora os resultados das análises dos algoritmos para 3 grupos, podemos destacar novamente uma divergência entre *K-Means* e *Fuzzy C-Means* com os tuítes do Obama. Na Tabela 3 pode-se observar o grupo 2 com 15 tuítes, presentes na Figura 4b em círculo azul, com uma razão de 3,03 para *retweetlike*, o que mostra que para cada curtida existem três compartilhamentos por tuítes, o que demonstra um maior nível de aceitabilidade pelos seguidores. Contudo na análise com o *Fuzzy* este grupo não manteve esta mesma característica, passando a ter 547 tuítes com valor médio de 0,816 para a mesma variável.

Porém, novamente os dois algoritmos apresentaram similaridades nos resultados, com três grupos para os tuítes do Trump, que podem ser vistos na Tabela 4. Em destaque o grupo 2, que apresenta valor médio acima de 1,9 para a razão *respretweet*, o que nos traz a ideia de que para cada compartilhamento existem duas respostas para cada tuíte presentes neste grupo. Um maior número de respostas no *Twitter* traz um indício de que o conteúdo pode ser polêmico. O que está sendo comum para os tuítes do Trump no período de 2016 a 2017 presentes nesta base de dados.

Uma forma de identificar se existe diferença entre os tuítes presentes em cada grupo, é por meio da medida de distância entre os centros dos grupos. Se a distância entre os grupos for nula, significa que os tuítes possuem as mesmas características (ANSARI, 2020). Contudo se a distância entre os grupos for maior isso demonstra uma diferença entre as características das variáveis e dos tuítes. O que pode ser facilmente interpretado como sendo uma grande variedade na popularidade dos tuítes presentes nos grupos (SINAGA; YANG, 2020). Estas medidas foram obtidas pelo algoritmo *Fuzzy C-Means* e são apresentadas nas Tabelas 5 e 6.

Tabela 5: Distância entre grupos *Fuzzy-Cmeans* (K=2).

	Obama	Trump
	0	0
1	0,112	1,224

Tabela 6: Distância entre grupos *Fuzzy-Cmeans* (K=3).

	Obama		Trump	
	0	1	0	1
1	0,085	0	2,831	0
2	0,271	0,179	0,481	0,990

Logo, na primeira análise com dois grupos, 0 e 1 pelo *Fuzzy*, pode-se notar a diferença entre as distâncias dos centros dos grupos entre os dados do Trump e Obama. Entre o 0 e 1 dos tuítes do Trump a distância foi de 1,224 o que significa uma maior diferença entre as razões para estes dois grupos. Para os tuítes do Obama a distância medida entre os dois grupos foi de 0,112, o que significa maior similaridade entre as razões dos dois grupos. Estas medidas são corroboradas pelo valor médio das razões apresentadas nas Tabelas 3 e 4.

Para a análise com 3 grupos realizadas pelo *Fuzzy C-Means* a distância entre os grupos para os tuítes do Trump são ainda maiores, obtendo valor de 2,831 para a distância entre o grupo 0 e 1, presentes na Tabela 5, o que traz a ideia de que estes grupos de tuítes possuem diferenças entre as razões, com grande probabilidade de serem opostos com relação a popularidade. Para os tuítes do Obama, o agrupamento com 3 grupos não obteve nenhuma distância expressiva entre os grupos.

Conclusão

Com o objetivo de realizar uma análise de *clustering*, baseados nas razões das métricas de tuítes, com o intuito de encontrar grupos com características populares e polêmicas, esta análise foi bem sucedida por meio dos algoritmos *K-Means* e *Fuzzy C-Means*. Pode-se concluir que houve diferenças e similaridades entre as análises de agrupamento realizadas pelos dois algoritmos utilizados, para os tuítes do Obama e Trump. Mesmo com a divergência nos resultados para os tuítes do Obama, foi possível notar que o Trump apresentou tuítes mais polêmicos quando comparados aos do Obama, com um maior valor para a razão respostas/retuítes. Já os tuítes do Obama, em sua grande maioria não apresentaram grandes valores para esta razão. Os tuítes do Obama, apresentaram somente na análise do *K-Means* o grupo 2 de tuítes com maior valor para a razão compartilhamento/curtidas. O algoritmo *K-Means* se mostrou melhor em localizar pequenos grupos de tuítes semelhantes, quando comparado ao *Fuzzy C-Means*. Contudo ambos os algoritmos foram capazes de realizar o agrupamento dos tuítes por meio das razões entre as métricas. Trabalhos futuros poderão repetir esta análise em novos tuítes de outras contas do *Twitter* e também de outras redes sociais, bem como estudar o conteúdo textual dos tuítes classificados como polêmicos e populares.

Acknowledgements

Agradeço a FAPEMIG pelo financiamento, ao Prof Dr. Ricardo Menezes Salgado em ministrar a disciplina de Ciências dos Dados e ao meu orientador Prof Dr. Eric Batista Ferreita.

References

ANGER, I.; KITTL, C. Measuring Influence on Twitter ACM International Conference Proceeding Series, p. 4-7, 2011

ANSARI, M, Z. Analysis of Political Sentiment Orientations on *Twitter*, ScienceDirect, v.167, n.1, p. 1821-1828, 2020.

DataReportal (2020), "Digital 2021 Global Digital Overview," retrieved from https://datareportal.com/reports/digital-2021-global-digital-overview

FERRARO, M. B., GIORDANI, P. A toolbox for fuzzy clustering using the R programming language. Fuzzy Sets and Systems, 279, 1–16, 2015.

GARG, Neha; RANI, Rinkle. Analysis and visualization of Twitter data using k-means clustering. In: 2017 International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2017. p. 670-675.

ISAAC, M . 2018. The Ratio Establishes Itself on *Twitter* https://www.nytimes.com/interactive/2018/02/09/technology/the-ratio-trends-on-*Twitter*.html

KUŠEN, Ema; STREMBECK, Mark. Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections. Online Social Networks and Media, v. 5, p. 37-50, 2018.

MINOT, J. R. et al. Ratioing the President: An exploration of public engagement with Obama and Trump on Twitter. PloS one, v. 16, n. 4, p.e0248880, 2021.

MULIAWATI, T.; MURFI, H. Eigenspace-based *Fuzzy C-Means* for sensing trending topics in *Twitter*. AIP Conference Proceedings, p. 1-7, 2017.

NEIL, L. 2017. How to Know if you've sent a horrible tweet
https://www.esquire.com/news-politics/news/a54440/ *Twitter*-ratio-reply/

OKTARINA, C.; NOTODIPUTRO, K.; INDAHWATI, I. Comparison of *K-Means* Clustering method and K-Medoids on *Twitter* data. Indonesian Journal of Statistics and Its Applications, v.4, n.1, p. 189-202, 2020.

ROTH, D. 2017 The Ratio is the triple crown of bad tuítes
https://deadspin.com/the-ratio-is-the-triple-crown-of-bad-tuítes-1798441271

ROEDER, O. et al. 2017. The Worst Tweeter In Politics Isn't Trump
https://fivethirtyeight.com/features/the-worst-tweeter-in-politics-isnt-trump/

RIQUELME, F.; CANTERGIANI, P. Measuring user influence on *Twitter*: A survey Information Processing and Management, v52, n.5, p. 949-975, 2016.

SINAGA, K, P.; YANG, M. Unsupervised *K-Means* Clustering Algorithm. IEEE Access v.8, n.1, p. 80716-81727, 2020.

SUNGE, A S. Analysis of Popularity Sentiment in Opinion Presidential Election 2019 on Twitter. 2021.

YAQUB, Ussama et al. Analysis of political discourse on twitter in the context of the 2016 US presidential elections. Government Information Quarterly, v. 34, n. 4, p. 613-626, 2017.

YE, Shaozhi; WU, S. Felix. Measuring message propagation and social influence on Twitter. com. In: International conference on social informatics. Springer, Berlin, Heidelberg, p. 216-231, 2010.

ZADEH, L, A. ABBASOV, S, A Analysis of Twitter hashtags: Fuzzy clustering approach, Annual Conference of the North American Fuzzy Information Processing Society - NAFIPS, sept 2015.