UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Tecnologia

Mariana Albino de Queiroz Andrade Antonio

Análise Exploratória dos Sentimentos das Notícias e a Reação dos Usuários no Facebook

Limeira
2021

Mariana Albino de Queiroz Andrade Antonio

Análise Exploratória dos Sentimentos das Notícias e a Reação dos Usuários no Facebook

Monografia apresentada à Faculdade de Tecnologia da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Bacharela em Sistemas de Informação.

Orientador: Prof. Dr. Ulisses Martins Dias

Este exemplar corresponde à versão final da Monografia defendida por Mariana Albino de Queiroz Andrade Antonio e orientada pelo Prof. Dr. Ulisses Martins Dias.

Limeira
2021

FOLHA DE APROVAÇÃO

Abaixo se apresentam os membros da comissão julgadora da defesa de dissertação para o Título de Bacharela em Sistemas de Informação na área de concentração, a que se submeteu a aluna Mariana Albino de Queiroz Andrade Antonio, em 14 de janeiro de 2021 na Faculdade de Tecnologia – FT/UNICAMP, em Limeira/SP.

Prof. Dr. Ulisses Martins Dias
Presidente da Comissão Julgadora

Prof. Dr. Plínio Roberto Souza Vilela
FT/UNICAMP

Profa. Dra. Ieda Geriberto Hidalgo
FT/UNICAMP

Ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós Graduação da FT.

Resumo

Esta monografia é uma análise exploratória das reações e os sentimentos encontrados em textos de notícias publicadas no *Facebook*. Os dados, compostos basicamente de textos e números de reações dadas por usuários, são pré-processados para omitir ruídos e dados irrelevantes, e processados de forma que estejam nas formas corretas para a análise. As análises são divididas em duas etapas: a primeira tem uma perspectiva mais particular em relação ao conteúdo de notícias, incluindo as reações dadas nas notícias, seus textos e o sentimento gerado. A segunda tem uma visão mais geral das notícias e seus sentimentos, sem considerar as reações dadas. Ambas as etapas usam o algoritmo de clusterização K-Means, sendo K escolhido usando do Método do Cotovelo. Cada palavra pertencente a um *cluster* é processada pelo analisador de sentimento VADER. Os resultados sugerem *insights* sobre o sentimento gerado para cada *cluster*, a ligação entre reações e sentimentos dados a notícias e assuntos mais frequentes em notícias.

Palavras-chave: Clusterização; Análise de Sentimentos; Notícias; Facebook.

Abstract

This monograph is an exploratory analysis of the reactions and feelings found in news texts published on *Facebook*. The data, basically composed of texts and number of reactions given by users, are pre-processed to omit noise and irrelevant data, and processed so that they are in the correct form for the analysis. The analyzes are divided into two stages: the first has a more particular perspective in relation to the news content, including the reactions given in the news, its texts and the sentiment generated. The second has a more general view of the news and its given feelings, not considering the reactions given. Both steps use the K-Means clustering algorithm, which K being chosen using the Elbow Method. Each word belonging to a *cluster* is processed by the VADER sentiment analyzer. The results suggest insights into the sentiment generated for each *cluster*, the link between reactions and feelings given to the news and most frequent news topics.

Keywords: Clustering; Sentiment Analysis; News; Facebook.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ABC	Artificial Bee Colony
EM	Expectativa-Maximização
PCA	Principal Component Analysis
SVD	Singular Value Decomposition
TF-IDF	Term Frequency–Inverse Document Frequency
VADER	Valence Aware Dictionary and sentiment Reasoner

Sumário

Capítulo 1

Introdução

Canais de comunicação usam das redes sociais para divulgar notícias que são posteriormente compartilhadas pelos usuários, esses que, além de publicarem sobre suas vidas pessoais, divulgam e opinam acerca das notícias. O *Facebook* [1], rede social mais usada no mundo inteiro segundo Statista (2020), acessado por volta de 66% da população brasileira e 73% da população norte-americana de acordo com Usage e Statistics (2020), faz parte deste compartilhamento de notícias na internet.

Com as possibilidades de reagir e comentar publicações compartilhadas no *Facebook*, a rede tem se tornado um espaço para protagonização de conflitos pessoais, muitas vezes por posicionamento político, ou mesmo por qualquer outra razão.

Apesar de ainda ser a rede social mais usada, o *Facebook* tem perdido usuários no Brasil. Em um estudo do Instituto Qualibest (2019), o *Facebook* foi a rede social que mais perdeu usuários em comparação a *Instagram* e *Youtube*, 46% dos entrevistados diminuíram a frequência na rede, 81% concordam que a rede é mais usada para se expor o que pensa, e 50% alegam ser a rede com mais postagens chatas.

O conflito gerado em discussões relativas às postagens no *Facebook* leva a um questionamento sobre a qualidade do ambiente desta rede social: essas discussões fazem sentido com o conteúdo da postagem? Quando se trata de notícias, as pessoas tendem a reagir mal apenas à notícias ruins, ou conflitos são gerados independente do tipo de notícia?

Com informações à mão que indicam o *Facebook* como uma rede social que tem se tornado desagradável, porém, ainda assim muito utilizada, uma análise sobre o ambiente da rede social pode tornar-se interessante.

[1] https://www.facebook.com/

1.1 Objetivo

O intuito desta monografia é, por meio de uma análise exploratória, usando de análise de sentimento, clusterização e relacionamento de informações através da visualização dos dados, realizar uma análise da qualidade do ambiente dentre notícias postadas no Facebook.

A clusterização combinada com a análise de sentimento formam a parte principal do trabalho, que exibirá o que há em comum entre as notícias mais compartilhadas, e qual o sentimento de cada conjunto. Todos os passos e componentes dessa análise exploratória serão explicados no decorrer da monografia.

1.2 Organização da Monografia

A motivação da monografia, como visto, foi apresentada neste capítulo, que será seguido pelo Levantamento Bibliográfico, no Capítulo 2, mostrando referências que foram usadas para a realização do trabalho.

As ferramentas, bibliotecas, métodos, procedimentos e a ordem em que foram empregados no trabalho são apresentados em Materiais e Métodos, no Capítulo 3. A forma como foram empregados e o resultado obtido apresentam-se no decorrer do Capítulo 4. As conclusões do trabalho estão no Capítulo 5.

Capítulo 2

Levantamento bibliográfico

Para a realização deste estudo, foi realizado um levantamento bibliográfico de materiais inseridos no mesmo contexto, aqui citados em ordem cronológica de acordo com suas respectivas datas de publicação.

Lilian Lee, Pang e Vaithyanathan (2002) descobriram que as técnicas de aprendizado de máquina padrão superam os patamares produzidos pelo homem. Contudo, as técnicas conhecidas como *Naive Bayes*, classificação de Entropia Máxima e Máquinas de Vetores de Suporte não performam tão bem na classificação de sentimento quanto a categorização baseada em tópicos tradicionais.

Sabe-se que métodos de aprendizagem semi-supervisionada constroem classificadores utilizando amostras de dados de treinamento rotuladas e não rotuladas. Mesmo que as últimas ajudem a melhorar a precisão dos modelos treinados, ainda há dificuldades quando os dados rotulados não são suficientes e enviesados em relação à distribuição de dados subjacentes. Dessa maneira, Zeng et al. (2003) concluíram que um *cluster* baseado em abordagem de classificação (CBC) supera os algoritmos já existentes quando o tamanho do conjunto de dados rotulado é muito pequeno.

Yessenov e Misailovic (2009) optaram por estudar a eficácia das técnicas de aprendizado de máquina em classificar mensagens de texto por significado semântico. Para tal, utilizaram o modelo conhecido como "saco de palavras" (ou " *Bag of Words*") e avaliaram seu efeito nos métodos *Naive Bayes*, Árvores de Decisão, Entropia Máxima e Clusterização K-Means.

Sob a ótica de Li e Fei Liu (2010), em sua pesquisa, a análise de sentimento baseada em clusterização é mais vantajosa que técnicas simbólicas e métodos de aprendizagem supervisionada.

Em relação à utilização de *emojis* como forma de comunicação no mundo *online*, Tian et al. (2017) apresentaram um extenso conjunto de dados de postagens do *Facebook* de páginas da mídia pública de quatro países, resultado de um estudo sobre o sentimento transmitido por um *emoji*.

Riaz et al. (2017) realizaram um estudo pelo viés do desafio que as empresas enfrentam na análise da grande quantidade de dados para extrair e avaliar a opinião dos clientes. Para enfrentar isso, fizeram a análise de sentimento sobre os dados do mundo real das avaliações do clientes para descobrir suas preferências, analisando expressões subjetivas. Por fim, compararam os resultados com a classificação por estrelas atribuídas pelos próprios usuários e encontraram uma mudança drástica nos resultados. Além disso, forneceram uma representação visual dos resultados para auxiliar na tomada de decisão.

De acordo com Sisi Liu e Ickjai Lee (2018), a análise de sentimento de *e-mail* não foi estudada e examinada exaustivamente, embora seja um dos meios de comunicação mais onipresentes na contemporaneidade. Dessa maneira, propuseram uma análise de sentimento híbrida, com técnicas de Frequência do Termo - Frequência Inversa do Documento ou Term Frequency–Inverse Document Frequency (TF-IDF) para extração de características, clusterização e rotulamento com o modelo K-Means e classificador SVM para classificação de sentimentos. A conclusão desse estudo indica resultados de classificação comparativamente melhores com a estrutura proposta do que outras combinações.

Recentemente, a análise de sentimento forneceu novos caminhos para entender a ligação entre as mídias sociais e os eleitores. A pesquisa de Sandoval-Almazán e Valle-Cruz (2018) explora os dados coletados de postagens do *Facebook* e seus *emoticons* que revelam algum tipo de sentimento dos usuários em relação a uma campanha política no México. Os resultados obtidos mostram que os partidos políticos têm um grande impacto com poucas postagens, mas, em geral, este artigo revela que a percepção dos eleitores de candidatos no *Facebook* é ruim para o partido político vencedor, pois apesar dessa situação, o partido político com melhor impacto de sentimento não poderia ganhar as eleições.

Enquanto isso, Korovkinas, Danėnas e Garšva (2019) sugeriram uma técnica híbrida a fim de melhorar a precisão da classificação da máquina de vetores de suporte por meio do emprego da clusterização para selecionar dados de treinamento e ajuste dos parâmetros. Sua conclusão foi a obtenção de melhores resultados ao utilizar este método proposto, em comparação aos resultados do método apresentado em seu trabalho anterior.

Rodriguez, Argueta e Chen (2019) preferiram se aprofundar no tema da análise de sentimento no que diz respeito à problemática dos discursos de ódio amplamente disseminados nas mídias sociais. Assim, empregaram técnicas de análise de sentimento, emoção e grafos a fim de agrupar e analisar postagens em páginas do *Facebook*. Tal proposta é capaz de identificar quais páginas promovem o discurso de ódio nas seções de comentários sobre tópicos delicados.

A análise das percepções e dos sentimentos dos movimentos antivacinas nas mídias sociais foi feita à luz de Garay, Yap e Sabellano (2019). Para isso, foi utilizada a ferramenta de análise de sentimento conhecida como Valence Aware Dictionary and sentiment Reasoner (VADER). Segundo esse estudo, para avaliar os resultados do K-means, a pontuação da silhueta é determinada para indicar a que distância um ponto está de outros *clusters* próximos e, nesse artigo, houve o indício de que tais pontos estão próximos dos limites de decisão.

Giuntini et al. (2019) também estudaram a análise de sentimentos em relação ao *emojis* utilizados por internautas nas redes sociais. Para tal, implantaram um algoritmo de Expectativa-Maximização (EM), e os resultados encontrados permitiram determinar a polaridade e a correlação das reações com as *emoticons*. Isso sugere que o uso de reações em algoritmos de análise de sentimentos pode aumentar a confiança na determinação da emoção que o conteúdo reflete e do estado emocional dos internautas.

Por fim, Orkphol e Yang (2019) deliberaram acerca da análise de sentimentos de *microblogging*, um tipo de blog usado por pessoas para expressar suas opiniões em relação a entidades com uma curta mensagem, pois ter o conhecimento de tais sentimentos seria benéfico para a tomada de decisão, planejamento e visualização. Dessa maneira, utilizaram o método K-means, técnicas de *TF-IDF*, decomposição de valor singular ou Singular Value Decomposition (SVD) e Colônia Artificial de Abelhas ou Artificial Bee Colony (ABC), algoritmo inspirado em colônias de abelhas para otimização numérica. Essa pesquisa concluiu que a combinação de várias técnicas pode melhorar significativamente o resultado do K-means.

Todos os trabalhos citados nesse capítulo serviram de inspiração, apoio ou referência, por serem de alguma forma relacionados a esta monografia. Boa parte dos trabalhos usa clusterização, análise de sentimentos, dados coletados de redes sociais em geral, ou quaisquer outras ferramentas que também neste foram aplicadas.

Capítulo 3

Materiais e Métodos

Apresentado o levantamento bibliográfico para o desenvolvimento desta monografia, os materiais e métodos utilizados são discutidos neste capítulo. Na primeira seção há um alinhamento do conteúdo e organização do trabalho, seguido pela Seção 3.2 que indica a fonte dos dados analisados, e, em seguida, a Seção 3.3, indicando as ferramentas empregadas para a análise. As Seções 3.4, 3.5, 3.6 e 3.7 compõem o processamento dos dados, seguido pela explicação sobre Análise de Sentimento, na Seção 3.8.

As Seções 3.9, 3.10 e 3.11 discutem sobre auxiliares para padronização, vetorização e otimização dos dados para uso posterior. Na Seção 3.12 é mostrado o Método Cotovelo e seu uso para a clusterização. Antes de finalizar, passamos pela Seção 3.13 que denota como os dados são transformados para serem utilizados para a clusterização, esta que fecha este capítulo na Seção 3.14.

3.1 Visão Geral

Este trabalho visa utilizar as notícias captadas de postagens no Facebook de páginas de canais de comunicação importantes, como *CNN, New York Times, Fox News etc.,* seguindo um script como explicado na Seção 3.2, para o estudo de análise de sentimentos, clusterização e todas as transformações necessárias aplicadas ao *dataset* com a finalidade de analisar informações e gráficos provenientes dos processos aplicados.

A abordagem proposta da análise foi arquitetada em dois *pipelines*, demonstrados nas figuras que seguirão. Para os dois *pipelines* são utilizados os mesmos processos, se

diferenciando na ordem em que agem e suas finalidades para as análises. O intuito desta abordagem é a comparação dos processos e suas análises relativas.

As etapas são principalmente compostas por: pré-processamento dos dados, processamento dos dados, cálculo de Análise do Componente Principal ou Principal Component Analysis (PCA), criação e análise de gráfico usando o Método do Cotovelo, vetorização de texto, cálculo de frequência de documento inverso, aplicação de análise de sentimentos e clusterização. O pré-processamento dos dados se dá por escolha do tamanho da fração do *dataset*, suas colunas e retirada de espaços em branco e termos desnecessários. O processamento de dados é constituído por remoção de pontuação, tokenização, remoção de *stop words* e lematização.

Para a primeira etapa, é realizado o *pipeline* descrito na Figura 3.1. Todos os conteúdos nos blocos das figuras serão apresentados ainda neste capítulo.

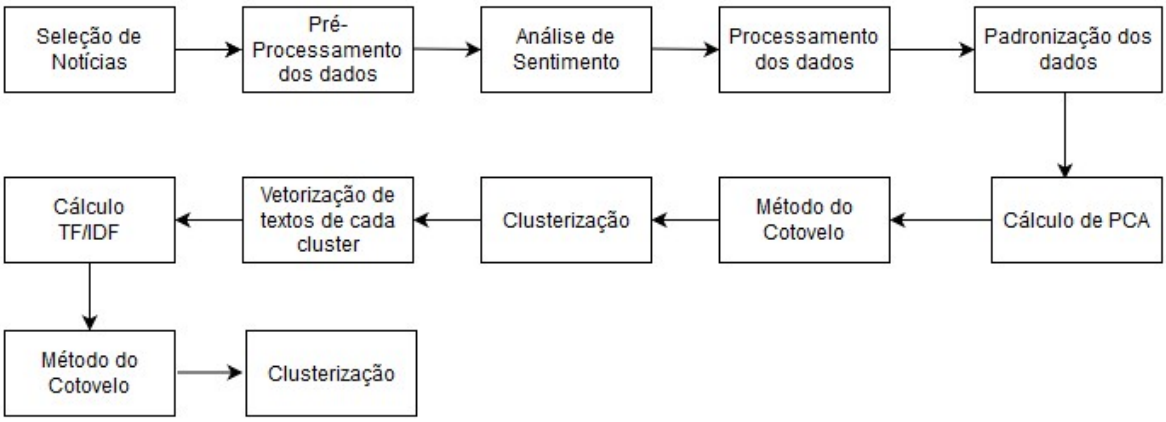

Figura 3.1: Pipeline da Primeira Etapa.

Para a segunda etapa, é realizado o *pipeline* descrito na Figura 3.2.

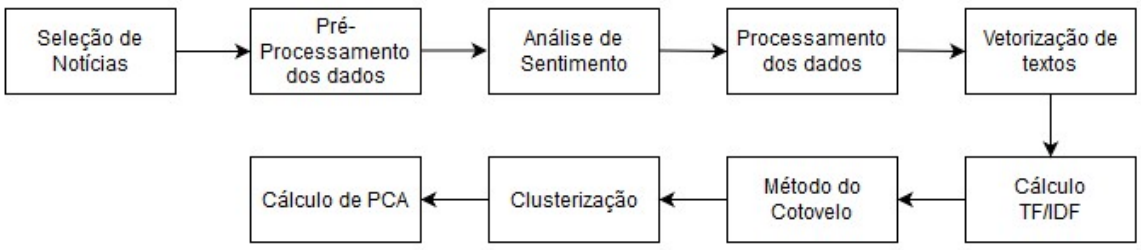

Figura 3.2: Pipeline da Segunda Etapa.

As ferramentas e métodos utilizados são explicados no decorrer deste capítulo. Os processos e seus motivos serão explicados ao longo do Capítulo 4.

3.2 Seleção das notícias

Os dados utilizados nesta análise são oriundos de um *dataset* criado pelo Cientista de Dados Sênior John Bencina, que criou um script para a obtenção de publicações de notícias no Facebook. Esse script percorre um dicionário de *ids* de páginas do Facebook (como o portal da *CNN*, *Fox News*, *New York Times* etc.) e recupera as últimas *N* postagens e até 100 comentários para cada postagem. Os resultados são opcionalmente armazenados como arquivos de dados individuais e, por fim, armazenados em tabelas: uma para notícias, e uma para comentários. Eles podem ser vinculados pelo campo *post_id* comum (BENCINA, 2020). Neste trabalho, será utilizada a tabela de notícias.

A tabela de notícias contém 19,850 linhas e 14 colunas, e a tabela de comentários tem 1,025,403 linhas e 5 colunas. As colunas escolhidas da tabela de notícias serão discutidas posteriormente. As notícias foram coletadas em 14 de julho de 2017. (BENCINA, 2020)

3.3 Ferramentas

O código da análise foi escrito e compilado no Jupyter Notebook, usando de um processador Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz 2.60 GHz, 8,00 GB de RAM e Sistema Operacional Windows 10 de 64 bits. As bibliotecas utilizadas (todas da linguagem Python) e as versões se apresentam na seguinte tabela:

Tabela 3.1: Bibliotecas utilizadas.

Biblioteca	Versão
Matplotlib	3.3.1
NLTK	3.5
NumPy	1.19.1
Pandas	1.0.5
Python	3.7.1
ScikitLearn	0.23.2
Seaborn	0.11.0
vaderSentiment	3.3.2

3.4 Remoção de pontuação

As pontuações contidas nos textos são removidas visando um processamento apenas de palavras para análise de sentimento. Para futuras transformações, pontuações seriam apenas um ruído nos dados.

3.5 Tokenização

Tokenização é o processo em que se quebra um texto em pedaços, denominados *tokens*. Para o NLP, o *token* é a menor unidade de texto que a máquina entende (BURNS, 2019). Os textos em análise foram tokenizados a nível de palavras, sendo cada palavra um *token*.

3.6 Remoção de stop words

Stop words são palavras neutras comumente utilizadas e normalmente ignoradas pelo fato de serem usadas com frequência. Por exemplo: *the, in, on* etc. (BURNS, 2019). Por gerarem ruído entre as informações importantes e não fazerem diferença numa análise de sentimento, removemos esse tipo de palavra dos textos.

3.7 Lematização

Lematização é o processo aplicado a cada *token*, a fim de encontrar a raíz da palavra, também conhecida como *lemma*, colocando em consideração o contexto da palavra na frase (BURNS, 2019). Por exemplo: o *lemma* das palavras *"playing", "played"* e *"plays"* é *"play"*.

3.8 Análise de Sentimento

Análise de sentimento é um ramo da NLP que lida com identificação e extração de opiniões de um texto. O objetivo é medir os sentimentos, emoções, e o que mais se referir ao campo de tais expressões por quem falou ou escreveu (BURNS, 2019).

Com o propósito de procurar por uma relação entre o conteúdo sentimental das notícias e as reações dadas por usuários para estas, a análise de sentimento é feita por meio da ferramenta *VADER*.

VADER significa *Valence Aware Dictionary and sentiment Reasoner*. É uma ferramenta de análise de sentimento baseada em regras lexicais para analisar sentimentos provenientes de redes sociais. *VADER* não só indica se o texto é positivo, negativo ou neutro, mas o quanto o texto é positivo, negativo ou neutro. Cada uma dessas pontuações mostra a proporção de texto que se enquadra nessas categorias. Todas essas pontuações devem somar 1. A pontuação composta (ou *compound*, termo apresentado no uso da ferramenta) denota uma métrica que calcula a soma de todas as classificações do léxico com normalização entre -1 (o negativo mais extremo) e +1 (o positivo mais extremo). A ferramenta funciona apenas para textos em Inglês, para outras línguas é necessário que o texto seja traduzido antes (BURNS, 2019; HUTTO; GILBERT, 2014; RODRIGUEZ; ARGUETA; CHEN, 2019).

Para calcular as porções sentimentais do texto, *Vader* verifica os textos das chamadas das notícias em busca de características sentimentais conhecidas, modifica a intensidade e a polaridade de acordo com suas regras, soma as pontuações das características encontradas no texto e normaliza a pontuação final entre -1 e 1 usando a função:

$$\frac{x}{\sqrt{x^2 + \alpha}}$$

onde α é definido como 15, o que se aproxima do valor máximo esperado de x.

Texto	Saída do VADER
Hey, Sabrina! I really hope you have a wonderful day!!	{'neg': 0.0, 'neu': 0.436, 'pos': 0.564, 'compound': 0.8303}

Figura 3.3: Exemplo de entrada de texto e saída do *VADER*.
Fonte: Própria (2020).

3.9 Padronização

A padronização é a técnica que dimensiona cada variável de entrada separadamente, subtraindo a média (chamada de centralização) e dividindo pelo desvio padrão para mudar a distribuição e, assim, obter média zero e desvio padrão um.

A ferramenta utilizada para padronização é a classe *StandardScaler* da biblioteca *scikit-learn*, que padroniza uma característica ou série de características subtraindo a média e, em seguida, escalando para a variância da unidade. A variância da unidade significa dividir todos os valores pelo desvio padrão.

StandardScaler resulta em uma distribuição com um desvio padrão igual a 1. A variância também é igual a 1, pois variância é igual ao desvio padrão ao quadrado. Ocorre então a média da distribuição igual a 0. Cerca de 68% dos valores estarão entre -1 e 1.

Normalização:

$$\zeta = \frac{x - \mu}{\sigma}$$

Com a média:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} (xi)$$

E desvio padrão:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (xi - \mu)^2}$$

Escala min-max:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

3.10 Vetorização de Texto

A vetorização de texto tokeniza o texto e retorna um vetor codificado com um comprimento de todo o vocabulário e uma contagem inteira para o número de vezes em que cada palavra apareceu no documento.

O vocabulário de palavras conhecidas é formado, o qual também é utilizado para codificar texto não visto posteriormente.

A ferramenta utilizada para vetorização de texto é a classe CountVectorizer da biblioteca scikitlearn. Abaixo está a representação da vetorização de texto para a frase *"The quick brown fox jumps over the lazy dog"*.

The	quick	brown	fox	jumps	over	lazy	dog
2	1	1	1	1	1	1	1

Figura 3.4: Exemplo de saída do *CountVectorizer*.

3.11 TF-IDF

Tf significa *term-frequency* (frequência de termo), enquanto *tf-idf* significa *term-frequency times inverse document-frequency* (frequência de termo vezes frequência de documento inversa). Este é um esquema de ponderação de termos comum na recuperação de informações, que também tem bom uso na classificação de documentos.

O objetivo de utilizar *tf-idf* ao invés das frequências brutas de ocorrência de um *token* em um determinado texto é reduzir o impacto dos *tokens* que ocorrem com muita frequência em uma determinada coleção de documentos textuais, e que são, portanto, menos informativos do que os recursos que ocorrem em um pequena fração da coleção de documentos textuais de treinamento.

Para um termo t em um documento d, a equação para o cálculo é: $tf - idf(t, d) = tf(t, d) \times idf(t)$ sendo idf calculado como: $df(t) = \log [n/df(t)] + 1$, onde n é o número total de documentos no conjunto de documentos e $df(t)$ é a frequência de documentos de t. É adicionado 1 caso haja termos com idf igual a 0, assim não serão ignorados.

3.12 Método do Cotovelo

O "Método do Cotovelo" ou *"Elbow Method"* é usado para a definição da quantidade de clusters na clusterização de um conjunto de dados. É um método visual de análise que consiste em visualizar um gráfico que comece com $K = 2$ em sua coordenada abscissa, e vá aumentando em passos de 1, calculando seus *clusters* e o custo associado ao treinamento. Em algum valor de K, o custo cai drasticamente e, depois disso, atinge um platô. Usamos o valor de K indicado na queda do gráfico para o número de clusters escolhido.

Para a imagem do exemplo, o número de K escolhido seria o número 6.

3.13 PCA

Inventado por Pearson (1901) e Hotelling (1933), o *Principal Component Analysis* (PCA) ou Análise do Componente Principal, é um método para redução de dados que contém muitas variáveis por um conjunto menor de variáveis derivadas a partir do conjunto original com uma perda mínima de informação.

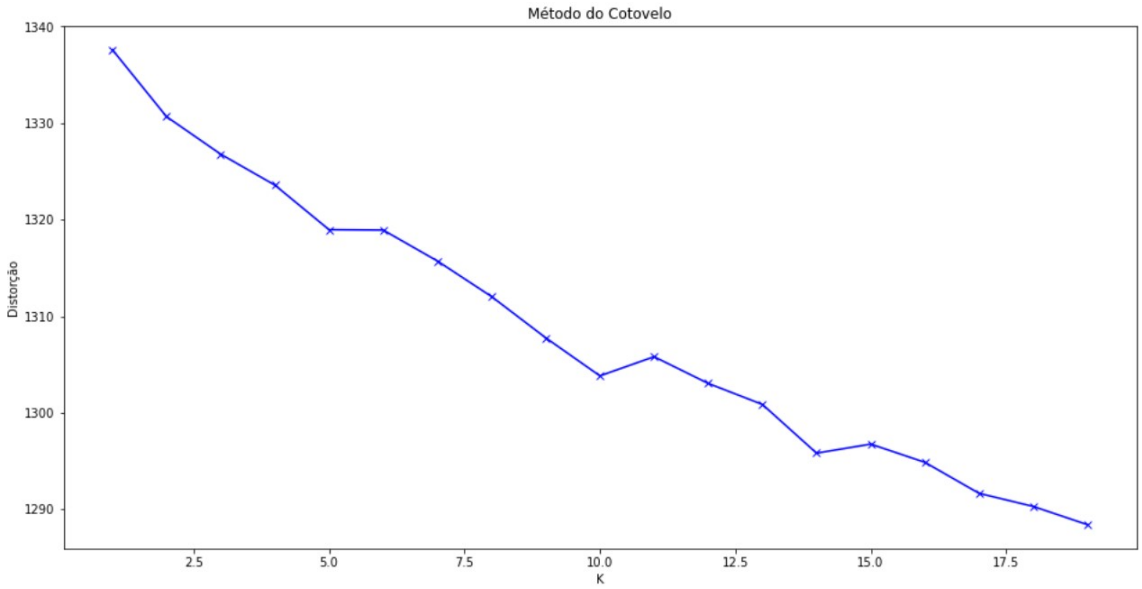

Figura 3.5: Exemplo de gráfico usando Elbow Method.
Fonte: Própria (2020).

O número de componentes principais se torna o número de variáveis consideradas na análise, mas geralmente as primeiras componentes formadas são as mais importantes, já que explicam a maior parte da variação total.

Os componentes principais em geral são extraídos via matriz de covariância, mas essa extração também pode ocorrer via matriz de correlação (DUNTEMAN, 1989).

3.14 Clusterização

A Clusterização de Dados tem por objetivo agrupar automaticamente por aprendizado não supervisionado os n casos da base de dados em k grupos, geralmente denominados *clusters* ou agrupamentos.

O uso da técnica de clusterização traz benefícios ao agrupar dados similares, podendo descrever de forma mais eficiente e eficaz as características peculiares de cada um dos grupos identificados, também fornecendo um maior entendimento do conjunto original de dados.

O algoritmo de clusterização escolhido é o K-Means, que procura por um número pré-determinado de clusters em um conjunto de dados multidimensional não rotulado. O "centro do *cluster*" é a média aritmética de todos os pontos pertencentes ao *cluster*. Cada ponto está mais próximo do centro de seu próprio *cluster* do que do centro dos outros *clusters*.

O K-Means envolve uma abordagem iterativa intuitiva conhecida como maximização da expectativa, que presume o centro dos *clusters* antes de tudo, e, posteriormente, se divide em duas etapas que se repetem até convergir: a "Etapa E" ou "Etapa de expectativa" (*Expectation step*), que é assim denominada pois envolve a atualização de nossa expectativa referente a qual cluster cada ponto pertence, e a "Etapa M" ou "Etapa de maximização" (*Maximization step*), que realiza a maximização de alguma função de aptidão que define a localização dos centros do cluster - neste caso, essa maximização é realizada tomando uma média simples dos dados em cada cluster (VANDERPLAS, 2016).

Capítulo 4

Desenvolvimento

4.1 Pré-processamento dos dados

Antes de começar o tratamento dos dados, é feita uma escolha das colunas utilizadas para a análise, assim, durante as etapas de processamento e análise, há menos gasto computacional desnecessário. Da tabela de notícias, são selecionadas 5000 linhas e as colunas demonstradas na seguinte tabela:

Tabela 4.1: Tabela de colunas e conteúdos do *dataset* de notícias

Coluna	Conteúdo
message	Texto de chamada da notícia na publicação
react_angry	Quantidade de reações Angry na publicação
react_haha	Quantidade de reações Haha na publicação
react_like	Quantidade de reações Like na publicação
react_love	Quantidade de reações Love na publicação
react_sad	Quantidade de reações Sad na publicação
react_wow	Quantidade de reações Wow na publicação
shares	Quantidade de compartilhamentos da publicação

As linhas foram selecionadas de acordo com os maiores valores da coluna *shares*, assim a tabela é formada apenas pelas notícias mais compartilhadas entre o *dataset*. Após a seleção de colunas pertinentes para a análise, são retiradas todas as linhas que contenham algum valor nulo para qualquer coluna da tabela.

Algumas linhas da coluna *message* possuem o link de suas fontes inclusas no texto, então antes de clusterizar, é feita a limpeza de qualquer palavra que comece com *"http"* desses textos, e assim evita-se que se formem clusters apenas com links de notícias.

4.2 Análise de Sentimento

A análise de sentimento é aplicada em dois momentos. Nesse primeiro momento, o *VADER* é aplicado aos textos das notícias para gerar um valor de *compoud* baseado num mapeamento de características lexicais mais fiéis, antes de passar pelo processamento dos dados, já que a ferramenta é sensível à pontuação, gírias e até *emojis* (BURNS, 2019; HUTTO; GILBERT, 2014). Os valores *compound*, provenientes da saída do *VADER* aplicado à cada linha da coluna de texto das notícias, ficam guardados uma nova coluna.

Em um segundo momento, a análise de sentimento será aplicada individualmente aos *clusters* formados na primeira etapa do processo, que será explicado posteriormente.

4.3 Processamento dos dados

O processamento de texto é importante, pois colabora para que os dados sejam melhor compreendidos, tanto por quem analisa quanto pela máquina (BURNS, 2019). Para aprimorar a perfomance da análise, foram realizados alguns processos para o tratamento dos textos contidos na coluna de texto *message* da tabela de notícias.

A remoção de pontuação é o primeiro processo realizado, retirando todas as pontuações do texto que criariam ruído e dificultariam a vetorização e cálculos aplicados aos dados. Com textos limpos, tokenizamos os textos para prepará-los para a lematização, e então removemos as *stop words*, a fim de deixar o texto apenas com palavras que façam sentido para a análise. A lematização então é aplicada para reduzir cada palavra dos textos à sua palavra raíz. Por fim, as palavras que estavam separadas em *tokens*, são transformadas de volta em frases, para as futuras transformações.

4.4 Primeira Etapa

Como foi visto na Seção 3.1 do capítulo anterior, o desenvolvimento do trabalho é dividido em dois *pipelines*. O que foi apresentado até o momento neste capítulo são os primeiros passos comuns entre os *pipelines*.

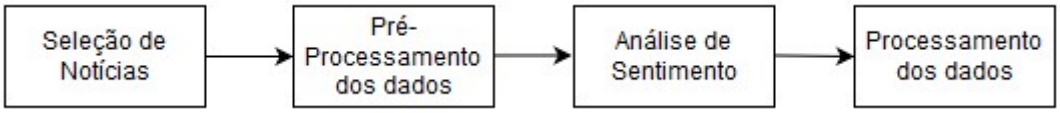

Figura 4.1: Partes comuns entre *pipelines*.

Esta primeira etapa, ou primeiro *pipeline*, visa clusterizar as colunas referentes às reações dadas para notícias e o valor *compound* dado para o texto da notícia, então por enquanto uma nova tabela é formada apenas com esses valores. Após a clusterização, será feita uma análise qualitativa do contéudo em texto das notícias de cada *cluster*. Analisar cada elemento de cada *cluster* seria um processo custoso, então é feita uma segunda clusterização conjunta de análise de sentimento do conteúdo do *cluster*, assim agrupando o conteúdo em menores partes facilmente legíveis.

Os passos para este primeiro *pipeline* serão descritos nas próximas subseções.

4.4.1 Visualização de Dados Brutos

Antes de preparar os dados para a clusterização, um *Pair Plot* foi gerado utilizando a função *pairplot()* da biblioteca seaborn, para visualizar a distribuição de cada coluna, assim como as relações entre pares de colunas (na forma de *scatter plots*), e a densidade dos gráficos representada pela coluna *compound*.

Pelo gráfico na Figura 4.2 podemos observar uma concentração maior de pontos mais escuros (indicando um valor de *compound* mais próximo de 1, indicando sentimento positivo), nas relações entre pares de colunas *react_love* e *react_like*, e pontos mais claros, indicando o sentimento contrário para pares de colunas *react_sad* e *react_angry*.

4.4.2 Padronização e PCA

Após uma visualização dos dados brutos, os dados passam pelo processo de padronização visto na Seção 3.9. Os valores da coluna *compound* já vêm padronizados pela ferramenta *VADER*, então não sofrem mudanças neste passo.

Figura 4.2: *Pair Plot* para visualização dos dados principais.

Como há várias variáveis correspondentes à cada tipo de reação dada para notícias, aplica-se então o método PCA para redução da quantidade de variáveis. Das 7 variáveis, são escolhidas apenas as duas primeiras geradas pelo PCA para a clusterização.

4.4.3 Método do Cotovelo e Clusterização

Para a escolha da quantidade de *clusters*, é aplicado o Método do Cotovelo com valor máximo de K (quantidade de *clusters*), igual a 20. O gráfico obtido apresenta-se na figura 4.3.

Não houve uma queda drástica neste gráfico, então o valor escolhido para K foi o número 5. A clusterização então é realizada através da classe K-Means da biblioteca scikit-learn, com saída definida de 5 *clusters* e valor 2020 como parâmetro para o número de gerações de

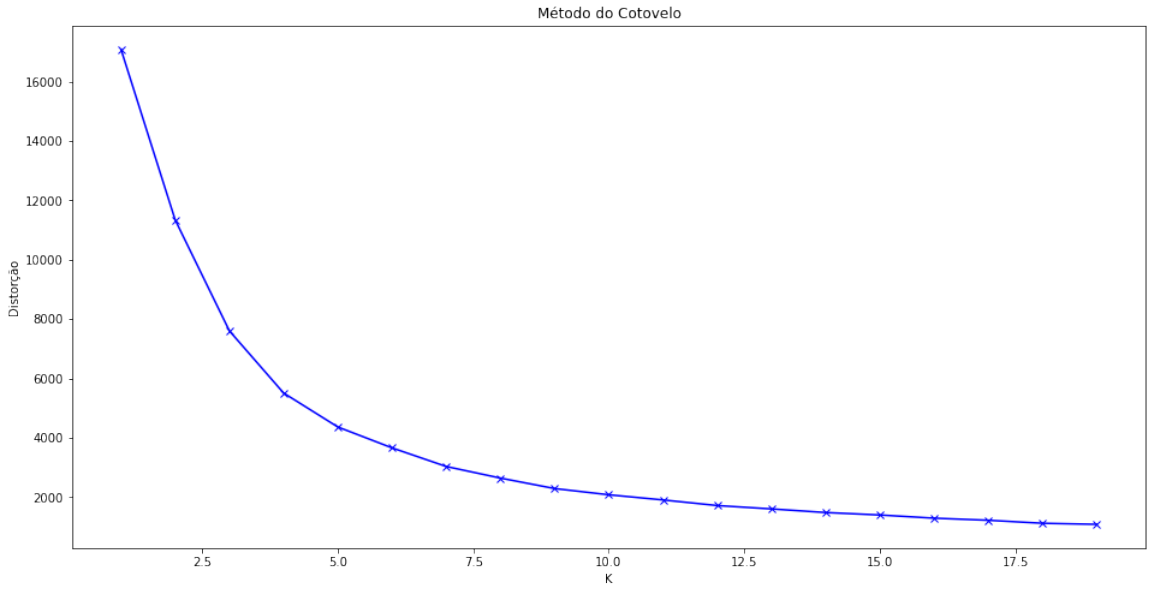

Figura 4.3: Gráfico do Método do Cotovelo.

inicialização de centróides. Para a visualizar a relação entre os componentes PCA gerados e *clusters*, os agrupamentos foram nomeados "segmentos" e recebem um valor ordinal e uma cor no gráfico para melhor identificação. Antes da visualização do gráfico, concatena-se cada linha da tabela com seu texto de notícia correspondente atráves do índice das tabelas. A figura 4.4 mostra a distribuição dos segmentos por componentes PCA.

4.4.4 Análise do conteúdo de cada cluster

Para cada segmento dos 5 criados, é criada uma tabela, facilitando o segundo processo de clusterização. Como mencionado, analisar o texto de cada notícia contida nos agrupamentos já formados seria custoso. O primeiro segmento, por exemplo, tem 1346 linhas, cada linha contendo uma notícia. Um segundo agrupamento de elementos de cada segmento faz-se necessário.

Os textos de cada segmento da primeira clusterização são vetorizados, utilizando a classe CountVectorizer da biblioteca scikit-learn, e então é aplicado o cálculo TF-IDF para ponderar os termos dos vetores através da classe TfidfTransformer, também da biblioteca scikit-learn. Antes da clusterização, o número K de *clusters* tem que ser escolhido, então o Método do Cotovelo é utilizado. A tabela 4.2 contém, para cada nova tabela referente a um segmento, o seu número K.

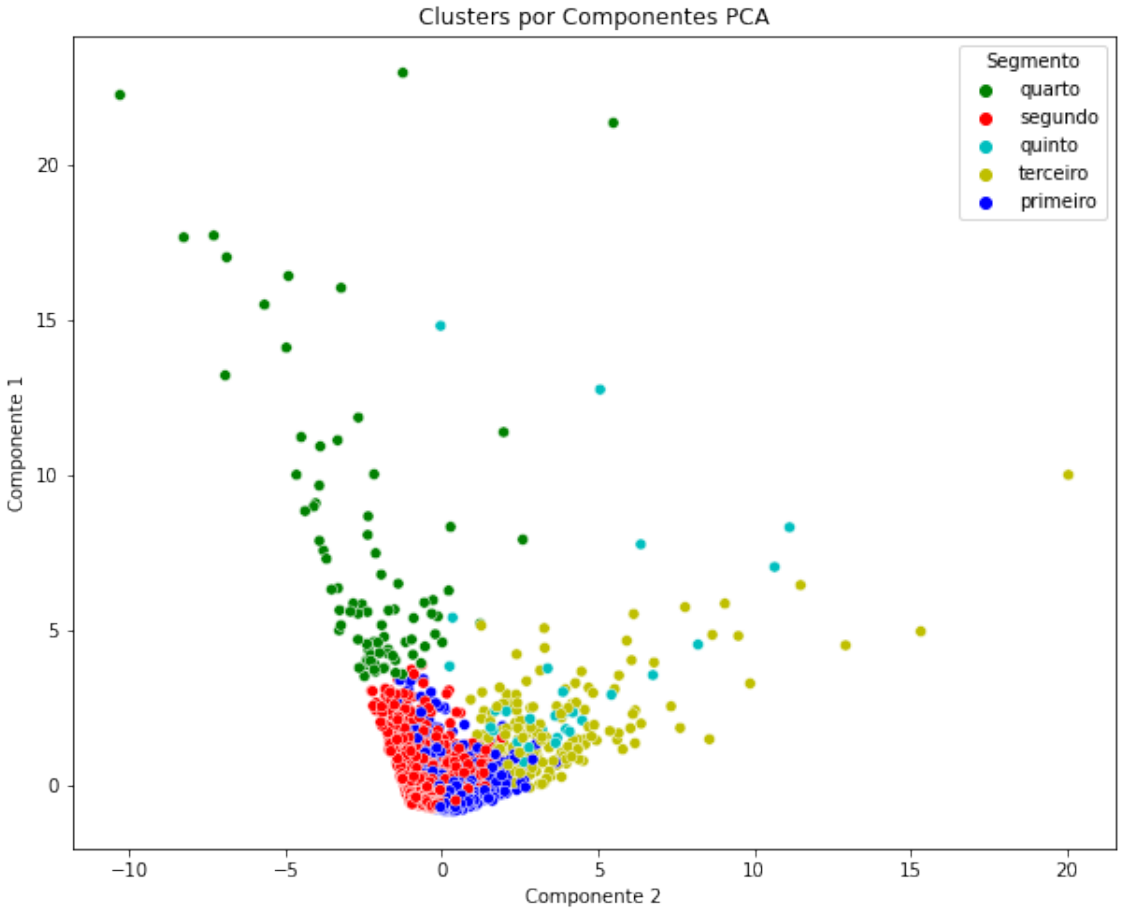

Figura 4.4: Gráfico da distribuição dos segmentos por componentes PCA.

Tabela 4.2: Quantia de *clusters* por segmento.

Tabela	Clusters
Tabela do Primeiro Segmento	9
Tabela do Segundo Segmento	7
Tabela do Terceiro Segmento	7
Tabela do Quarto Segmento	8
Tabela do Quinto Segmento	5

Logo ao aplicar o K-Means para a clusterização, obtém-se em cada agrupamento um vetor de palavras em comum. Cada vetor passa pela função *polarity_scores()* da ferramenta *VADER*, recebendo o valor *compound* que indica o sentimento proeminente.

4.5 Resultados e Discussões da Primeira Etapa

Todos os resultados nas próximas subseções são análises qualitativas dos agrupamentos resultados do algoritmo K-Means para cada segmento definido na Subseção 4.4.3, com um conjunto K definido pelo Método do Cotovelo, para otimizar o resultado, seguido do seu valor de sentimento *compound*.

4.5.1 Primeiro Segmento

A saída de cada *cluster* dos 9 gerados do primeiro segmento são mostrados na Figura 4.5. A maioria dos *clusters* envolve política e o presidente dos Estados Unidos da América, Donald Trump. Um único *cluster* obteve sentimento neutro, e como pode ser observado, é composto em maior parte por substantivos relacionados à política e comunicação. Os *clusters* 1, 4 e 6 envolvem notícias ruins relacionadas à polícia, já para 0, 3 e 5 são notícias diretamente relacionadas à presidência e atos do presidente. O último *cluster* se trata de notícias relacionadas ao aquecimento global e quebra de *icebergs* na Antártida.

```
[0 : new, president, america, republicans, democrats, don, russia, know, yor
k, white, senate, house, disgusting, left] compound: -0.5267
[1 : say, said, police, hell, man, day, federal, right, wrong, child, proble
m, going, city, sad] compound: -0.926
[2 : cnn, meme, news, internet, war, officially, declared, organization, dea
d, trump, fake, destroys, warning, new] compound: -0.9538
[3 : news, attack, fake, president, claim, trump, reporter, bad, media, esta
blishment, mexican, trumps, tourist, fox] compound: -0.8658
[4 : people, state, killed, say, plane, country, new, theyre, mississippi, p
olice, group, think, military, official] compound: -0.6705
[5 : dont, collusion, trump, think, want, know, conclusion, russian, hate, p
eople, care, times, new, talk] compound: -0.0516
[6 : stop, holy, crap, read, destroying, laughing, whatsoever, foul, lying,
play, note, suicide, officer, whining] compound: -0.9118
[7 : trump, donald, president, jr, lawyer, russian, meeting, email, informat
ion, clinton, administration, russia, time, hillary] compound: 0.0
[8 : seriously, iceberg, global, ice, warming, antarctica, broken, shelf, de
laware, broke, size, recorded, largest, scientist] compound: -0.7184
```

Figura 4.5: Saída da clusterização do primeiro segmento.

4.5.2 Segundo Segmento

A saída de cada *cluster* dos 7 gerados do segundo segmento são mostrados na Figura 4.6. O segundo segmento teve seus elementos definidos entre neutros e positivos. Os conteúdos dos

clusters 1, 4 e 6 se tratam de notícias relacionadas à movimentações políticas, enquanto as outras não têm assuntos definidos.

```
[0 : good, make, new, time, say, like, said, year, busted, world, look, wow,
news, really] compound: 0.8481
[1 : donald, trump, jr, russian, meeting, lawyer, email, jrs, clinton, hilla
ry, campaign, met, information, government] compound: 0.0
[2 : breaking, busted, zukerberg, exist, exciting, exclusive, exclusively, e
xcuse, executive, exemption, exercising, exhausted, exhibit, exhibition] com
pound: 0.3612
[3 : cartoon, mcphail, daily, todays, noth, paul, loper, brendan, sipress, l
ewis, chatfield, byrnes, cheng, lautman] compound: 0.0
[4 : trump, president, donald, french, watch, russia, lady, administration,
macron, said, putin, day, trumps, paris] compound: 0.0
[5 : history, meet, hope, museum, american, lesson, year, world, day, decisi
on, commission, learn, new, heres] compound: 0.4404
[6 : people, cnn, think, say, trump, thing, million, news, like, clown, heal
th, great, fact, insurance] compound: 0.765
```

Figura 4.6: Saída da clusterização do segundo segmento.

4.5.3 Terceiro Segmento

A saída de cada *cluster* dos 7 gerados do terceiro segmento são mostrados na Figura 4.7. O terceiro segmento tem agrupamentos distribuídos entre neutros, positivos e negativos. O *cluster* 0 não apresenta assunto definido. Os *clusters* 1, 3 e 5 são relacionados à notícias políticas, o 2 se trata de assuntos que envolvem propagação de doenças sexualmente transmissíveis, e por fim, o *cluster* 6 envolve denúncia de assédio e abuso sexual.

```
[0 : time, gross, brimstone, normal, hmmm, nope, ugh, grabyourwallet, black,
gop, think, near, just, heres] compound: -0.1945
[1 : trump, president, leader, french, lady, macron, house, known, prayer, l
aying, surrounded, bowed, head, hand] compound: 0.4939
[2 : say, satisfying, teenager, oral, expert, dangerous, sex, condom, gonorr
hoea, producing, decline, use, helping, spread] compound: 0.2732
[3 : said, republican, party, street, wall, safer, democrat, make, presiden
t, trump, goes, wanna, deport, linda] compound: 0.6705
[4 : woman, homeless, officer, ground, nasty, beat, violently, policeman, ar
rest, shown, striking, bed, thought, california] compound: -0.8689
[5 : donald, trump, jr, email, government, breaking, president, russian, new
s, told, hillary, clinton, say, aid] compound: 0.0
[6 : called, political, stunt, japan, sulfur, president, victim, harassment,
worland, burning, fake, accused, sexual, assault] compound: -0.9287
```

Figura 4.7: Saída da clusterização do terceiro segmento.

4.5.4 Quarto Segmento

A saída de cada *cluster* dos 8 gerados do quarto segmento são mostrados na Figura 4.8. O quarto segmento, assim como o terceiro, foi formado por *clusters* de sentimentos variados. A maior parte dos *clusters* não apresenta assunto em específico. Pode-se relacionar o *cluster* 3 à assuntos presidenciais, e o 5 à notícias sobre imagens captadas de animais marinhos na Austrália.

```
[0 : youre, drop, mic, place, tbh, relationship, goal, lot, stop, like, afgh
an, happy, getting, agency] compound: 0.4404
[1 : time, went, second, new, mon, laferte, buy, right, didn, 36yearold, dad
dy, man, powerful, america] compound: 0.4215
[2 : update, boom, true, day, doctors, baby, lion, cell, save, story, counte
rprotesters, klan, largely, member] compound: 0.7184
[3 : president, trump, donald, white, home, photo, macron, wife, meme, mad,
cnn, latest, hopping, house] compound: -0.4939
[4 : say, lesser, trump, hope, thurston, douchebag, von, hot, fuckface, sea
t, shitbag, wouldn, quite, evil] compound: -0.8999
[5 : play, footage, drone, capture, australia, coast, appears, dolphin, west
ern, whale, pod, moment, cast, thwarted] compound: 0.3182
[6 : people, think, drowning, today, island, moment, male, player, mean, cub
e, ice, teachable, watching, eyerolling] compound: 0.0
[7 : news, daughter, dad, like, fifth, react, son, bbc, treat, delivered, ba
dly, needed, california, good] compound: 0.6124
```

Figura 4.8: Saída da clusterização do quarto segmento.

4.5.5 Quinto Segmento

A saída de cada *cluster* dos 5 gerados do quinto segmento são mostrados na Figura 4.9. Todos os agrupamentos do quinto segmento receberam valores sentimentais negativos. Este é o único segmento sem qualquer *cluster* relacionado à notícias presidenciais dos Estados Unidos da América, e tem apenas o *cluster* 2 composto por termos definidos, que levam à interpretação de notícias relacionadas a mortos e sobreviventes de tragédias.

4.6 Segunda Etapa

Para a segunda etapa, os primeiros passos do *pipeline* em comum com a primeira etapa foram executados, como mencionado.

Nessa etapa, são analisados os textos de todas as notícias da tabela em *clusters* e seus respectivos sentimentos atribuídos. Após a clusterização, os textos vetorizados serão

```
[0 : night, owner, toilet, flee, chewy, las, threemonth, relationship, airpo
rt, vegas, abusive, abandoned, old, nana] compound: -0.802
[1 : battle, cancer, bradley, life, bbcin2udlo6n, malnourished, left, mosul,
orphan, year, rio, dy, ferdinand, tragedy] compound: -0.9081
[2 : marine, corps, mississippi, plane, say, 16, survived, came, county, lef
lore, airplane, crash, dead, official] compound: -0.5719
[3 : antarctica, ice, away, thats, size, block, broken, giant, fourtimes, lo
ndon, build, jimmy, carter, president] compound: -0.7184
[4 : legacy, mother, child, life, terrible, let, daughter, bbc, education, q
uality, bad, feel, facing, doctor] compound: -0.765
```

Figura 4.9: Saída da clusterização do quinto segmento.

transformados em componentes PCA para a visualização dos dados em gráfico. Os passos de cada transformação e resultados são discutidos nas próximas subseções.

4.6.1 Vetorização e Cálculo TF/IDF

Como visto na Subseção 4.4.4, para que se possa clusterizar dados em texto, deve-se vetorizar as notícias, usando a classe *CountVectorizer* da biblioteca scikit-learn, e ponderar os vetores usando da classe TfidfTransformer. As palavras de cada vetor são guardadas em uma variável usando da função *get_feature_names()*, da classe *CountVectorizer*, assim é possível recuperá-las.

4.6.2 Método do Cotovelo e Clusterização

Com textos de notícias transformados em vetores ponderados, a quantidade de *clusters* é escolhida baseando-se no gráfico do Método do Cotovelo, este gerado com a quantidade máxima K igual a 20.

O número de *clusters* escolhido é 12, então são formados os agrupamentos de palavras comuns entre notícias juntamente ao sentimento geral encontrado em cada grupo. A análise dos *clusters* será discutida na Seção 4.7, com base na Figura 4.11.

4.6.3 PCA e Visualização

À título de visualização, os textos vetorizados foram transformados em componentes PCA, e desses componentes foram escolhidos 2 para serem representados no gráfico da Figura 4.10. O gráfico representa em segmentos todos os 12 *clusters* de notícias, formados no capítulo anterior, em cores diferentes. A quantidade de *clusters* é maior que da primeira etapa, e os segmentos são constituídos de palavras em comum por contexto de notícia. Pode-se observar

que os agrupamentos na figura aparecem mais dispersos, mas ainda é possível visualizar alguns *clusters* mais definidos, como o primeiro, segundo, décimo e décimo segundo.

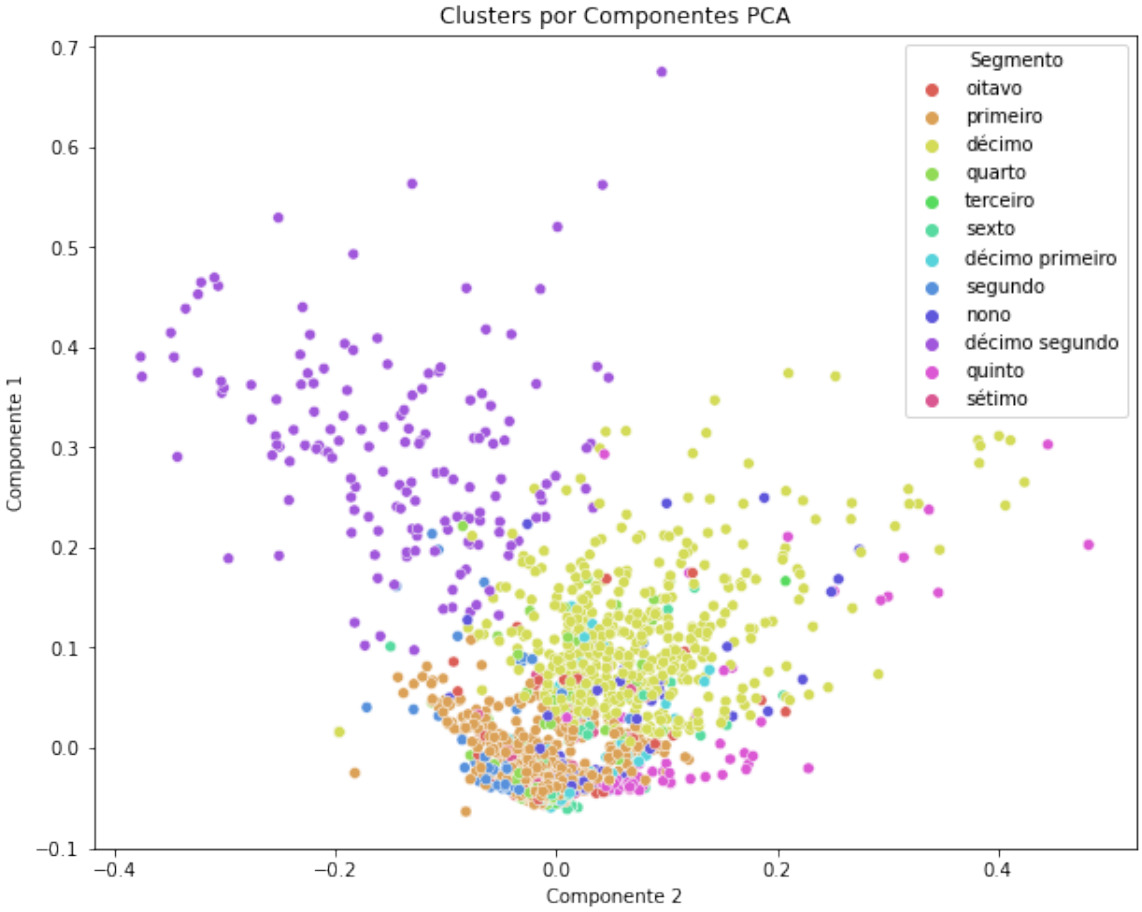

Figura 4.10: Gráfico da clusterização de notícias.

4.7 Resultados e Discussões da Segunda Etapa

Na Figura 4.11 apresentam-se os 12 *clusters* e seus valores *compound* atribuídos.

Os sentimentos dos grupos variam dentre o espectro, mas a maior parte recebe valores para sentimento positivo. Quase todos os grupos citam o presidente Donald Trump, alguns citando outros presidentes e ex-presidentes. O assunto de cada *cluster* não é bem definido, entende-se apenas que a maioria trata-se de movimentações políticas.

Partindo dos *clusters* analisados, pode-se observar que para as 5000 notícias mais compartilhadas dentre o *dataset*, a maioria retrata acontecimentos políticos. Embora algum assunto em específico não tenha sido identificado, os grupos recebem palavras pontuais que não se repetem em outros grupos.

```
[0 : said, year, news, day, new, busted, state, going, cnn, woman, think, li
ke, man, police] compound: 0.3612
[1 : york, new, times, opinion, section, los, angeles, city, trump, obamacar
e, state, men, reported, better] compound: 0.4404
[2 : right, come, political, civil, news, trump, year, change, need, attack,
climate, gun, ll, law] compound: -0.6705
[3 : say, care, health, trump, senate, like, want, dont, donald, said, presi
dent, just, house, child] compound: 0.7184
[4 : good, shape, lady, french, trump, news, youre, coffee, beautiful, love
r, president, tell, lord, donald] compound: 0.891
[5 : make, stuff, safer, president, street, dont, difference, america, trum
p, food, sure, new, car, break] compound: 0.6249
[6 : look, daily, kos, member, community, like, cartoon, todays, behindthesc
enes, expatgirl, amendment, user, church, right] compound: 0.3612
[7 : people, think, country, say, help, group, trump, today, thing, know, th
eyre, life, president, million] compound: 0.4019
[8 : time, trump, president, cover, magazine, long, hard, child, big, prett
y, olympics, race, went, tough] compound: 0.3182
[9 : trump, president, donald, breaking, russia, trumps, house, white, said,
administration, putin, macron, french, watch] compound: 0.0
[10 : world, obama, got, g20, leader, trump, barack, war, gas, nation, highe
r, president, administration, stage] compound: -0.5994
[11 : donald, trump, jr, russian, lawyer, meeting, email, clinton, hillary,
jrs, information, campaign, met, russia] compound: 0.0
```

Figura 4.11: Saída da clusterização da segunda etapa.

Capítulo 5

Conclusões

O processo de análise exploratória desenvolvido no capítulo anterior mostrou-se eficiente para que algumas informações sobre o *dataset* de notícias utilizado pudessem ser extraídas.

A quantidade média de *clusters* escolhidos usando do Método do Cotovelo foi 8, calculando o método com K máximo igual a 20.

A análise mostra que a clusterização K-Means pode ser usada como uma abordagem para agrupar um corpus de dados qualitativos, como notícias publicadas no *Facebook*. Além disso, a análise também mostra que métodos quantitativos em aprendizado de máquina, particularmente o algoritmo K-Means, podem ser usados para analisar dados qualitativos, como notícias e reações.

O uso do *VADER* para análise de sentimento mostrou-se suficiente para comparação de sentimentos dados em reações e o sentimento dado pelas palavras nos textos.

Numa análise mais individual realizada na primeira etapa, foi possível observar que a maior parte dos *clusters* gerados de cada segmento obtiveram uma pontuação negativa para o sentimento gerado. O mesmo não ocorre numa visão mais geral dada pela análise na segunda etapa, onde é mostrado que ao se reunir todos os textos de notícias de uma vez e clusterizá-los, sem antes separá-los por reações, a maior parte recebe uma pontuação positiva para o sentimento gerado em cada *cluster*.

Levando em consideração a abordagem da primeira etapa, que engloba tanto as reações quanto os sentimentos das notícias, pode-se observar que o ambiente em relação à notícias no *Facebook* é negativo. A segunda etapa demonstrou um resultado diferente, porém desconsidera a contextualização dos sentimentos em relação a o que as reações indicam sobre o conteúdo dessas notícias.

Os métodos nesta análise podem ser replicados em outros domínios práticos. O mesmo método de agrupar dados e encontrar sentimentos para cada cluster formado pode ser aplicado com dados como avaliações e feedbacks. Isso abre possibilidades de novos tipos de caso estudos que exigem a interpretação do feedback, extraindo recursos e compreendendo o sentimento associado.

Referências bibliográficas

BENCINA, J. (Ed.). **Facebook News Scraper**. 2020. Disponível em: <https://github.com/jbencina/facebook-news>. Acesso em: 12 jan. 2020.

BENGFORT, B.; BILBRO, R.; OJEDA, T. **Enabling Language-Aware Data Products with Machine Learning**. 1. ed. 1005 Gravenstein Highway North, Sebastopol, CA 95472: O'Reilly Media, Inc., 2018. ISBN 9781491963043.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural Language Processing with Python**. 1. ed. 1005 Gravenstein Highway North, Sebastopol, CA 95472: O'Reilly Media, Inc., 2009. ISBN 0596516495.

BURNS, S. **Natural Language Processing: A Quick Introduction to NLP with Python and NLTK**. 2. ed. 1005 Gravenstein Highway North, Sebastopol, CA 95472: GlobatechNTC, 2019. ISBN 1699028451.

COLLOMB, A.; BRUNIE, L.; COSTEA, C. A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation. In:

DUNTEMAN, G. H. **Principal components analysis**. [S.l.]: Sage, 1989.

GARAY, J.; YAP, R.; SABELLANO, M. J. G. An analysis on the insights of the anti-vaccine movement from social media posts using k-means clustering algorithm and VADER sentiment analyzer. In: p. 304–309. DOI: 10.1088/1757-899X/482/1/012043.

GIUNTINI, F. T. et al. **How Do I Feel? Identifying Emotional Expressions on Facebook Reactions Using Clustering Mechanism**. [S.l.], abr. 2019. DOI: 10.1109/ACCESS.2019.2913136.

HUTTO, C.; GILBERT, E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In: ICWSM. [S.l.: s.n.], 2014.

KOROVKINAS, K.; DANĖNAS, P.; GARŠVA, G. **SVM and k-Means Hybrid Method for Textual Data Sentiment Analysis**. Kaunas, Lithuania, jan. 2019. DOI: 10.22364/bjmc.2019.7.1.04.

LEE, L.; PANG, B.; VAITHYANATHAN, S. Thumbs up? Sentiment Classification using Machine Learning Techniques. In: p. 79–86. DOI: 10.3115/1118693.1118704.

LI, G.; LIU, F. **A Clustering-based Approach on Sentiment Analysis**. Hangzhou, China, nov. 2010. DOI: 10.1109/ISKE.2010.5680859.

LIU, S.; LEE, I. Email Sentiment Analysis Through k-Means Labeling and Support Vector Machine Classification. **CYBERNETICS AND SYSTEMS: AN INTERNATIONAL JOURNAL**, v. 49, n. 3, p. 181–199, 2018. DOI: 10.1080/01969722.2018.1448242.

ORKPHOL, K.; YANG, W. **Sentiment Analysis on Microblogging with K-Means Clustering and Artificial Bee Colony**. [S.l.], jul. 2019. DOI: 10.1142/S1469026819500172.

QUALIBEST, I. (Ed.). **Estudo: Redes Sociais no Brasil**. 2019. Disponível em: <https://www.institutoqualibest.com/download/redes-sociais-brasil/>. Acesso em: 15 nov. 2020.

RIAZ, S.; FATIMA, M.; KAMRAN, M.; NISAR, M. W. **Opinion mining on large scale data using sentiment analysis and k-means clustering**. Wah Cantt, Pakistan, jul. 2017. DOI: 10.1007/s10586-017-1077-z.

RODRIGUEZ, A.; ARGUETA, C.; CHEN, Y.-L. **Automatic Detection of Hate Speech on Facebook Using Sentiment and Emotion Analysis**. Okinawa, Japan, fev. 2019. DOI: 10.1109/ICAIIC.2019.8669073.

SANDOVAL-ALMAZÁN, R.; VALLE-CRUZ, D. **Facebook Impact and Sentiment Analysis on Political Campaigns**. Toluca, Mexico, mai. 2018. DOI: 10.1145/3209281.3209328.

STATISTA (Ed.). **Most popular social networks worldwide as of July 2020, ranked by number of active users.** 2020. Disponível em: <https://www.statista.com/statist%20ics/272014/global-social-networks-ranked-by-number-of-users/>. Acesso em: 14 nov. 2020.

TIAN, Y. et al. Facebook Sentiment: Reactions and Emojis. In: p. 11–16. DOI: 10.18653/v1/W17-1102.

USAGE, I. W. S.; STATISTICS, P. (Ed.). **Internet Usage, Facebook Subscribers and Population Statistics for all the Americas World Region Countries July 31, 2020**. 2020. Disponível em: <https://www.internetworldstats.com/stats2.htm>. Acesso em: 14 nov. 2020.

VANDERPLAS, J. **Python Data Science Handbook**. 1. ed. 1005 Gravenstein Highway North, Sebastopol, CA 95472: O'Reilly Media, Inc., 2016. ISBN 9781491912058.

YESSENOV, K.; MISAILOVIC, S. **Sentiment Analysis of Movie Review Comments**. [S.l.], mai. 2009.

ZENG, H.-J. et al. CBC: Clustering Based Text Classification Requiring Minimal Labeled Data. In: p. 443–450. DOI: 10.1109/ICDM.2003.1250951.