

# Análise comportamental a partir de publicações feitas no Twitter

Fabrizio Guilherme Beatrice de Oliveira<sup>1</sup>, Pedro Henrique Cacique Braga<sup>1</sup>

<sup>1</sup> Faculdade de Computação e Informática – Universidade Presbiteriana Mackenzie  
Rua da Consolação, 930 - Consolação, São Paulo - SP, 01302-907

guilhermo.fabricio@gmail.com, pedro.braga@mackenzie.br

**Abstract.** *Mining data consists of extracting valuable information implicit in environments, which are not imagined that it could be contained. Thus, looking at the profitable environment in which the major social networks present themselves and studies about psychology and human relations, Twitter appears to be, apparently, a very rich source of information and opinions that can serve as a basis for conclusions about a certain audience. or product, which can help even in the creation of advertising pieces or general understanding of the market. Thus, the present Course Conclusion Paper seeks to mine and intelligently handle publications on Twitter in search of information that aggregates about the behavior of the desired audience.*

**Keywords:** *Aining. Data. Information. Analysis. Text.*

**Resumo.** *Minerar dados consiste em extrair informações valiosas implícitas em ambientes, os quais não se imagina que elas poderiam estar contidas. Desse modo, observando o meio lucrativo em que as grandes redes sociais se apresentam e estudos acerca da psicologia e relações humanas, o Twitter mostra-se, aparentemente, uma fonte muito rica de informações e opiniões que podem servir como base para conclusões sobre determinado público ou produto, o que pode ajudar até mesmo na criação de peças publicitárias ou entendimento geral de mercado. Assim, o presente Trabalho de Conclusão de Curso busca minerar e tratar de forma inteligente publicações no Twitter em busca de informações que agreguem a respeito do comportamento do público desejado.*

**Palavras-chave:** *Mineração. Dados. Informação. Análise. Texto.*

## 1. Introdução

Com o mundo globalizado, o qual, segundo o levantamento “Internet Usage Worldwild” (Statista, 2021), são mais de 3,79 bilhões de pessoas conectadas entre si por meio da Internet, o acesso à informação e divulgação da mesma ficou extremamente mais eficiente do que há um século atrás. Atualmente, uma pessoa com acesso a uma rede de “internet” pode compartilhar opiniões gerais sobre os mais diversos temas, com as redes sociais tornando isso cada vez mais popular e acessível.

O Twitter é um exemplo de mídia que permite de forma rápida e fácil o compartilhamento de opiniões abertas sobre os mais diversos temas (filmes, política, jogos...). Tratando-se de uma rede social que se caracteriza como “microblogging”, ou seja, um espaço para expressar pensamentos com um número limitado de caracteres, o compartilhamento de opiniões nessa mídia se popularizou ao redor da Terra. Assim, existe uma

quantidade enorme de informações presente gratuitamente na “internet”, ainda em 2013 o fluxo de novas publicações no Twitter possuía uma média diária de 500 milhões de publicações (Twitter, 2013), torna-se o lugar ideal para análise sentimental, visto a tração da plataforma e todo o objetivo de comentários curtos que expressam opiniões e desejos.

### **1.1. Contextualização e relevância do tema**

Seguindo a analogia descrita por Leandro Nunes e Daniel Gomes (2016) no livro “Introdução a Mineração de Dados”, pode-se descrever o ato de minerar dados como a extração de minerais valiosos, como ouro e pedras preciosas a partir de uma mina. Embora esses materiais não possam ser cultivados ou produzidos artificialmente, eles existem de maneira implícita, podendo ser extraídos. Esse processo requer acesso à mina, utilização de ferramentas adequadas, a extração do conteúdo propriamente dita e o preparo para comercialização.

De forma enxuta, a mineração de dados segue o mesmo fluxo de serviço que a mineração convencional já conhecida: explora-se uma base de dados já existente de forma implícita, não replicável artificialmente (mina), utilizamos algoritmos (ferramentas) específicos para extrair o conhecimento ou a base desejada (minerais preciosos). Todavia, vale lembrar que tais dados, sem o tratamento adequado, são joias brutas sem um valor necessariamente agregado.

Com a chegada da “internet” e o avanço no desenvolvimento de sistemas complexos, a computação passou a ser a força motriz da sociedade moderna, o que a torna cada vez mais interdisciplinar. Com a mineração de dados não é diferente, a mesma envolve diversas áreas como estatística, matemática, engenharia, inteligência artificial, bancos de dados, sistemas de informação, visualização. . .

Além disso, a mineração de dados está contida em um conjunto de um processo maior, conhecido como descoberta de conhecimento em bases de dados (knowledge discovery in databases, ou KDD). Sendo assim, com finalidade de estudos e aprofundamentos, este trabalho seguirá respeitando o KDD e seguindo todos os quatro passos presentes:

- Base de dados: coleção organizada de dados, conceitualmente falando, a parte mais básica de todo o processo, as informações de forma bruta, sem tratamento algum, em sua forma originária;
- Pré-processamento de dados: processo anterior a mineração de dados propriamente dita, o qual consiste na preparação da base para uma análise eficaz e eficiente. Nessa etapa, há remoção de ruídos e elementos inconsistentes, combinação de conceitos obtidos de múltiplas fontes, escolhas das informações relevantes a análise e transformação dos fatos para um formato apropriado a mineração;
- Mineração de dados: aplicação de algoritmos específicos responsáveis por extrair o conhecimento implícito nos dados preprocessados na etapa anterior;
- Avaliação ou validação do conhecimento: avaliação final dos resultados obtidos, objetivando extrair conhecimentos verdadeiramente úteis e não triviais.

Para fins de análises comportamentais e aprofundamento nos estudos acerca de mineração de dados, este trabalho irá analisar comentários reproduzidos na plataforma (Twitter), utilizando técnicas de “text mining” para futuramente trata-los através de algoritmos ou expressões regulares de modo a tirar informações interpretáveis e processá-las para tirar valor das mesmas.

### **1.1.1. Contextualização da plataforma**

Twitter é uma rede social e um servidor para “microblogging”, criado em Março de 2006 por Jack Dorsey, Evan Williams, Biz Stone e Noah Glass, que permite aos usuários enviar e receber atualizações pessoais de outros contatos (em textos de até 280 caracteres, conhecidos como “tweets”).

As atualizações são exibidas no perfil de um usuário em tempo real e também enviadas a outros usuários seguidores que tenham assinado para recebê-las. As atualizações de um perfil ocorrem por meio do sítio do Twitter, por RSS, por SMS ou programa especializado para gerenciamento. O serviço é gratuito pela “internet” podendo ser acessado de qualquer navegador, entretanto, usando o recurso de SMS pode ocorrer a cobrança pela operadora telefônica.

Na plataforma, pessoas do mundo todo podem iniciar um fio (sequência de postagens relacionadas a um mesmo assunto) ou até mesmo postagens individuais a qualquer momento. Além disso, é possível resumir todo o ambiente em três ações fundamentais para disseminação de ideias e opiniões: o “responder” que permite um usuário da plataforma interagir com uma publicação fazendo comentários a respeito, o chamado “retweet” termo criado e popularizado pelo próprio Twitter para representar o compartilhamento de uma publicação podendo ser esse compartilhamento com comentário ou sem, e a função de “curtir” que permite que usuários demonstrem apoio e ajudem no engajamento de outras publicações.

O meio mais comum de interação com os produtores de conteúdo é o botão de curtir e o “retweet”. Importante frisar que, com o perfil configurado com a privacidade pública, todos os seus compartilhamentos e postagens pessoais ficam visíveis para qualquer outro usuário que tenha acesso ao seu identificador único (chamado popularmente por arroba) ou que vejam na própria linha do tempo da rede social caso o sigam ou através de interações.

### **1.1.2. Contextualização do problema da pesquisa**

Durante grandes campanhas publicitárias, atos políticos ou grandes eventos num geral, pode-se observar dezenas de milhares de comentários sendo publicados por minuto, com as mais variadas opiniões possíveis. Tudo isso se tora facilmente trassável através de hashtags (um termo associado a assuntos ou discussões que se deseja indexar em redes sociais, inserindo o símbolo da cerquilha [#] antes da palavra, frase ou expressão. Quando a combinação é publicada, transforma-se em um hiperlink que leva para uma página com outras publicações relacionadas ao mesmo tema, popularmente chamado de *trand*). Todavia, é importante tomar cuidado com a quantidade elevada de ruídos presentes no meio desses textos, como destacado no artigo “Text Mining of Audience Opinion in eSports Events” (Bruno Omella, 2017), vale ressaltar a presença de emotes: pequenas imagens que são exibidas embutidas em mensagens de texto quando certas palavras são digitadas. Cada uma dessas notas tem um certo significado implícito, que pode variar dependendo do contexto de uso. Alguns emotes representam opiniões ou sentimentos, como aprovação ou alegria, enquanto outros têm significados muito específicos. Esses memorandos desempenham um grande papel na maneira como o significado é transmitido

por meio de mensagens de bate-papo. Além disso, outro grande ruído presente na extrassão de mensagens do Twitter são os arrobas e as menções a outros “tweets”.

Através da análise dos comentários presentes dentro de um determinada fio é possível tirar conclusões preciosas não só sobre o público que está opinando, mas também sobre a campanha num geral, isso ocorre pelo fato de que pessoas com gostos semelhantes acabam assistindo as mesmas coisas e se interessando por pessoas que apresentam traços semelhantes aos seus, como destacado pelo artigo “Behavior-Based Personality-Assessment Method Reveals Personality Similarity Among Couples and Friends” (Youyou Wu, 2017). Dessa maneira, o Twitter por se tratar de uma plataforma gratuita e com o tráfego que possui, consegue replicar esse comportamento da vida real. Desse modo, através de técnicas de text-mining (mineração de dados textuais) e posteriormente aplicando um tratamento inteligente, é possível analisar comportamentos de grupos ou bolhas específicas centralizadas em uma hashtag do site ou até mesmo em um perfil específico.

Neste contexto, a pergunta que será respondida nesta pesquisa é: como a análise de perfis do Twitter através de text-mining e tratamentos inteligentes podem inferir em conclusões?

## **2. Objetivos**

Visando consolidar o conhecimento de técnicas de mineração de dados, este trabalho tem o propósito buscar valor agregado em publicações feitas por pessoas na rede social Twitter, desenvolvendo uma aplicação na linguagem de programação Python capaz de pegar as últimas publicações de um determinado perfil na rede social e mapear as palavras mais citadas, o humor do perfil e, por fim, a subjetividade do mesmo.

## **3. Referencial teórico**

Para produção textual referente a este capítulo será usado um fluxo de produção semelhante ao desenvolvimento do programa complementar a esta obra. Assim, a primeira obra a ser tratada será o livro que embasa todo esse TCC: “Introdução a Mineração de Dados”, escrito pelo professor Dr. Leandro Nunes de Castro com coautoria de Daniel Gomes Ferrari, nesta obra é possível analisar toda uma introdução referente ao tema, onde é apresentado técnicas e termos ligados a mineração de dados e tudo que será visto no livro, aprende-se as principais tarefas a serem feitas para se obter um resultado eficiente e que possa agregar a estudos e análises, é possível ver aplicações concisas e práticas das técnicas a serem ensinadas como: predição de produtividade de grãos, análise de sentimento em redes sociais (destaca-se este como um dos mais relevantes capítulos a serem estudados e agregados a este Trabalho de Conclusão de Curso), detecção de fraudes em cartões de crédito, combate a perdas não técnicas de energia elétrica, segmentação de curvas de carga em sistemas de energia elétrica e modelagem de processos siderúrgicos.

Em trechos subsequentes, é possível aprender as técnicas seguidas com base no, já citado anteriormente no capítulo 1.1 deste trabalho, knowledge discovery in databases, ou KDD, com início no pré-processamento de dados, nomenclaturas e tipos, todo o processo referente a limpeza dos dados, valores ausentes dados ruidosos (entendidos como outliers), os quais podem danificar ou distorcer fortemente o resultado final, dados inconsistentes, integração dos dados, compressão de atributos, redução no número de dados, padronização, normalização, discretização e por fim a redução. Desse modo, teremos uma

base de informações sólidas e concisa para, após o tratamento para visualização adequada, chegarmos ao resultado aguardado e, possivelmente, validação da hipótese apresentada.

Tudo isso levando em consideração fundamentos matemáticos essenciais para a análise, os quais podem ser divididos em conjuntos e operações com conjuntos, vetores e matrizes, operações elementares usando essas estruturas de dados, espaço vetorial linear, produto interno, produto externo, norma, semi-norma, e quase-norma, ângulo entre dois vetores, ortogonalidade e ortonormalidade entre dois vetores e espaços ortogonais. Conceitos básicos de estatística são essenciais, entre eles podem ser ressaltados os conceitos de população, amostra, variáveis, probabilidade, variáveis aleatórias, medidas, entropia da informação, curva normal e, por fim, intervalo de confiança. Além disso, computacionalmente falando é importante entender processos elementares presentes em teoria dos grafos, como definições básicas como árvores. Com uma base sólida dos conceitos apresentados, faz-se importante compreender pontos de base biológica das RNAs (ácido ribonucleico), as quais permitiram um tratamento avançado, o qual ajudará na compreensão final e conclusão do trabalho.

Vale ressaltar a importância do conhecimento em técnicas de otimização computacional via métodos de busca, o qual, junto com a base já citada, norteará o desenvolvimento e os resultados apresentados por este trabalho.

Com todos os pontos apresentados e estudados através do livro base, é possível partir para segunda obra que servirá como referência para o desenvolvimento subsequente, o artigo “Text Mining of Audience Opinion in eSports Events” apresentado na SBGames 2017. Nele, em suma, o autor usou técnicas de mineração de texto para formar uma base de dados sólida, recolhida através dos bate-papo ao vivo da transmissão da *North American League Championship Series* (NALCS) e *European League Championship Series* (EULCS), dois dos maiores campeonatos de eSports do mundo. Além disso, o autor aplicou os conceitos de mineração de dados para extrair conclusões a respeito da percepção do público espectador dos times competidores.

No artigo é possível ver todo o processo e ferramentas utilizadas, assim como a justificativas para cada uma que foi apresentada. Outro ponto destaque são as dificuldades apresentadas por Bruno Omella no momento de fazer a filtragem de comentários relevantes ao estudo, pois assim como já citado anteriormente, a linguagem de internet apresentada em bate-papos semelhantes ao de transmissões de grandes jogos ao vivos apresenta algumas barreiras que podem causar grande quantidades de ruído, como a presença dos “emotes”, os quais fazem referências diretas a objetos popularizados no cenário, jogadores ou até mesmo algum sentimento relacionado ao atual momento da jogada.

Todavia, há grandes vantagens em uma linguagem tão informal neste meio de comunicação como, por exemplo, o uso das chamadas “hashtags”, que é um composto de palavras-chave, ou de uma única palavra, que é precedido pelo símbolo cerquilha (#). Esse tipo de marcação, utilizada nas redes sociais e em outros meios, serve para associar uma informação a um tópico ou discussão. Geralmente essas hashtags tornam-se links indexáveis pelos mecanismos de busca. Isso permite que os demais usuários possam clicar nelas ou procurá-las e visualizarem todas as informações, imagens, vídeos etc relacionados as mesmas. Com a ampla utilização delas em bate-papos gerais, o levantamento de relevância se torna um pouco mais fácil.

Além disso, outro grande pilar serão estudos que permitirão maior compreensão a respeito do comportamental e da mente humana. Um dos exemplos, que está sendo usado como base, é o artigo “Birds of a feather do flock together: Behavior-based personality-assessment method reveals personality similarity among couples and friends”, o qual tenta mostrar, através de experimentos e levantamentos, como pessoas que apresentam gostos similares ou afinidades por coisas específicas tendem a se relacionar, seja como amigos ou até mesmo em algum vínculo amoroso. Isso é muito importante para começar a entender como as pessoas se juntam para comentarem a respeito de determinado ato ou até mesmo interagirem com a sua bolha social, trazendo reflexões como: “elas estão atrás de algo específico?”.

Com esses três pilares da literatura, será possível desenvolver o objetivo apresentado. Em suma, há todos os conceitos básicos para desenvolvimento da meta apresentada no capítulo 2, um exemplo prático de um processo semelhante com descrições de ferramentas utilizadas e experiências com dificuldades e facilidades e, por fim, um estudo sobre comportamento humano para embasar conclusões futuras.

A partir do que já foi descrito, outras grandes referências que serão utilizadas giram em torno de APIs que fazem monitoramentos constantes, como “Twitter API” a qual pode dar informações preciosas de quem é a pessoa mais seguida, resgatar “tweets”.

#### **4. Metodologia**

Iniciou-se com a revisão da literatura específica geral a cerca da mineração de dados. O objetivo principal desta análise inicial foi aprofundar-se nos conceitos do tema, entender sua importância em meio a sociedade, as práticas para limpar ruídos e extrair informações. Estudo comportamental e ligado a bibliotecas de análise sentimental, como TextBlob, também tiveram relevância no estudo, uma vez que foram fortemente aplicados no desenvolvimento do programa.

Posteriormente analisou-se detalhadamente da API disponibilizada pelo *Twitter* e como é feito o estudo de comportamento, no final, uma aplicação prática simples de um programa que analise os sentimentos empregados em publicações de uma determinada conta na rede social e atenda a base do objetivo proposto, provando a viabilidade da hipótese levantada.

Em suma, o estudo se enquadra nas seguintes classificações clássicas de pesquisas científicas:

- Pesquisa experimental: Pode ser considerada de natureza aplicada, dado que visa aplicação prática, concluindo com a construção de um programa capaz de mostrar as palavras mais citadas por um determinado perfil e analisar o sentimento que norteia as postagens do mesmo;
- Pesquisa quantitativa: Norteada pelo positivismo, considerando que a realidade só pode ser compreendida por meio na análise de dados brutos. Assim, recorrendo à matemática para descrever algum fenômeno e analisar as informações recolhidas.
- Pesquisa bibliográfica: Levantamento de informações acerca de mineração de dados a partir de materiais bibliográficos.

De modo a estudar e buscar provar os pontos levantados nesse trabalho de conclusão de curso foi construído um projeto prático, programado em Python, o qual, para estudo, foi feito em proporções menores, ou seja, ao invés de analisar um conjunto de “hashtags” ou

um grupo de perfis, a análise sentimental foi aplicada sobre o perfil do ex presidente dos Estados Unidos: Barack Obama. Sendo assim, nesse capítulo será feita a documentação do projeto prático sobre o perfil do político.

#### **4.1. Bibliotecas usadas**

Para facilitar a construção do programa, foram utilizadas diversas bibliotecas disponibilizadas em Python, dentre elas:

- tweepy: A principal para guiar o estudo, pois ela que permite resgatar de forma simples tweets publicados na rede social;
- textblob: Responsável por ajudar na interpretação e análise dos textos extraídos;
- wordcloud: Para fins ilustrativos, foi utilizada para montar uma nuvem de palavras utilizadas nos textos;
- pandas: tratamento numérico e organizacional dos dados;
- matplotlib: exibição gráfica dos dados;
- re: utilização de expressões regulares, as quais ajudaram a limpar os textos minerados para posteriormente serem tratados.

#### **4.2. Permissões para uso dos dados presentes no Twitter**

Um passo importante para o desenvolvimento do projeto prático foi a permissão para utilização dos dados presentes no Twitter, envolvendo todas as publicações com a privacidade pública feitas na história. Desse modo, fez-se necessário acessar o site do Twitter Developers, o qual exige um relatório justificando o uso da API disponibilizada pela empresa. Após a permissão ser concedida são liberadas quatro chaves identificadoras únicas, que precisam ser utilizadas para validação do uso das funções criadas pelo próprio Twitter.

Com isso, foi possível resgatar todos os “tweets” publicados por uma conta pública através da função “user\_timeline”. Nesse caso, a função foi preenchida com as informações da conta verificada do democrata Barack Obama. Além do nome do usuário, outros parâmetros que essa função exige são: quantidade de “tweets” a serem resgatados (em ordem cronológica, do mais novo para o mais antigo), linguagem a qual as publicações são feitas e como os textos serão resgatados (para esse estudo utilizei a versão mais completa das publicações).

#### **4.3. Expressões regulares**

Os textos resgatados através da função disponibilizada pela plataforma trazem diversos ruídos, dentre eles menções ou até mesmo referências aos usuários que publicaram o texto originalmente. As mesmas aparecem com a seguinte formatação no meio do texto: @exemplo, isso torna-se um “outlier” da informação desejada, visto que não agregam valor semântico a publicação feita na rede, assim como a indicação de compartilhamento, a qual vem com o texto “RT” no início do “tweet”. Outros exemplos que não agregam valor analítico ao texto são as próprias “hashtags” e hiperlinks, os quais podem facilitar na requisição das informações, mas não trazem significado a frase analisada e precisam ser removidas da publicação original.

Para a remoção de todos esses componentes que chegam junto com o texto quando é feita uma requisição, foi escolhido a utilização de expressões regulares, removendo objetos indesejados. Assim, foi criada uma função nomeada de “cleanTxt”.

```

def cleanTxt(text):
    text = re.sub(r"@[A-Za-z0-9_]+", "", text) # Remove menções
    text = re.sub(r"#", "", text) # Remove hashtags
    text = re.sub(r"RT[\s]+", "", text) # Remove a marcação de retweets
    text = re.sub(r"https?:\/\/\S+", "", text) # Remove hiperlink

    return text

```

**Figura 1. Função criada para remover o ruído dos textos das publicações.**

**Fonte: Autor.**

#### 4.4. Análise sentimental

Para um dos pontos mais cruciais propostos, como já citado anteriormente, foi utilizada a biblioteca “TextBlob”. Assim, há duas categorias de análises que podem ser feitas: a análise sentimental, que dentro da biblioteca é tida na função “polarity”, a qual retorna um número no intervalo [-1, 1], sendo -1 uma sentença extremamente negativa e 1 uma frase extremamente positiva. Além disso, podemos fazer a análise de subjetividade, ou seja, quantificar a quantidade de opinião dentro do respectivo pedaço de texto. No caso da subjetividade, utiliza-se a função “subjectivity”, a qual retorna um número no intervalo [0, 1], sendo 0 completamente isento de opinião e 1 representando o extremo oposto.

Utilizando algumas funções específicas da biblioteca matplotlib foi possível extrair de forma gráfica um resumo das publicações, tanto na perspectiva sentimental quanto na visão opinativa, feitas pelo perfil do Barack Obama no twitter.

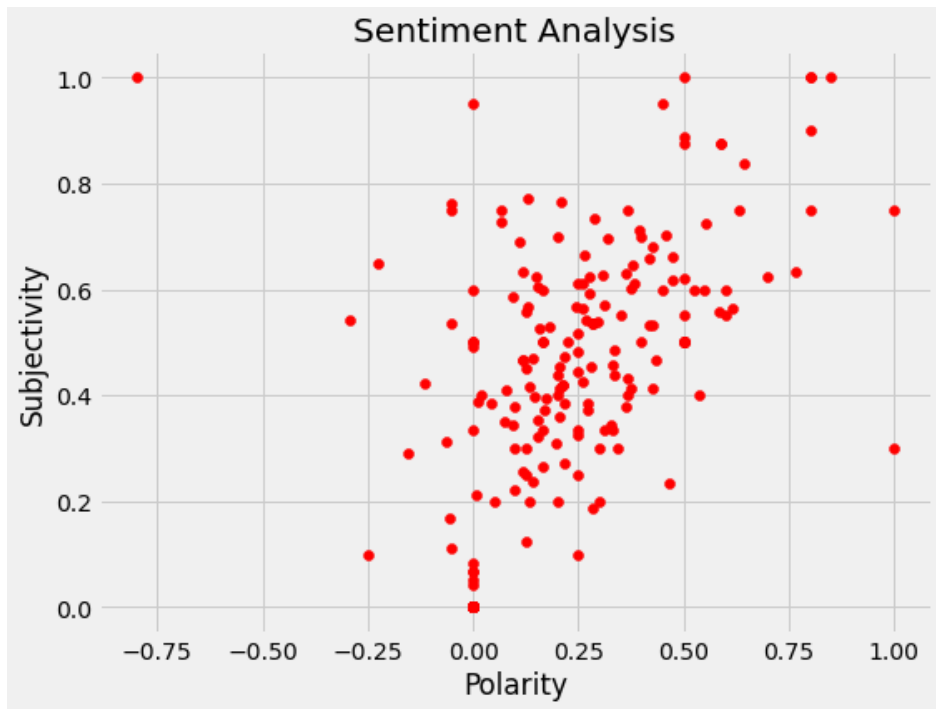
Além disso, podemos tirar de forma quantitativa as informações ligadas aos sentimentos das publicações, ou seja, mostrar diretamente quantas das publicações textuais resgatadas se enquadram como positivas, quantas neutras e quantas negativas, como na imagem 3. Vale ressaltar, que com o retorno menor que 0 a sentença será classificada como negativa, com retorno 0 a sentença é considerada neutra e maior que 0 é positiva.

## 5. Resultados

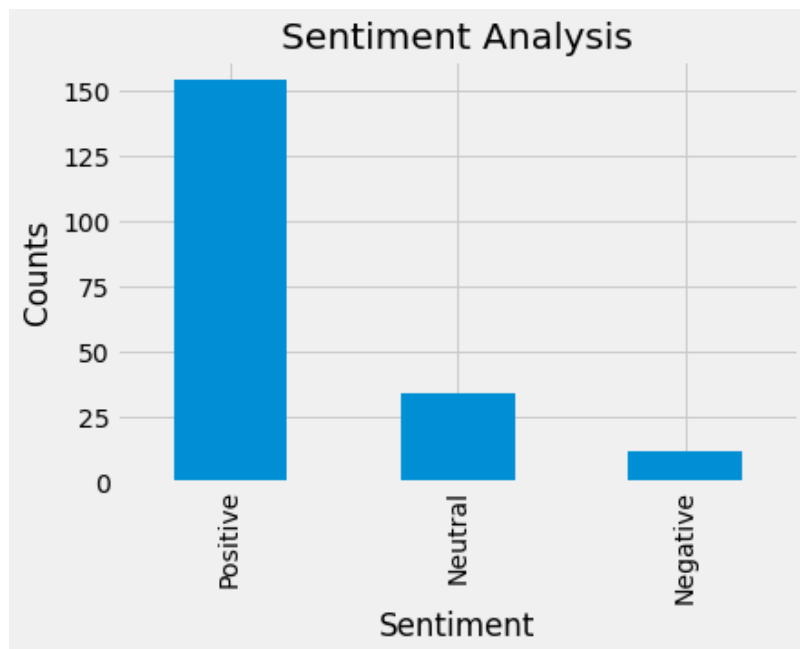
Quando o perfil do Barack Obama é colocado sobre análise no programa, é possível perceber através do gráfico de dispersão apresentado na figura 2, uma tendência de publicações majoritariamente positivas, seguidas por uma sequência neutra, tendência reforçada na exibição utilizada na figura 3. Isso aplicado ao momento em que os democratas se encontravam quando a análise foi feita, no dia 24 de fevereiro, é possível compreender o teor sentimental das publicações quando colocadas sob uma visão geopolítica global.

Além disso, o ex-presidente, detentor de um dos maiores perfis da rede social, publica fortemente com sua opinião, dado o seu alinhamento e profissão. Tendência que pode ser observado ao olhar para o eixo da subjetividade das publicações.





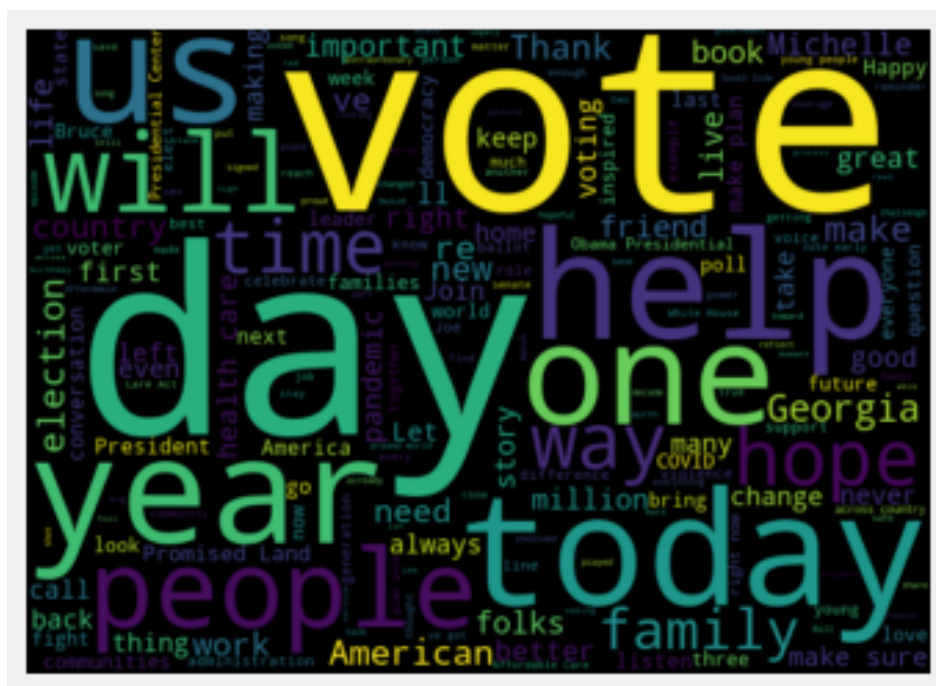
**Figura 2. Gráfico de dispersão da relação entre polaridade e subjetividade das publicações do Barack Obama.**  
**Fonte: Autor.**



**Figura 3. Gráfico de barras mostrando a quantidade de cada grupo de polaridade nas publicações do Barack Obama.**  
**Fonte: Autor.**

Como forma complementar, buscando atribuir sentido aos sentimentos presentes

nos “Tweets” analisados, pode-se observar a nuvem de palavras, a qual centraliza as palavras mais utilizadas nos textos publicados, sendo as que aparecem em mais destaque representando as mais repetidas no texto. Isso, alinhado aos gráficos e a um contexto histórico, quando analisado, é possível atribuir significado e valor ao que se foi recolhido e compreendido.



**Figura 4.** Nuvem de palavras mais citadas no perfil do Barack Obama.

Fonte: Autor.

Em suma, analisai-se publicações com opiniões pessoas fortes, a maioria voltada para um cunho positivo e quando observadas juntamente ao contexto da época em que foram feitas e olhando as palavras mais citadas (voto, ajuda, dia, experênça. . . ), pode-se entender melhor as sensações passadas pelas últimas postagens do democrata.

## 6. Conclusão

Com os dados apresentados nos capítulos anteriores é possível observar que caminhan-se para um mundo cada vez mais conectado, onde as pessoas compartilham cada vez mais seus posicionamentos e opiniões. Todo esse compartilhamento presente, torna-se cada vez mais fácil a extração desses dados.

Com os dados coletados, o próximo passo é a limpeza de ruído dos mesmos. Assim, teremos a informação necessária disponível, pronta para uma análise mais limpa e fiel à realidade, a qual pode ser demonstrada de diversas maneiras, com conclusões diversas. Os gráficos que são apresentados depois do processamento podem ser utilizados de forma complementar e alinhados a notícias e acontecimentos históricos podem trazer informações valiosas para análise.

Em suma, possui-se uma quantidade enorme de informações presentes gratuitamente na “internet”, muitas que podem aparentar não possuir nenhum valor, como uma

simples publicação no twitter. Todavia, quando organizamos os dados e tratamo-os de forma apropriada, podemos extrair valores intrínsecos presentes no meio.

No projeto prático documentado neste capítulo, foi feita a análise de um perfil com o objetivo de provar a capacidade de extrair informações relevantes. Pode-se observar, com uma análise mais superficial dos gráficos, otimismo e esperança, ligados aos mais diversos fatores: eleições, vacinação, povo americano, projetos do ano e futuro. Isso apenas com uma análise gráfica e temporal do perfil do ex-presidente dos EUA, o que mostra o potencial para analisar impactos de possíveis propagandas, influência de pessoas públicas, recepção das eleições, etc. As possibilidades, alinhadas ao atual poder computacional e conexão, são praticamente ilimitadas.

## 7. Trabalhos futuros

Existem inúmeras possibilidades quando o tema é mineração de dados e os benefícios que podem trazer a sociedade. Sendo assim, a aplicação desenvolvida pode ser melhorada para trazer comparações e um painel de informações em tempo real, com o propósito de acompanhar o impacto de uma campanha publicitária, por exemplo. Para isso, será necessário utilizar bibliotecas específicas para esse monitoramento de forma precisa e até mesmo a criação de um banco de dados para armazenamento das informações recolhidas e analisadas.

Além disso, busca-se realizar uma adaptação para que a biblioteca utilizada analise com mais eficiência textos produzidos na língua portuguesa.

Por fim, realizar os ajustes necessários para gerar uma aplicação executável que possa ser utilizada por profissionais de outras áreas além da computação. Como, por exemplo, psicólogos ou até mesmo publicitários, visto que a aplicação da solução abrange diversos interesses.

## 8. Referências

BUSINESS OF APPS. **Twitter Revenue and Usage Statistics (2021)**. Disponível em: <https://www.businessofapps.com/data/twitter-statistics/>. Acesso em: 28 abr. 2021.

CASTRO, L. N. D; FERRARI, Daniel Gomes. **Introdução à mineração de dados: Conceitos básicos, algoritmos e aplicações**. 1. ed. Rua Henrique Shaumann, 270, Pinheiros - São Paulo - SP: Saraiva, 2016.

GANDOMI, Amir; HAIDER, Murtaza. Beyond the hype: Big data concepts, methods, and analytics. **International Journal of Information Management**, ed Rogers School of Management, Ryerson University, Toronto, Ontario M5B 2K3, Canada, v. 35, n. 1, p. 137-144, dez./2014. Disponível em: [https://www.researchgate.net/publication/276248376\\_Beyond\\_the\\_hype\\_Big\\_data\\_concepts\\_methods\\_and\\_analytics](https://www.researchgate.net/publication/276248376_Beyond_the_hype_Big_data_concepts_methods_and_analytics). Acesso em: 20 jan. 2021.

HILBERT, Martin. Big Data for Development: A Review of Promises and Challenges. University of California, Davis, jan./2016. Disponível em: [https://www.researchgate.net/publication/286907720\\_Big\\_Data\\_for\\_Development\\_A\\_Review\\_of\\_Promises\\_and\\_Challenges](https://www.researchgate.net/publication/286907720_Big_Data_for_Development_A_Review_of_Promises_and_Challenges). Acesso em: 10 fev. 2021.

MAINIERI, B. O. *et al.* Text Mining of Audience Opinion in eSports Events. R. da Consolação, 930 - Consolação, São Paulo - SP, 01302-907, nov./2017. Disponível

em: [https://www.researchgate.net/publication/320471151\\_Text\\_Mining\\_of\\_Audience\\_Opinion\\_in\\_eSports\\_Events](https://www.researchgate.net/publication/320471151_Text_Mining_of_Audience_Opinion_in_eSports_Events). Acesso em: 18 nov. 2020.

NEW Tweets per second record, and how!. [S. l.], 5 ago. 2013. Disponível em: [https://blog.twitter.com/engineering/en\\_us/a/2013/new-tweetsper-second-record-and-how](https://blog.twitter.com/engineering/en_us/a/2013/new-tweetsper-second-record-and-how). Acesso em: 5 out. 2021

PYTHON DOCS. **Regular expression operations**. Disponível em: <https://docs.python.org/3/library/re.html>. Acesso em: 15 abr. 2021.

STATISTA. **Twitter - Statistics & Facts**. Disponível em: <https://www.statista.com/topics/737/twitter/#dossierSummary>. Acesso em: 28 abr. 2021.

TWITTER DEVELOPER. **How to analyze the sentiment of your own Tweets** . Disponível em: <https://developer.twitter.com/en/docs/tutorials/how-to-analyze-the-sentiment-of-your-own-tweets>. Acesso em: 5 abr. 2021.

TWITTER DEVELOPER. **Introducing a new and improved Twitter API**. Disponível em: [https://blog.twitter.com/developer/en\\_us/topics/tools/2020/introducing\\_new\\_twitter\\_api.html](https://blog.twitter.com/developer/en_us/topics/tools/2020/introducing_new_twitter_api.html). Acesso em: 25 mar. 2021.

YOUYOU, W. *et al.* Birds of a Feather Do Flock Together: Behavior-Based Personality-Assessment Method Reveals Personality Similarity Among Couples and Friends . **Sage Journals**, v. 28, n. 3, p. 276-284, jan./2017. Disponível em: <https://journals.sagepub.com/doi/10.1177/0956797616678187>. Acesso em: 12 nov. 2020.