Centro Universitário do Instituto Mauá de Tecnologia
Escola de Engenharia Mauá
Engenharia de Controle e Automação

Pesquisa de Intenção de Voto em Redes Sociais

Giovanni Lemos Maddalena
Matheus Yudy Kazama
Vivian Monaco Savioli

São Caetano do Sul, SP
2021

Giovanni Lemos Maddalena
Matheus Yudy Kazama
Vivian Monaco Savioli

Pesquisa de Intenção de Voto em Redes Sociais

Trabalho de Conclusão de Curso apresentado à Escola de Engenharia Mauá do Centro Universitário do Instituto Mauá de Tecnologia como requisito parcial para a obtenção dos títulos de Engenheiro de Controle e Automação.

Instituto Mauá de Tecnologia - IMT

Orientador: Prof. Ms. Alexandre Harayashiki Moreira

São Caetano do Sul, SP
2021

Lemos Maddalena, Giovanni

Pesquisa de Intenção de Voto em Redes Sociais/ Giovanni Lemos Maddalena, Matheus Yudy Kazama, Vivian Monaco Savioli. – São Caetano do Sul: CEUN-IMT, 2021.

35 p.

Trabalho de Conclusão de Curso - Escola de Engenharia Mauá do Centro Universitário do Instituto Mauá de Tecnologia, São Caetano do Sul, SP, 2021

Orientador: Prof. Ms. Alexandre Harayashiki Moreira

1. Pesquisa de intenção de voto. 2. Processamento de linguagem natural. 3. Computação em nuvem. 4. Machine Learning. 5. Eleições. I. Kazama, Matheus Yudy. II. Monaco Savioli, Vivian. III. Instituto Mauá de Tecnologia. Escola de Engenharia. V. Pesquisa de Intenção de Voto em Redes Sociais

Giovanni Lemos Maddalena
Matheus Yudy Kazama
Vivian Monaco Savioli

Pesquisa de Intenção de Voto em Redes Sociais

Trabalho de Conclusão de Curso apresentado à Escola de Engenharia Mauá do Centro Universitário do Instituto Mauá de Tecnologia como requisito parcial para a obtenção dos títulos de Engenheiro de Controle e Automação.

Prof. Ms. Alexandre Harayashiki Moreira
Orientador

Prof. Ms. Tiago Sanches
Avaliador

Prof. Dr. Fernando Silveira Madani
Avaliador

São Caetano do Sul, SP
2021

Agradecimentos

Aos familiares e amigos por todo o apoio e incentivo ao longo desses meses, que foram de vital importância para nos manter motivados e focados no desenvolvimento deste projeto.

Ao orientador, Professor Alexandre Harayashiki Moreira, pela dedicação em compartilhar seus conhecimentos, experiências e ideias. Além de sua dedicação e pronta disposição, ajudando plenamente em todas as fases do desenvolvimento deste trabalho. Sua ajuda foi indispensável não apenas na parte técnica, mas também em todos os assuntos pertinentes ao regulamento do Instituto Mauá de Tecnologia.

Por fim, a todas as pessoas que direta ou indiretamente contribuíram de alguma forma no desenvolvimento deste trabalho.

"We can only see a short distance ahead, but we can see plenty there that needs to be done."
(Alan Turing)

Resumo

Atualmente, durante o período eleitoral, são realizadas diversas pesquisas de intenção de voto afim de demostrar o cenário político. Estas pesquisas eleitorais são realizadas por diferentes empresas e a metodologia para a realização varia de acordo com a precisão desejada. Basicamente, o custo destas pesquisas é composto pelo como o eleitor será entrevistado (via telefone ou presencialmente) e a quantidade de eleitores a serem entrevistados. Com isso, este trabalho propõe uma metodologia alternativa para realizar uma pesquisa de intenção de voto. Esta nova metodologia consiste em coletar as opiniões políticas dos usuários da rede social Twitter e através de técnicas de processamento de linguagem natural, extrair se o conteúdo destas opiniões são favoráveis ou não ao candidato ao qual ela se refere. Para isso, foram utilizadas ferramentas disponíveis na plataforma de computação em nuvem da Amazon (AWS) e para compilar e facilitar a visualização dos resultados, os dados foram apresentados em um *dashboard* desenvolvido no software Microsoft PowerBI. Apesar das 160 mil publicações utilizadas na análise representarem apenas uma coleta em um período de 2 horas, foi possível observar grande correlação das intenções de votos destas publicações com os resultados do 1º turno das eleições de 2018. Assim, pode-se concluir que esta metodologia pode ser utilizada na realização de pesquisas de intenção de voto com a possibilidade de alcançar um número maior de pessoas, além de permitir a realização destas pesquisas quase que em tempo real.

Palavras-chaves: Pesquisa de intenção de voto. Processamento de linguagem natural. Computação em nuvem. Machine learning. Eleições.

Abstract

Currently, during the electoral period, several polls are conducted to show the political scenario. These electoral polls are carried out by different companies, and the methodology used varies according to the desired accuracy. Basically, the cost of these surveys is composed of how the voter will be interviewed (by phone or in person) and the number of voters to be interviewed. With this in mind, this paper proposes an alternative methodology to conduct a survey of voting intention. This new methodology consists in collecting the political opinions of Twitter social network users and, through natural language processing techniques, extracting whether the content of these opinions is favorable or not to the candidate to whom it refers. For this, tools available on Amazon's cloud computing platform (AWS) were used, and to compile and facilitate the visualization of the results, the data were presented in a dashboard developed in Microsoft PowerBI software. Although the 160,000 publications used in the analysis represent only one collection in a 2-hour period, it was possible to observe a great correlation of the voting intentions voting intentions of these publications with the results of the 1st round of the 2018 elections. Thus, it can be concluded that this methodology can be used in the conduct of polls of voting intentions with the possibility of reaching a larger number of people, besides the possibility of reaching a larger number of people, as well as allowing these surveys to be conducted almost in real time.

Keywords: Voting intention poll. Natural Language Processing. Cloud Computing. Machine learning. Elections.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Em todo o mundo, pesquisas de intenção de voto são feitas antes das eleições para mostrar o cenário atual das intensões de voto. Mas ao longo dos anos, muitas destas pesquisas mostraram-se falhas. O método precário de selecionar os eleitores a serem entrevistados, tende a colocar pessoas do mesmo sexo, mesma idade e mesma região como eleitores de mesma opinião, tornando o método falho. Além disso, o número de pessoas a serem entrevistadas depende do veículo de comunicação que está solicitando esta pesquisa.

De acordo com AgenciaBrasil (2014), órgãos de pesquisa como o Vox Populi, por exemplo, divide sua amostra em cinco estratos: idade, sexo, escolaridade, renda e ocupação, enquanto o Ibope divide em apenas quatro: idade, sexo, escolaridade e ramo de atividade.

Com isso, (MENEZES, 2018) e (G1, 2020) apresentam casos onde as previsões apontadas pelas pesquisas não representaram a real intensão de voto dos eleitores, mostrando que a metodologia utilizada pelos órgãos de pesquisa apresenta falhas.

Além disso, o custo envolvido na elaboração destas pesquisas podem ser muito alto, conforme apresentado na Figura 1

Figura 1 – Custo das Pesquisas Eleitorais de 2018.

Instituto	Registro	Divulgação	Valor	Código	Contratante	Amostra	Método	Custo/entrevistado
IPESPE	21/09/2018	27/09/2018	R$ 60.000,00	BR-00526/2018	XP Investimentos	2000	Telefone	R$ 30,00
MDA	24/09/2018	30/09/2018	R$ 179.467,00	BR-03303/2018	CNT	2002	Presencial	R$ 89,64
Vox Populi	18/09/2018	24/09/2018	R$ 83.441,00	BR-00793/2018	Folhamax	1000	Presencial	R$ 83,44
Ibope	25/09/2018	01/10/2018	R$ 347.653,33	BR-08650/2018	Globo, Estado	3010	Presencial	R$ 115,50
Paraná	20/09/2018	26/09/2018	R$ 67.500,00	BR-03512/2018	Empiricus	2020	Telefone	R$ 33,42
Data World	21/09/2018	27/09/2018	R$ 90.000,00	BR-06051/2018	Data World	1500	Presencial	R$ 60,00
Datafolha	22/09/2018	28/09/2018	R$ 398.344,00	BR-08687/2018	Folha, Globo	9072	Presencial	R$ 43,91
DataPoder360	21/09/2018	27/09/2018	R$ 56.853,00	BR-09543/2018	DataPoder360	3000	Telefone	R$ 18,95
FSB	25/09/2018	01/10/2018	R$ 65.000,00	BR-05879/2018	BTG Pactual	2000	Telefone	R$ 32,50

Fonte: MoneyTimes (2018)

Conforme observado na Figura 1, as pesquisas são realizadas de forma presencial ou por telefone. Porém, com o crescimento das redes sociais, muitos usuários dessas redes começaram a expressar suas opiniões políticas com maior frequência, além dos partidos também utilizarem a plataforma para expor suas campanhas políticas.

1.1 Objetivos

Realizar uma pesquisa de intenção de voto através de comentários extraídos de redes sociais.

A fim de alcançar esse objetivo, alguns objetivos específicos foram definidos:

- Desenvolvimento de uma metodologia para classificação dos comentários de redes sociais utilizando técnicas de *Natural Language Processing* (NLP);
- Desenvolvimento de um dashboard para visualização dos dados.

1.2 Justificativa

Realizar pesquisas de intenção de voto em redes sociais, além de trazer facilidade no acesso às opiniões de um número maior de eleitores de diferentes localidades, também possibilita entender os impactos causados por notícias recentes sobre um determinado candidato com maior velocidade.

1.3 Organização do Trabalho

Este trabalho está dividido em cinco grandes capítulos, com o primeiro sendo a introdução, motivação e objetivos deste trabalho. O segundo capítulo consiste em um estudo sobre trabalhos relacionados às pesquisas eleitorais em redes sociais e um estudo sobre NLP. O terceiro capítulo apresenta a metodologia proposta. Por fim, os testes e resultados são apresentados no quarto capítulo e as conclusões e possíveis melhorias são descritas no último capítulo.

2 Revisão Bibliográfica

Neste capítulo serão apresentadas e comparadas as técnicas utilizadas em trabalhos semelhantes, afim de se propor uma nova metodologia para pesquisas eleitorais em redes sociais. Além disso, também serão apresentados os conceitos de NLP para a classificação de comentários das redes sociais.

2.1 Previsibilidade Eleitoral Utilizando Machine Learning

O Trabalho desenvolvido por Brito e Adeodato (2020) apresenta que para se obter um resultado acurado da previsibilidade eleitoral de um presidente no Brasil, são necessários cinco passos: entendimento de negócios, entendimento, preparação, modelagem e avaliação dos dados.

Para o entendimento de negócios, um estudo é feito nos candidatos que estão concorrendo a presidência e quais os mais revelantes para a votação, ou seja, aqueles que receberam mais de 1% dos votos. Neste caso, foram separados para o estudo cinco candidatos.

Com os candidatos definidos, o entendimento de dados passa a ser o próximo passo do estudo, nesta etapa é visualizado a interação destes em redes sociais, sendo elas Twitter, Facebook e Instagram, levando em conta curtidas, comentários e compartilhamentos de suas publicações, mostrado na Figura 2.

Figura 2 – Interações das postagens dos candidatos.

Interaction	Total	Mean / Post	Median / Post	Std. Deviation
Facebook likes	81,395,302	13,481	4,131	29,987
Facebook shares	27,125,269	4,492	1,057	14,391
Facebook comments	12,561,981	2,080	421	17,491
Twitter likes	21,309,015	2,184	380	6,294
Twitter retweets	5,442,402	558	120	1,460
Instagram likes	100,777,503	31,691	6,826	85,591
Instagram comments	3,416,665	1,074	196	4,106

Fonte: Brito e Adeodato (2020)

Com esses dados apurados, é feita a preparação para cada candidato e cada *dataset* é criado com diferentes janelas agregadas para evitar uma seleção arbitrária de janelas. Foi aplicado análise de componentes principais (PCA) para que não ocorresse problemas de alta dimensionalidade e violação da dimensão VC durante os testes.

Após a criação do novo *dataset*, este foi executado em uma rede neural artificial com multicamadas que teve a implementação com a biblioteca scikit-learn em Python. Dois métodos de seleção de parâmetros foram testados: selecionar os parâmetros manualmente

e grade de busca por parâmetro. Dez previsões foram feitas para cada método utilizado e os critérios utilizados para a rede neural são citados na Figura 3.

Figura 3 – Critérios utilizados para cada rede neural artificial.

Parameter	Values
Hiddel Layer Sizes	3, 4, 5, 10
Activation Function	Logistic, Tanh
Solver	SGD, L-BFGS, ADAM
Alpha	0.00001, 0.001, 0.01, 0.05, 0.1
Learning Rate	Constant, Adaptive

Fonte: Brito e Adeodato (2020)

A partir dos resultados, foi feita uma avaliação comparando os métodos utilizados com os resultados parciais da eleição e com os obtidos através de uma Regressão Linear, conforme apresentado na Figura 4.

Figura 4 – Comparação de resultados de cada análise.

Candidate	Vote Share (%)	Pollsters (%)		Linear Regression (%)		ANN Fixed Param. (%)		ANN Grid Search(%)	
		Ibope	Datafolha	Mean	Median	Mean	Median	Mean	Median
Bolsonaro	32.6	36.0	36.0	35.8	35.9	32.6	32.5	33.4	33.5
Haddad	20.8	22.0	22.0	25.9	25.9	21.0	20.9	21.0	21.1
Gomes	8.8	11.0	13.0	10.9	11.1	10.2	10.3	10.2	10.4
Alckimin	3.4	7.0	7.0	6.4	6.4	5.7	5.7	5.4	5.5
Amoêdo	1.8	2.0	3.0	2.5	2.4	2.4	2.3	2.3	2.3
MAE		2.13	2.73	2.82	2.88	**0.90**	0.93	1.00	1.10
MAPE		0.32	0.48	0.37	0.38	0.24	0.24	**0.22**	0.24

Fonte: Brito e Adeodato (2020)

Com esta comparação, o resultado que chegou mais próximo ao real, foi do método com parâmetros inseridos manualmente. As siglas MAP e MAPE são para erro médio absoluto e erro percentual médio absoluto respectivamente, que são duas métricas que foram utilizadas para distinguir qual método teve o menor erro comparado com o resultado real.

2.2 Redes Sociais e Análise de Sentimentos

2.2.1 Eleições Presidenciais na Nigéria

Conforme Oyebode e Orji (2019), para que a previsibilidade eleitoral na Nigéria fosse feita a partir de redes sociais, dois métodos de detecção de sentimentos seriam testados, *Lexicon-based* e *Machine Learning*.

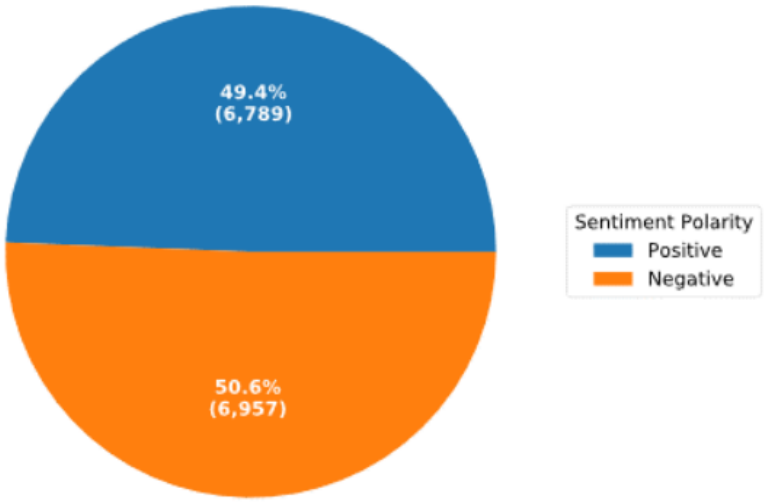
A primeira etapa para se completar a pesquisa é a coleta de dados, neste caso, por ser em um país que o Twitter não é muito utilizado, os dados foram obtidos da rede social Nairaland. Esta rede social não possui ferramentas como API para a coleta de dados, mas neste trabalho foi desenvolvido um *script* de *web scraping* feito em Python.

Com os dados coletados, o próximo passo é o pré-processamento destes, onde basicamente gerou um novo *dataset* com mudanças que ficam mais acessíveis para o computador ler, trocando siglas por palavras, excluindo espaços desnecessários e mais algumas mudanças para uma melhor leitura.

A próxima etapa é a detecção de sentimentos de cada comentário no novo *dataset*. Os sentimentos foram definidos como positivo, negativo e neutro para ambos métodos. Por meio de uma análise mais profunda entre *Machine Learning* e os adentos de *Lexicon-based*, foi constatado que o melhor método é o VADER-EXT abordado em *Lexicon-based* (OYEBODE; ORJI, 2019, p. 4).

O primeiro candidato observado, Mohammadu Buhari, obteve os sentimentos mostrados na Figura 5 após a execução do método VADER-EXT.

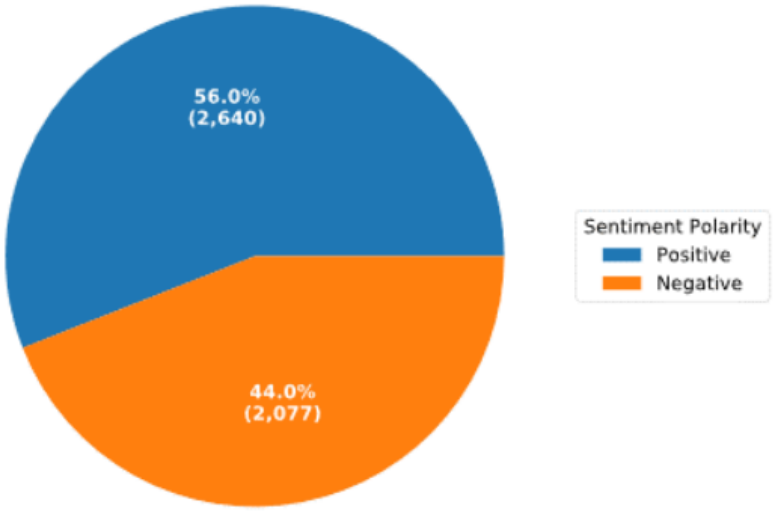
Figura 5 – Comparação de sentimentos positivos e negativos do candidato Mohammadu Buhari.



Fonte: Oyebode e Orji (2019)

Já o segundo candidato observado, Atiku Abubakar, obteve os sentimentos mostrados na Figura 6 após a execução do método VADER-EXT.

Figura 6 – Comparação de sentimentos positivos e negativos do candidato Atiku Abubakar.



Fonte: Oyebode e Orji (2019)

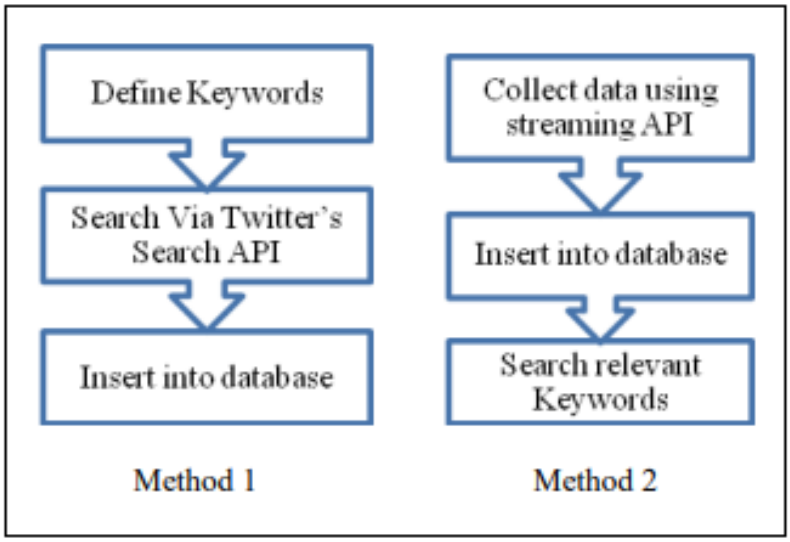
A comparação entre o resultado real da eleição e o resultado desta pesquisa não pode ser conclusiva, pois foi publicado anteriormente a eleição e não teve uma conclusão sucinta, mas de acordo com os dados apurados, o possível ganhador da eleição seria o candidato Atiku Abubakar, mas segundo a matéria publicada no veículo de comunicação Exame (2021), o presidente eleito foi o candidato Mohammadu Buhari, o que mostrou certas falhas na pesquisa feita pelo artigo.

2.2.2 Eleições Presidenciais nos Estados Unidos da América

Segundo Salunkhe e Deshmukh (2017), para obter a previsibilidade eleitoral para os Estados Unidos da América (EUA), são feitos quatro passo: coleta de dados, pré-processamento dos dados, análise dos sentimentos e apresentação dos resultados.

Para a coleta dos dados, o Twitter foi a rede social escolhida, onde pode-se filtrar e coletar esses dados por API, ou selecionando os *tweets* a partir de palavras chaves, como mostra a Figura 7 com os dois métodos de coleta de dados. O método escolhido foi por API, que selecionou os comentários relacionados aos dois candidatos que concorriam a presidência dos EUA, Hilary Clinton e Donald Trump.

Figura 7 – Organograma dos métodos de coleta de dados.

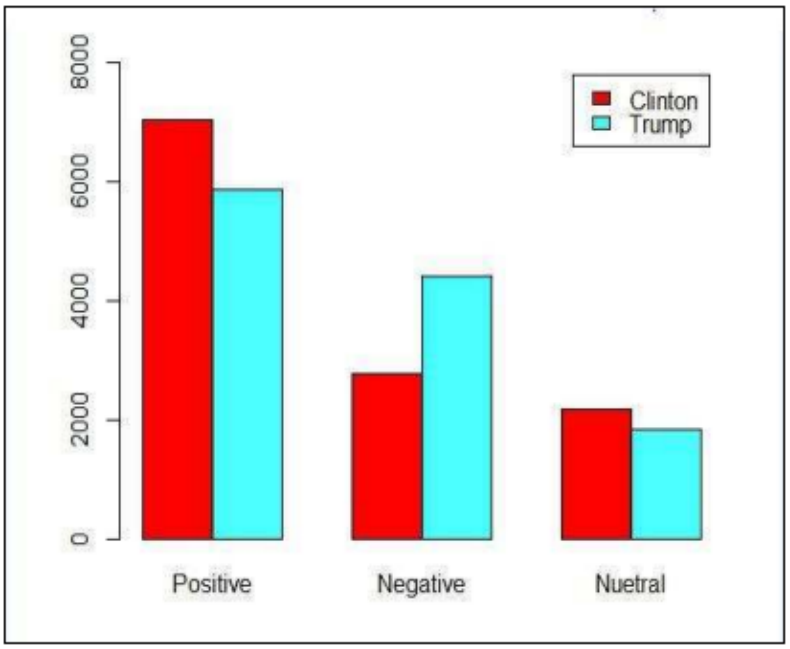


Fonte: Salunkhe e Deshmukh (2017)

Com a coleta desses dados, a fase de pré-processamento se inicia, onde transforma os comentários mais legíveis para os computadores, ou seja, convertendo todos os textos para caracteres minúsculos, removendo pontuações e números, removendo os espaços em branco e mais algumas técnicas que são utilizadas no pré-processamento criando um novo *dataset*.

Com o novo *dataset* pronto, a detecção de sentimentos foi feita pela abordagem *Lexicon-based*, os sentimentos foram divididos em três categorias, positivos, negativos e neutros. Os resultados obtidos foram dispostos em um gráfico, apresentado na Figura 8.

Figura 8 – Resultados dos sentimentos para cada candidato.



Fonte: Salunkhe e Deshmukh (2017)

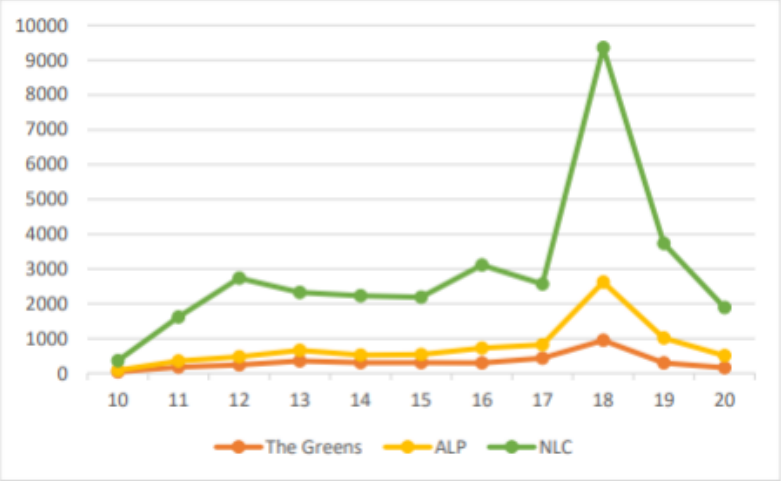
Os resultados obtidos nesta pesquisa coloca Hilary Clinton como presidente eleita, mas na realidade, quem ganhou esta eleição foi Donald Trump. Este erro pode ser previsível no caso dos EUA, pois o voto não é obrigatório como é no Brasil, a maioria dos usuários do Twitter são pessoas mais jovens e a própria eleição é feita de um jeito bem diferente, mesmo Hilary Clinton tendo recebido mais votos nas urnas, o que aconteceu, ela acabou perdendo a eleição por conta dos delegados que compõe o Colégio Eleitoral. Com esta curta explicação, os resultados fazem sentido, mas como o presidente não é eleito pelas urnas, a pesquisa acaba sendo falha para esta situação.

2.3 Previsibilidade Eleitoral Utilizando Erro Médio Absoluto e Raiz do Erro Quadrático Médio

De acordo com Budiharto e Meiliana (2018), para prever o resultado da eleição na Austrália segundo dados das redes sociais, foi utilizado dois métodos matemáticos: Erro Médio Absoluto (MAE) e a Raiz do Erro Quadrático Médio (RMSE).

A rede social utilizada para a coleta dos dados foi o Twitter, onde foi utilizado o API para filtrar e coletar os dados. Com isso, a Figura 9 mostra o fluxo de citações sobre os partidos políticos.

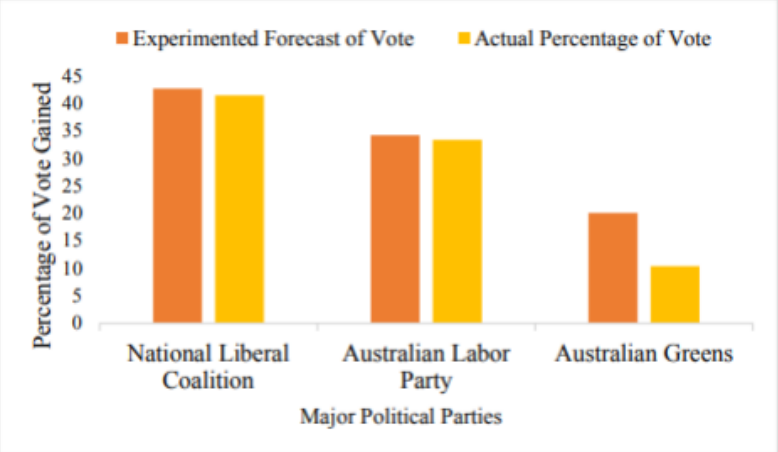
Figura 9 – Fluxo de citações sobre os partidos políticos.



Fonte: Budiharto e Meiliana (2018)

Com os dados coletados, dois métodos matemáticos foram utilizados para fazer as porcentagens de cada partido, estes resultados foram exibidos em forma gráfica comparando os resultados obtidos com os resultados reais das eleições, mostrado na Figura 10.

Figura 10 – Comparação entre resultados obtidos dos dados do Twitter e da eleição.



Fonte: Budiharto e Meiliana (2018)

Os resultados foram muito parecidos com o da votação, concluindo assim que o método utilizado foi bem eficaz apesar de sua simplicidade.

2.4 Comparativo

A seguir, será apresentada uma Tabela 1 que demonstra as técnicas usadas nos artigos apresentados para uma melhor visualização dos métodos utilizados e qual obteve o melhor resultado.

Tabela 1 – Comparativo entre os trabalhos apresentados.

Artigos	DM	RNA	Lexicon-based	ML	MM	Resultados Corretos
2.1	x	x		x	x	x
2.2.1	x		x			
2.2.2	x			x		
2.3	x				x	x

A Tabela 1 apresenta um comparativo entre as bibliografias apresentadas mostrando quais métodos cada uma delas apresenta como solução para a previsibilidade eleitoral. Os métodos desta são: *Data Mining* (DM), Redes Neurais Artificiais (RNA), *Lexicon-based*, *Machine Learning* (ML) e Modelos Matemáticos (MM). Como é ilustrado, os únicos métodos que não obtiveram um resultado acertivo ao da real eleição foram o 2.2.1 e o 2.2.2 que utilizaram análise de sentimentos.

Ambos os trabalhos utilizaram uma técnica de processamento de linguagem natural (NLP) manualmente, retirando caracteres especiais, espaços em branco e outros caracteres para que o computador compreendesse melhor os comentários, mas como forma feitos de forma manual, isso pode ter prejudicado no resultado destes.

2.5 Processamento de Linguagem Natural

O processamento de linguagem natural é a tecnologia usada para ajudar dispositivos tecnológicos a entenderem a linguagem do ser humano de maneira a responder suas demandas (TAKEBLIP, 2021).

Esta tecnologia é utilizada em assistentes pessoais, como por exemplo a SIRI do sistema IOS, *chatbots* e previsões de pesquisa, como no google. O NLP, pode entender intenções, reconhecer pessoas, lugares, datas e muitos outras entidades.

Neste trabalho, será feita algumas mudanças no método de classificação de sentimentos, pois foram os únicos que erraram a previsão, com isso, será implementado o método NLP para tentar obter um melhor resultado e o mais próximo ao real.

3 Metodologia

Todo o desenvolvimento técnico do trabalho foi realizado através da plataforma de computação em nuvem da Amazon (Amazon Web Service - AWS) (AMAZON, 2021). A AWS vende serviços como a locação de servidores e disponibiliza alguns serviços gerenciados como os que serão abordados ao longo deste trabalho.

Todas as interações com a AWS foram feitas pelo navegador web, através do console. Todos os serviços funcionam pela chamadas de APIs que podem ser feitas programaticamente (por código) ou no console web da plataforma.

As APIs são interfaces para programação de aplicações, e são utilizadas na integração de sistemas. Todas interações nas aplicações, como a leitura de textos, postagem de textos, encaminhamento de mensagens pode ser feita através de linhas de código, facilitando o desenvolvimento de aplicações que dependem de outras e dando a possibilidade da automatização de processos que seriam por interações humanas.

3.1 Implementação atual

O trabalho utilizou como *dataset* uma base de dados que se encontra no Kaggle que originalmente foi extraída pela API do Twitter, contendo tuítes do período eleitoral de 2018 e todos com algum conteúdo político. O *dataset* conta com apenas 3 colunas, sendo elas: ID, *timestamp* (hora do *tweet*) e o conteúdo do *(*tweet).

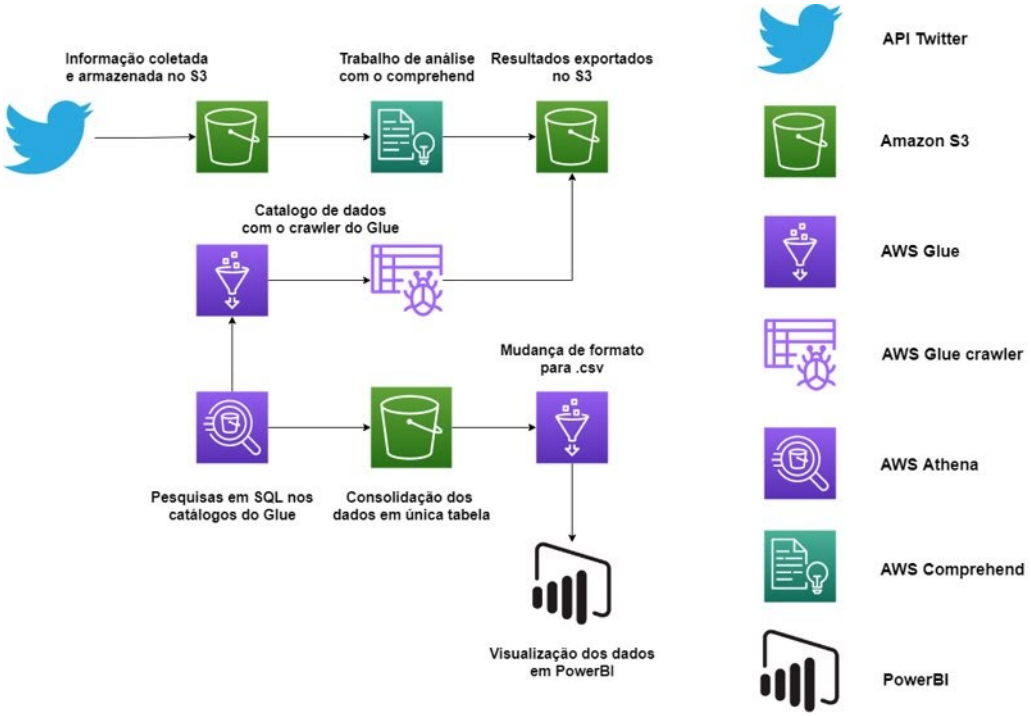
O Kaggle é um website que possui diversas bases de dados e códigos para serem trabalhados em projetos de ciência de dados. A plataforma conta com a colaboração de outros usuários para a distribuição destes conteúdos.

O *dataset* utilizado (NATANAEL, 2018), possui 187 mil *tweets* e todos estão relacionados com o cenário político do Brasil em 2018. A coleta tem como tempo inicial 01:11 05/10/2018 e término 02:34 05/10/2018 da janela de tempo da aquisição destes dados na API do Twitter.

3.1.1 Arquitetura final na AWS

A arquitetura na Figura 11 demonstra como todos os componentes abordados no capítulo 3 estão orquestrados na AWS.

Figura 11 – Arquitetura na AWS.



Fonte: Autores (2021)

3.2 Tratamento de dados

Dependendo de como os dados extraídos são estruturados é necessário realizar transformações antes de serem manipulados. Outro ponto é a filtragem de dados não relevantes ou incorretos que podem afetar a performance dos resultados. Dentre as técnicas para pré-processamento estão a remoção de valores em duplicata, remoção de valores nulos e a remoção de ruídos. O esquema de dados deve ser compatível com as integrações da aplicação.

Foi feita uma verificação da base de dados utilizada e foram verificados que diversos *tweets* possuíam caracteres de espaço e de quebra de linha. Arquivos no formato .csv reconhecem uma entrada de dado a cada linha, portanto é importante remover todos as quebras de linha da coluna de *tweets*. Além disso, a precificação de um dos serviços utilizados se baseia na quantidade de caracteres escaneados, portanto a remoção dos espaços extras é crucial já que não afeta a inferência de dados.

O arquivo foi importado para o Microsoft Excel no formato *comma delimited* e então foi utilizada a ferramenta de buscar e substituir primeiro nos caracteres de quebra de linha por nenhum caractere. Em seguida uma nova coluna foi adicionada com a função TRIM(), que normaliza os caracteres de espaço, referenciando a coluna de *tweets*. Por fim uma nova aba foi aberta e apenas a coluna de *tweets* foi colada ali, e o arquivo exportado como .csv.

3.3 Carregamento dos dados na AWS

Antes de qualquer operação com dados, eles devem ser carregados na nuvem, uma vez que toda a infraestrutura que detém os serviços também está na nuvem, para essa etapa será utilizado dos serviços de armazenamento de objetos Amazon S3 que armazena seus dados na nuvem.

Os dados podem ser carregados através do console com o limite de até 50GB por carregamento, de forma programática através da SDK ou pela linha de comandos (CLI).Os dados devem residir na nuvem por requerimento de execução nas próximas etapas.

Primeiramente deve-se criar uma especie de "repositório"no serviço S3, que são os *buckets*. Um *bucket* foi criado com as configurações padrões do serviço, nada foi alterado além do nome do *bucket*. O *dataset* adquirido no Kaggle possuía um tamanho inferior a 1GB e foi carregado manualmente via console pela opção de carregar arquivos dentro do *bucket*.

3.4 Inferência dos dados

Na parte de inferência dos resultados foi utilizado um serviço gerenciado pela AWS, o Amazon Comprehend. O Amazon Comprehend trata-se de um serviço de NLP com alguns modelos pré-treinados disponíveis, entre estes modelos 2 foram escolhidos para a realização do trabalho. Dentre os modelos pré-treinados, são estes: reconhecimento de entidades, frases-chave, PII (Informações pessoais sensíveis), idioma, sentimento e sintaxe. Os 2 modelos utilizados foram o reconhecimento de entidades e reconhecimento de sentimentos. Além disso, há a possibilidade de treinar um próprio modelo classificador, sendo necessário uma base de dados propriamente rotulada para treinar o modelo.

A Amazon não divulga dados sobre a rede neural utilizada pelo Comprehend nem outros dados como precisão e acurácia do modelo.

Todos os arquivos providenciados como base de dados devem estar em texto UTF-8 ou documentos semi estruturados, como documentos PDF e Word.

Na página do serviço Comprehend foram criados dois trabalhos de análise, um para análise de sentimentos e outro para análise de entidades. Ambos foram criados com as seguinte configurações: Fonte de dados - meus documentos, localização do S3 - (Nome do *bucket*), entrada de dados - arquivo .csv da primeira etapa, saída de dados - *bucket* criado na etapa anterior, formato de entrada - um documento por linha. Depois na parte de permissões foi criada uma função com permissões de leitura e escrita em *buckets* de S3. Por fim o trabalho de análise foi iniciado para cada finalidade, e concluído em cerda de 10 minutos.

3.4.1 Reconhecimento de entidades

O reconhecimento de entidades pode reconhecer diversas entidades em uma única entrada de texto. As entidades reconhecidas são: pessoas, lugares, itens comerciais, e referências precisas a medidas como datas e quantidades.

Para cada entidade reconhecida um valor de score normalizado é apresentado indicando o nível de confiança naquela predição. Esse *score* pode ser utilizado para computar o nível de certeza sobre o desenvolvimento do trabalho ou até mesmo fazer a filtragem de valores que possuem um *score* baixo. No trabalho todos os dados foram utilizados independente de seu *score*.

Outra saída de dado gerada é a categoria da entidade, classificando como pessoa, lugar, data, eventos, entre outros.

3.4.2 Reconhecimento de sentimentos

A detecção de sentimentos do Comprehend pode detectar entre sentimentos positivos, negativos, neutros e mistos. A detecção detecta apenas um teor de sentimento por entrada de texto, diferentemente do reconhecimento de entidades que podem ter diversas entidades em uma entrada.

Assim como o reconhecimento de entidades, cada sentimento reconhecido tem um *score* que representa a probabilidade do sentimento ter sido detectado corretamente.

3.5 Manipulação dos resultados

Os arquivos provenientes do processo de inferência possuem um formato JSON,são documentos contendo o resultado possuindo a estrutura chave valor. Antes da visualização gráfica dos resultados é necessário realizar algumas transformações. A primeira etapa consiste num mapeamento do esquema de dados obtidos com o resultado. O serviço AWS Glue conta com *crawlers* para mapeamentos de dados armazenados no S3. Um *crawler* foi criado para um mapeamento de todos os resultados em colunas de valores, essas colunas representam as chaves dos arquivos JSON. Este processo gera um catálogo de dados que serão utilizados na próxima etapa. O *crawler* foi criado especificando o caminho para o trabalho de *"crawl"* no *bucket* de S3 com os resultados extraídos do Comprehend, uma função foi criada para dar as devidas permissões de leitura de objetos no S3 e a agenda de execução foi configurada para sob demanda. Na parte de saída de dados, um banco de dados foi criado para guardar a saída dos *crawlers*. Essas operações geraram duas tabelas: *entities-results* e *sentiments-results*. O segundo passo é a execução de comandos SQL através do Amazon Athena que só é possível graças à primeira etapa. O Amazon Athena por si é o serviço de consultas SQL sobre arquivos que residem no S3 ou catálogos do Glue.

Primeiro foi feito o a mudança do esquema das tabelas *entities-results* e *sentiments-results*, desdobrando a coluna de dados de resultados em várias colunas separadas com os seguintes comandos no Athena:

```
CREATE TABLE sentiment_results_final AS
SELECT file, line, sentiment,
sentimentscore.mixed AS mixed,
sentimentscore.negative AS negative,
sentimentscore.neutral AS neutral,
sentimentscore.positive AS positive
FROM sentiment_results
```

```
CREATE TABLE entities_results_1 AS
SELECT file, line, nested FROM entities_results
CROSS JOIN UNNEST(entities) as t(nested)
```

```
CREATE TABLE entities_results_final AS
SELECT file, line,
nested.beginoffset AS beginoffset,
nested.endoffset AS endoffset,
nested.score AS score,
nested.text AS entity,
nested.type AS category
FROM entities_results_1
```

Agora com as duas tabelas bem formatadas podemos unir os resultados em uma única tabela com o seguinte comando:

```
CREATE TABLE results_final AS
SELECT *
FROM sentiments_results_final
LEFT JOIN entities_results_final
ON sentiments_results_final.line = entities_results_final.line
```

A última operação é consultar apenas valores que possuem a categoria de entidade igual a pessoas.

```
1 SELECT *
2 FROM results_final
3 WHERE results_final.category = PERSON
```

3.6 Visualização gráfica no Power BI

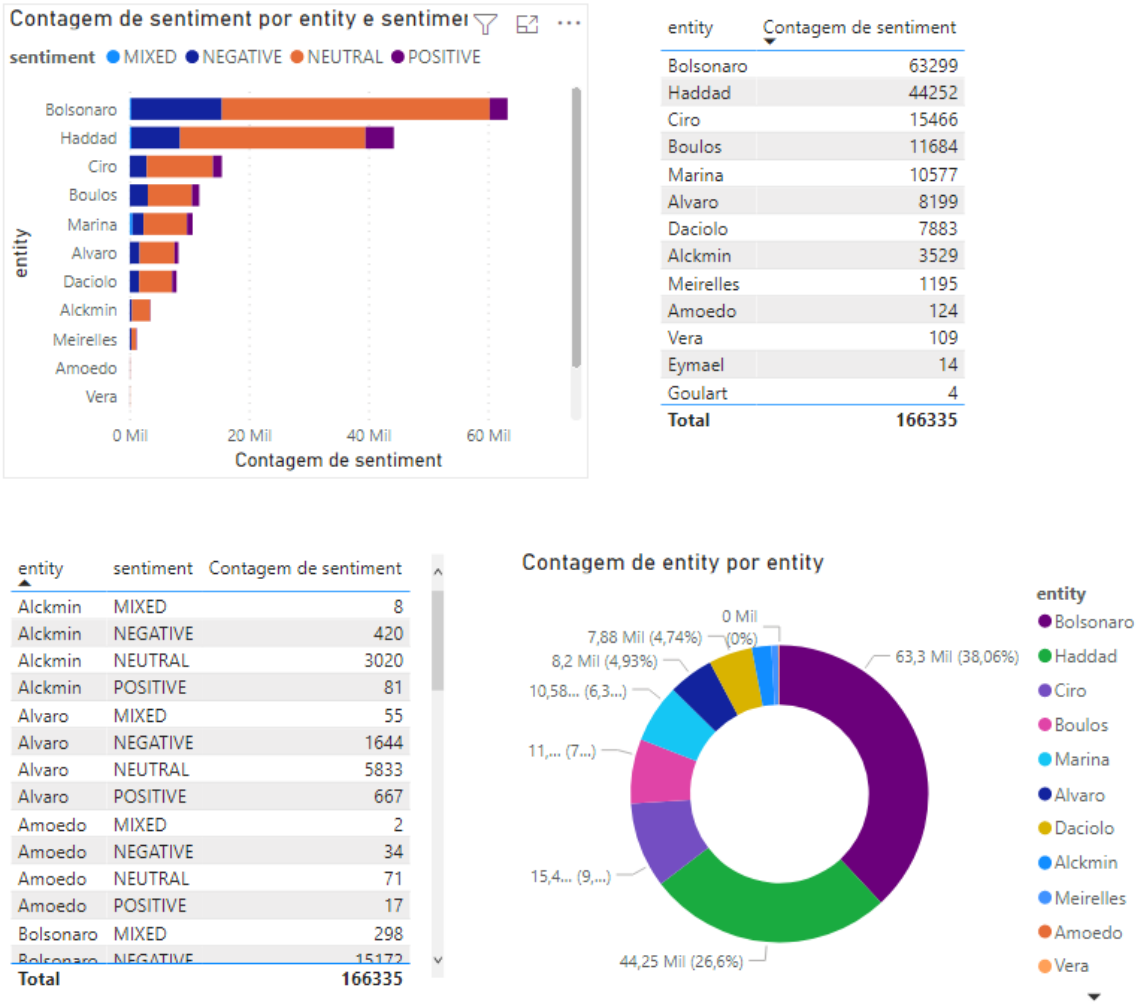
Com a tabela final é possível exportar como um arquivo .csv pelo console do Athena e finalizar a seleção de entidades que representam apenas os candidatos da eleição. Abrindo o arquivo no Microsoft Excel é gerada uma tabela dinâmica é feita a contagem de entrada de dados que tem-se por entidade encontrada no Comprehend. Esta etapa é necessária para retirar pessoas mencionadas que não são os candidatos e agrupar algumas menções feitas por pseudônimos e apelidos comuns a um único nome.

Foi estabelecido um limiar para apenas entidades que aparecem, no mínimo, 100 vezes na tabela dinâmica. Uma tabela dinâmica foi gerada para contagem de menção por entidade, e todas as entidades acime de 100 menções foram consolidadas com a ferramenta buscar e substituir em 13 candidatos, às entidades irrelevantes foram apagadas da base de dados final.

Após todo o tratamento de dados necessários, tem-se um total de 166 mil sentimentos para os 13 candidatos da eleição de 2018, alguns deles com números muito baixos pela sua baixa popularidade.

Para uma melhor visualização do resultado obtido na análise de sentimento foi desenvolvido um *dashboard* utilizando o software PowerBI (MICROSOFT, 2021), ilustrado na Figura 12.

Figura 12 – Dashboard elaborado em PowerBI.



Fonte: Autores (2021)

3.7 Custos para desenvolvimento do trabalho

A Figura 13 demonstra todos os custos envolvidos no projeto detalhados por cada serviço utilizado na plataforma.

Figura 13 – Custo total do trabalho.

Serviço	Out-19	Out-20	Out-21	Serviço Total
Custo total ($)	29.85	0.02	39.09	68.96
Comprehend ($)	29.83	0.00	38.19	68.02
Glue ($)	0.02	0.02	0.90	0.93
S3 ($)	0.00	0.00	0.01	0.01
Athena ($)			0.00	0.00

Fonte: Amazon (2021)

Cada serviço tem sua base de cálculo dependendo de seu uso.

3.7.1 Cálculo de custos com Comprehend

O Comprehend tem um custo baseado na quantidade de unidades que serão processadas pelos servidores. Uma unidade equivale a 100 caracteres de um arquivo de texto. A Tabela 2 demonstra os custos por unidade.

Tabela 2 – Preço por unidade (USD).

Recurso	Até 10 milhões	10 a 50 milhões	+ 50 milhões
Análise sentimentos	0,0001	0,00005	0,000025
Reconhecimento entidades	0,0001	0,00005	0,000025

3.7.2 Cálculo de custos com Glue

O Glue tem o custo baseado no tempo de execução dos *crawlers* e quantas unidades de processamento foram alocadas para certo serviço. O custo é de 0,44 USD por Unidade de processamento por hora, com a cobrança mínima de 10 minutos. Para o desenvolvimento do trabalho, os *crawlers* estavam em uma unidade de processamento e demoraram 1 minuto para a execução

Além disso, o custo para armazenar os catálogos de dados é gratuito para o primeiro milhão de dados, depois é de 1 USD por 100.000 dados. O trabalho teve um total armazenado de 663 mil dados, mantendo-se em nível gratuito.

3.7.3 Cálculo de custos com S3

O S3 tem o custo baseado no volume de dados armazenados no serviço e na quantidade e tamanho das requisições para recuperação dos arquivos. Existe um nível gratuito de 5GB de armazenamento mensal e 20.000 requisições para objetos. O trabalho não gerou nenhum custo com S3 já que os arquivos não ultrapassaram o tamanho de 1GB. Abaixo estão as tabelas 3 e 4 de custo por armazenamento e requisição.

Tabela 3 – Custo armazenamento S3.

S3 standard	preço por armazenamento
Primeiros 50TB/mês	0,023 USD por GB
Próximos 450TB/mês	0,022 USD por GB
Mais de 500TB/mês	0,021 USD por GB

Tabela 4 – Preço por requisição (USD).

	Solicitações PUT	Solicitações GET
S3 Standard	0,005 por milhar	0,0004 por milhar

3.7.4 Cálculo de custos com Athena

O Athena tem seu custo baseado no volume de dados escaneados por consulta SQL. A cobrança será feita por número de *bytes* verificados pelo Amazon Athena, arredondada para cima para o *megabyte* mais próximo, com um mínimo de 10MB por consulta. O preço é de 5 USD por TB escaneado.

4 Resultados

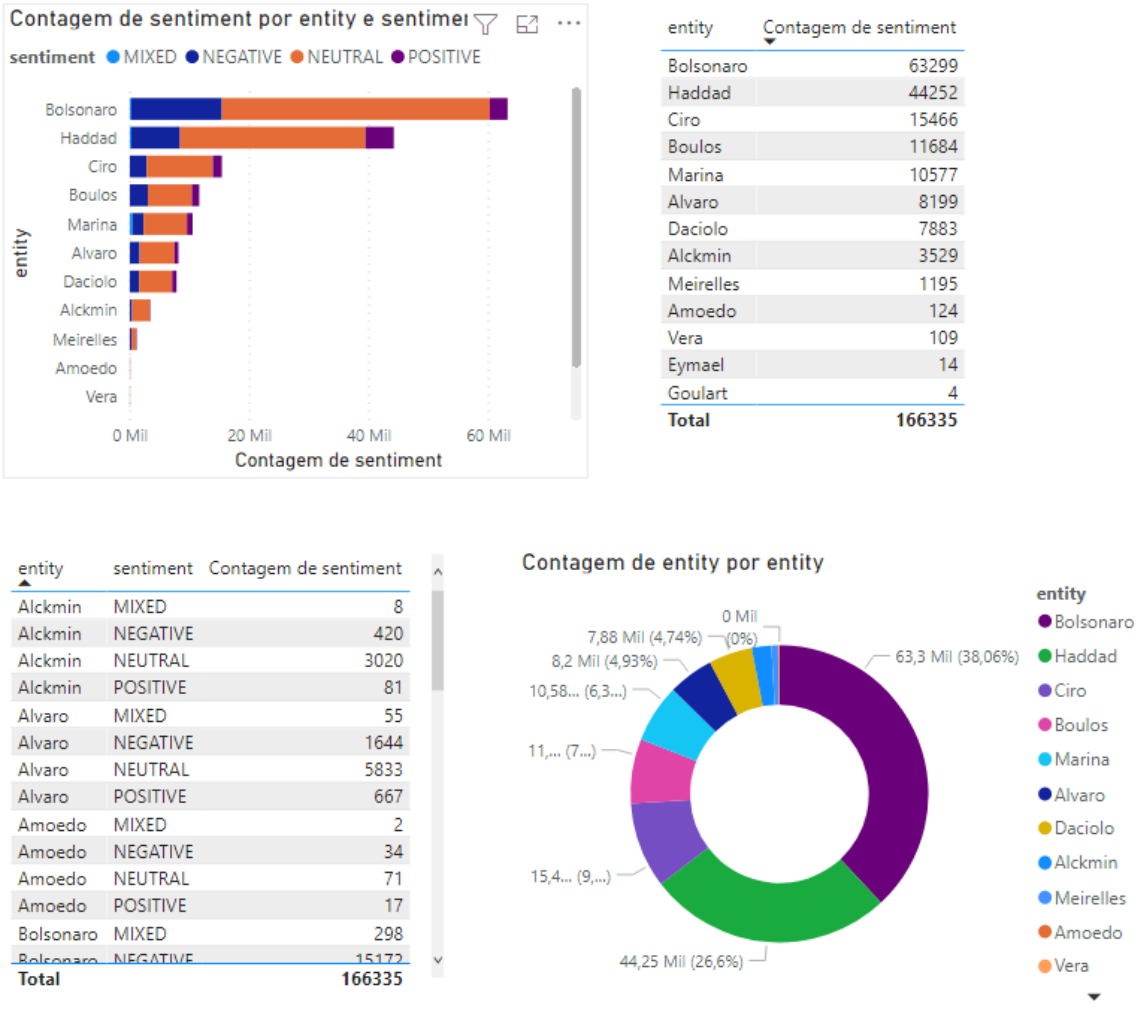
Para validar a tese do trabalho de pesquisa de intenção de voto em redes sociais, será utilizado o resultado real da eleição presidencial ocorrida em 2018, nessa o resultado levou a disputa para o 2º turno, pois o candidato mais votado não obteve mais de 50% dos votos validos computados (TSE, 2021). A Tabela 5 apresenta os resultados oficiais e a porcentagem de votos validos direcionados a cada candidato à presidência.

Tabela 5 – Resultados da votação no primeiro turno.

Candidato	Porcentagem
Jair Bolsonaro	46,03%
Fernando Haddad	29,28%
Ciro Gomes	12,47%
Geraldo Alckmin	4,76%
João Amoedo	2,50%
Cabo Daciolo	1,26%
Henrique Meirelles	1,20%
Marina Silva	1,00%
Alvaro Dias	0,80%
Guilherme Boulos	0,58%
Vera	0,05%
Eymael	0,04%
João Goulart Filho	0,03%

Ao longo do desenvolvimento do trabalho, foram realizadas análises de sentimento para cada comentário relacionado aos candidatos à presidência em 2018. Nessas análises era possível determinar se o comentário era positivo, negativo, neutro ou misto. Destes resultados uma análise foi feita e um *dashboard* foi montado com dois gráficos, um deles com os sentimentos dos comentários, positivos, negativos, neutros e mistos para cada candidato e um com a porcentagem de comentários positivos e neutros que cada um recebeu, como ilustrado na Figura 14.

Figura 14 – Dashboard elaborado em PowerBI.



Fonte: Autores (2021)

A Tabela 6 foi feita com base na análise de sentimento para cada candidato a partir dos comentários da rede social, contabilizando o número de cada sentimento dos comentários e o número total destes. Jair Bolsonaro, Fernando Haddad e Ciro Gomes foram os que mais receberam comentários respectivamente, ilustrando a popularidade destes candidatos, tanto positiva, quanto negativa.

Tabela 6 – Contagem e classificação de sentimentos por candidato

	Sentimentos				
Candidatos	Positivo	Neutro	Negativo	Misto	Total
Alckmin	81	3020	420	8	3529
Alvaro	667	5833	1644	55	8199
Amoedo	17	71	34	2	124
Bolsonaro	2989	44840	15172	298	63299
Boulos	1237	7372	3018	57	11684
Ciro	1506	11065	2828	67	15466
Daciolo	739	5526	1575	43	7883
Eymael	1	10	3	0	14
Goulart	0	3	1	0	4
Hadad	4750	31040	8143	319	44252
Marina	949	7280	1821	527	10577
Meirelles	106	750	338	1	1195
Vera	1	69	39	0	109

Com esta análise e somatória de sentimentos, foi possível retirar a porcentagem que cada candidato recebeu somando os comentários positivos e neutros e em seguida calculando a porcentagem sobre o total de postagens sobre cada respectivo político, ilustrado na Tabela 7.

Tabela 7 – Resultados da análise de sentimento em porcentagem.

Candidatos	Porcentagem de Sentimentos
Jair Bolsonaro	36,81
Fernando Haddad	27,55
Ciro Gomes	9,69
Guilherme Boulos	6,63
Marina Silva	6,33
Alvaro Dias	5,00
Cabo Daciolo	4,82
Geraldo Alckmin	2,39
Henrique Meirelles	0,66
João Amoedo	0,07
Vera	0,05
Eymael	0,004
João Goulaert Filho	0,001

Com essas porcentagens, foi possível comparar estes resultados com os do primeiro turno da eleição de 2018, apresentados na Tabela 8.

Tabela 8 – Comparação entre o resultado da eleição e do trabalho.

Eleição 2018		
Candidatos	Resultado 1º Turno (%)	Resultado análise de sentimento (%)
Jair Bolsonaro	46,03	36,81
Fernando Haddad	29,28	27,55
Ciro Gomes	12,47	9,68
Geraldo Alckmin	4,76	2,39
João Amoedo	2,50	0,07
Cabo Daciolo	1,26	4,82
Henrique Meirelles	1,20	0,66
Marina Silva	1,00	6,33
Alvaro Dias	0,80	5,00
Guilherme Boulos	0,58	6,63
Vera	0,05	0,05
Eymael	0,04	0,004
João Goulart Filho	0,03	0,001

É possível observar uma relação entre os resultados obtidos com o percentual de votos, porém pelo período de amostragem ser de apenas 2 horas, a porcentagem obtida pela análise de sentimento apresentou tal discrepância.

Além disso, também é possível relacionar que a popularidade dos candidatos na rede social é refletida na quantidade de votos obtida.

5 Conclusão

Atualmente as pesquisas de intenção de voto, dependem do cenário político em que foram realizadas, do número de pessoas entrevistadas e demandam um tempo para serem elaboradas. Além desses fatores, uma pesquisa de intenção de voto envolve altos custos.

Comparando os resultados entre o real e o retirado da análise de sentimento, é possível observar algumas semelhanças, como os três primeiros candidatos que receberam as maiores porcentagens de votos válidos também são os mais comentados na rede social observada, e que os três últimos, com a menor porcentagem de votos válidos, também foram os menos populares na rede social.

Desta forma, é possível observar que a metodologia utilizada pode representar a intenção de voto a partir das opiniões extraídas das rede social. Entretanto, algumas melhorias devem ser implementadas.

Uma primeira melhoria é a coleta de opiniões em tempo real para que após a divulgação de notícias que impactam diretamente a imagem de um candidato, o cenário de intenções de voto possa ser analisado de uma forma mais rápida.

Outra melhoria seria o desenvolvimento de algoritmos de NLP, além da comparação e utilização de hospedagem em diferentes serviços de computação em nuvem, diminuindo os custos envolvidos.

Por fim, desenvolver um algoritmo de *data mining* para a coleta de dados em diferentes fontes, como redes sociais, jornais, matérias, etc.

Referências

AGENCIABRASIL. *Saiba como são feitas as pesquisas de intenção de voto*. 2014. Disponível em: <https://agenciabrasil.ebc.com.br/politica/noticia/2014-09/saiba-como-sao-feitas-pesquisas-de-intencao-de-voto>.

AMAZON. *Amazon Web Service*. 2021. Disponível em: <https://aws.amazon.com>.

BRITO, K. dos S.; ADEODATO, P. J. L. Predicting brazilian and us elections with machine learning and social media data. In: IEEE. *2020 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2020. p. 1–8.

BUDIHARTO, W.; MEILIANA, M. Prediction and analysis of indonesia presidential election from twitter using sentiment analysis. *Journal of Big data*, Springer, v. 5, n. 1, p. 1–10, 2018.

EXAME. *Muhammadu Buhari é reeleito na Nigéria após eleição contestada*. 2021. Disponível em: <https://exame.com/mundo/nigeria-reeleito-presidente-promete-combater-a-inseguranca-e-a-corrupcao/>.

G1. *Eleições nos EUA: em quais estados as pesquisas acertaram e erraram*. 2020. Disponível em: <https://g1.globo.com/mundo/eleicoes-nos-eua/2020/noticia/2020/11/04/eleicoes-nos-eua-em-quais-estados-as-pesquisas-acertaram-e-erraram.ghtml>.

MENEZES, P. *Pesquisas no fim de agosto "erraram" últimas 4 eleições presidenciais*. 2018. Disponível em: <https://www.infomoney.com.br/colunistas/pedro-menezes/pesquisas-no-fim-de-agosto-erraram-ultimas-4-eleicoes-presidenciais/>.

MICROSOFT. *Power BI*. 2021. Disponível em: <https://powerbi.microsoft.com>.

MONEYTIMES. *Quanto custam as pesquisas eleitorais? Veja as mais caras*. 2018. Disponível em: <https://www.moneytimes.com.br/quanto-custam-as-pesquisas-eleitorais-veja-as-mais-caras/>.

NATANAEL. *Tweets das Eleições 2018 no Brasil*. 2018. Disponível em: <https://www.kaggle.com/natanaelsilva/tweets-eleicao2018>.

OYEBODE, O.; ORJI, R. Social media and sentiment analysis: The nigeria presidential election 2019. In: IEEE. *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. [S.l.], 2019. p. 0140–0146.

SALUNKHE, P.; DESHMUKH, S. Twitter based election prediction and analysis. *International Research Journal of Engineering and Technology (IRJET)*, v. 4, n. 10, p. 539–544, 2017.

TAKEBLIP. *Tudo sobre NLP: o que é processamento de linguagem natural e seus desafios na Inteligência Artificial*. 2021. Disponível em: <https://www.take.net/blog/tecnologia/nlp-processamento-linguagem-natural/>.

TSE. *Quando, afinal, há segundo turno em uma eleição?* 2021. Disponível em: <https://www.tse.jus.br/o-tse/escola-judiciaria-eleitoral/publicacoes/revistas-da-eje/artigos/revista-eletronica-eje-n.-6-ano-3/quando-afinal-ha-segundo-turno-em-uma-eleicao>.