UNIVERSIDADE FEDERAL FLUMINENSE

WILLIAM WANDERLEY DA SILVA

UM MODELO DE PREVISÃO DE RESULTADO DE ELEIÇÕES BASEADO EM COMENTÁRIOS, LIKES E DISLIKES EM MÍDIAS SOCIAIS DE CONTEÚDO EDITORIAL.

NITERÓI

2017

UNIVERSIDADE FEDERAL FLUMINENSE

WILLIAM WANDERLEY DA SILVA

UM MODELO DE PREVISÃO DE RESULTADO DE ELEIÇÕES BASEADO EM COMENTÁRIOS, LIKES E DISLIKES EM MÍDIAS SOCIAIS DE CONTEÚDO EDITORIAL.

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação.Área de concentração: ENGENHARIA DE SISTEMAS E INFORMAÇÃO

Orientadora:
Prof.ª Dr.ª ANA CRISTINA BICHARRA GARCIA

NITERÓI

2017

WILLIAM WANDERLEY DA SILVA

UM MODELO DE PREVISÃO DE RESULTADO DE ELEIÇÕES BASEADO EM COMENTÁRIOS, LIKES E DISLIKES EM MÍDIAS SOCIAIS DE CONTEÚDO EDITORIAL.

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: ENGENHARIA DE SISTEMAS E INFORMAÇÃO

Aprovada em Abril de 2017.

BANCA EXAMINADORA

Prof.ª Dr.ª ANA CRISTINA BICHARRA GARCIA
Orientadora, UFF

Prof. Luis Marti Orosa, UFF

Prof. Luis Miguel Parreira e Correia, Universidade de Lisboa

Niterói
2017

Ao meu pai que esteve comigo todo esse tempo (in memoriam).

Agradecimentos

A Deus, por ter iluminado meu caminho e me conduzido de volta à universidade após 15 anos de formado. Agradeço, também, à minha família, por todo o apoio durante este período de dedicação. À minha esposa Ana Paula e a meus filhos Manuela e Guilherme, por terem entendido os meus vários momentos de ausência; mas, que ainda assim, continuaram sempre me incentivando, pois somos um verdadeiro time. Vocês são a minha força. À minha orientadora, Ana Cristina Bicharra Garcia, por acreditar em mim, por sua sinceridade, incentivo, disponibilidade e, principalmente, pelos valiosos puxões de orelha quando necessário. Aos professores Luis Correia e Luis Marti pelas suas excelentes críticas e observações. Aos amigos Marcus Paulo Ferreia, Daniel Cinalli, Igor Natal e Felipe Montella, que me ajudaram muito durante esta jornada final do curso de mestrado. À todos os professores que contribuíram para a minha formação, em especial, Aline Paes e Carlos Ribeiro, por incentivarem a minha inscrição no processo de seleção. Agradeço também a Georgia Santoro, Adriana Franco e ao meu amigo Fleury pelo grande apoio na reta final deste trabalho.

Resumo

O processo de eleição é um dos mais importantes eventos para o futuro da sociedade, especialmente quando se trata de posições de prefeito, governador e presidente. O destino de uma nação pode ser drasticamente modificado pelo resultado de uma eleição. Efeitos na política e na economia são sentidos mesmo durante a campanha eleitoral. As pesquisas eleitorais são ferramentas poderosas que nos ajudam a minimizar as surpresas do resultado da eleição e a entender como está a percepção do candidato por meio do ponto de vista do eleitor. Fazer previsões é parte da natureza humana para antecipar acontecimentos e se sentir mais preparado para cenários futuros. A cultura de previsões está enraizada nas mais diversas áreas, tais como: mercado financeiro ou esportes, com a finalidade de preparar melhor as equipes para os adversários e, também, estar presente frequentemente nos ramos de apostas. Métodos tradicionais de previsão de resultado de eleição envolvem entrevistar uma amostra da população questionando sobre suas intenções de voto em relação aos candidatos que estão disputando determinada eleição. Estes métodos são caros, consomem tempo e estão sujeitos ao viés do entrevistador, que pode influenciar na opinião do eleitor apenas por sua linguagem corporal ao falar sobre os candidatos. O uso de mídias sociais para inferir a opinião dos eleitores tem se mostrado promissor e se utiliza do fato de que as pessoas estão cada vez mais expondo suas opiniões nas redes sociais, influenciando amigos e defendendo suas convicções de uma forma rápida e eficiente, por meio do uso de seus *smartphones*, que facilitam a exposição de suas opiniões no âmbito político e em redes sociais. Analisando este cenário, onde as opiniões dos eleitores estão de forma bastante aberta nas redes sociais, esta dissertação propõe um modelo de previsão de resultados de eleições baseado no método de regressão linear, no qual consideramos a polaridade dos comentários sobre candidatos realizados em conteúdo editorial em notícias políticas, bem como a aprovação e desaprovação deste comentário em forma de "*likes*" e "*dislikes*", respectivamente. Este modelo foi inspirado em estudos que mostram o poder de previsão utilizando *Twitter* e *Facebook* e tem como objetivo ser uma alternativa mais barata e ágil para os métodos tradicionais de pesquisa de opinião. Para a construção e teste do modelo proposto, coletamos comentários dos principais jornais *online* do Brasil, realizados, aproximadamente, há 4 meses antes das eleições municipais de 2016 em relação ao processo de eleição de 14 municípios, com um total de 70 candidatos.

Palavras-chave: Previsão de eleições. Jornais *online*. Análise de sentimentos. Mídias sociais. Inteligência coletiva.

Abstract

The election process is one of the most important events for a country, especially for the executive position such as mayor, governor, and president. The destiny of a nation may be drastically changed by the election results. Effects on economy and politics are felt even during the electoral campaign. Prediction pools are hired to post predictions so as to minimize surprises. Making predictions is part of human nature to anticipate events and thus get better prepared for possible future scenarios. The culture of prediction is rooted in the most diverse areas such as financial market, sports and sports betting. The forecasting process is an indispensable tool so as to develop and refine our decisions more assertively to obtain better results. Traditional methods of forecasting election results involve interviewing a sample of the population questioning their vote intentions. These methods are expensive, time consuming and subject to interviewer bias which can influence voter opinion only by their body language when talking about candidates.The use of social media to infer voters' opinions has been promising and people are increasingly exposing their opinions through social networks, influencing friends and defending their political beliefs by simply using smartphones. Taking into account this scenario where voters' opinions are widely available in social networks, this study proposes a model for forecasting election results based on linear regression method where we use the polarity of the comments about candidates in editorial content in political news, as well as the approval and disapproval of this comment in the form of likes and dislikes, respectively. This model was inspired by studies that show the power of forecasting using Twitter and Facebook and aims to be a cheaper and agile alternative for traditional forecast methods. For the construction of the proposed model and test, we collected comments from the main brazilian online newspaper conducted, approximately, 4 months before the municipal elections of 2016 to 14 municipalities and a total of 70 candidates.

Keywords: Election forecast. Online newspaper. Sentiment analysis. Social network. Collective intelligence.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

TSE	:	Tribunal Superior Eleitoral
AS	:	Análise de Sentimento
API	:	*Application Programming Interface* (Interface de Programação da Aplicação)
MAE	:	*Mean Absolute Error* (Erro Médio Absoluto)

Sumário

Capítulo 1

Introdução

O processo eleitoral é um dos acontecimentos mais importantes para a população de um país, pois a escolha de um determinado candidato pode mudar drasticamente o destino de uma nação ou até mesmo ter impacto no âmbito mundial. Durante o período eleitoral, a pesquisa de intenção de votos é um dos indicadores mais importantes, pois possibilita prevermos a preferência do eleitorado para determinados candidatos e entendermos o impacto da campanha eleitoral para cada perfil de eleitor. Este resultado também pode influenciar os eleitores indecisos [11].

Fazer previsões do futuro faz parte da natureza humana para se antecipar aos acontecimentos e se preparar melhor para os possíveis efeitos dos resultados. A cultura de previsão está enraizada nas mais diversas áreas, tais como: mercado financeiro, esportes, vendas e casas de apostas. O processo de previsão é uma ferramenta indispensável, para que possamos desenvolver e aprimorar as nossas decisões de forma mais assertiva, a fim de obter melhores resultados.

Embora a sociedade já aceite as previsões dos métodos tradicionais, estes têm se mostrado, por muitas vezes, frágil e incorreto. Temos como exemplo o caso da eleição presidencial americana do ano de 2016, em que a vitória da candidata Hillary Clinton era dada como certa pela grande maioria dos métodos tradicionais e, também, por empresas inovadoras, como a FiveThirtyEight[1], do famoso Nate Silver. Contrariando as previsões, o candidato Donald Trump foi eleito presidente dos Estado Unidos da América de uma maneira impressionante, surpreendendo o mundo.

Nesta sociedade, onde a informação está cada vez mais disponível, o tempo utilizado para gerar uma previsão do resultado da eleição precisa ser otimizado; pois a opinião dos

[1] https://fivethirtyeight.com/

eleitores pode sofrer mudanças que são refletidas nos resultados das pesquisas eleitorais. Fatos novos e relevantes podem ser considerados e, quanto mais rápida esta variação for detectada, maior serão as chances de um candidato ou partido se posicionar e atuar corrigindo os rumos da disputa política.

Atualmente, estamos vivendo em uma era onde as pessoas estão cada vez mais participativas. O acesso a *smartphones* facilita a exposição de opiniões no âmbito político em mídias sociais. Como exemplo, temos o *Facebook* e *Twitter*, onde os usuários compartilham suas opiniões, publicando milhões de *tweets* todos os dias. É possível, até mesmo, coletar opiniões das pessoas indiretamente, utilizando computação pervasiva, como previsto por Weiser [43] - onde a interação homem-máquina acontece de forma transparente e as preferências políticas podem ser mapeadas sem serem diretamente questionadas.

Nos métodos tradicionais de pesquisa, a opinião dos eleitores é coletada através de questionários criados de forma a minimizar influências do entrevistador. Com o presente estudo, apresentamos uma alternativa com um custo reduzido em relação à pesquisa tradicional, retirando o viés do entrevistador; pois a opinião do eleitor é coletada indiretamente, por meio de extração dos comentários realizados em notícias políticas de jornais *online*, onde a opinião está sendo criada sem o compromisso do eleitor se considerar participante de um processo de pesquisa eleitoral.

Neste estudo, é apresentado um modelo (baseado nas opiniões positivas, negativas ou neutras em comentários de jornais, bem como no apoio ou rejeição de demais usuários em forma de "likes"e "dislikes") que aplicamos ao primeiro turno do processo de eleição municipal no Brasil para o ano de 2016. Os resultados encontrados mostraram que existe uma relação importante destas variáveis independentes e o resultado oficial da eleição, no qual conseguimos prever os candidatos que estariam elegíveis no segundo turno.

Esta dissertação está organizada da seguinte forma:

- Capítulo 1: Introdução - Neste capítulo, é apresentado o trabalho, descrevendo os motivos para realizá-lo, destacando, assim, a importância do experimento.
- Capítulo 2: Fundamentação teórica. Este capítulo traz o modelo eleitoral brasileiro, pesquisas eleitorais, métodos tradicionais de previsão de eleição e os conceitos sobre a técnica de análise de sentimentos em textos não estruturados.
- Capítulo 3: Trabalhos relacionados à previsão de eleições, utilizando redes sociais. Este capítulo descreve e discute importantes estudos realizados sobre previsão de eleição, usando *Twitter* e *Facebook*.

- Capítulo 4: Estudo piloto. Neste capítulo, é apresentado o projeto que foi utilizado para a criação do modelo proposto.
- Capítulo 5: Modelo. Neste capítulo, é mostrado o modelo proposto para previsão da eleição.
- Capítulo 6: Experimento. Neste capítulo, relata-se o experimento realizado, utilizando o modelo proposto.
- Capítulo 7: Discussões e resultados. Neste capítulo, são exibidos os resultados computacionais encontrados e suas aplicações.
- Capítulo 8: Conclusões e Trabalhos futuros. Finalmente, neste último capítulo, as conclusões desta dissertação são apresentadas, ressaltando suas principais contribuições e trabalhos futuros relacionados.

1.1 Motivação

No mundo de hoje, devido à facilidade de acesso à informação, as pessoas estão criando o hábito de não mais aguardar por notícias dispostas em jornais impressos, que relatam, por vezes, informações desatualizadas. As notícias estão disponíveis quase que instantaneamente através do *Facebook*, *Twitter* ou *Whatsapp*. Assim, podemos perceber claramente a mudança de hábito dentro da nossa própria casa onde as crianças desde cedo possuem acesso à informação como nunca antes, por meio de dispositivos, como *smartphones*, *tablets* e *smart tvs*, em que escolhem o conteúdo que desejam sem mais se limitar à grade de programação definida por um canal de televisão. Com esta nova dinâmica, as pessoas estão deixando a posição passiva de simples consumidores de informação para uma posição de protagonista, no qual, a cada momento, é incentivado a expressar a sua opinião em relação às notícias publicadas e compartilhadas. As opiniões das pessoas expressas de forma aberta têm se tornando um ativo extremamente importante, que pode, ao mesmo tempo, criar celebridades ou destruir a credibilidade de empresas e produtos. Estas opiniões influenciam os indecisos e podem determinar o resultado de uma eleição.

Neste mesmo tempo, o processo eleitoral brasileiro está passando por um importante período de transição. Está saindo de uma era na qual candidatos são apoiados por doações de grandes grupos empresariais para um cenário onde este tipo de apoio está cada vez mais sendo questionado, devido a sua origem e objetivo, apesar de seguirem a constituição brasileira. Este fato tende a transformar o modo como são realizadas as campanhas

eleitorais e, além disso, já estamos vivendo uma tendência crescente de utilização de redes sociais por parte dos candidatos que já possuem contas com bastante atividades no *Twitter*, *Facebook*, *Youtube* e *Instagram*, onde interagem diretamente com os eleitores.

Com todos estes fatos, temos uma vasta quantidade de informações disponíveis e divulgadas em redes sociais e jornais digitais que refletem a opinião política dos usuários, tais opiniões podem ser uma fonte de informação para gerar previsões de eleições de forma rápida, barata e eficiente em comparação com as previsões realizadas pelos institutos tradicionais de pesquisa de opinião, que são geralmente caras, trabalhosas e demoradas.

1.2 Problema

Métodos tradicionais de previsão de eleição envolvem entrevistar uma amostra da população por meio de questionários predefinidos, para obter a intenção de voto em relação aos candidatos que estão disputando determinada eleição. Estes métodos são caros, consomem tempo e estão sujeitos ao viés do entrevistador, que pode influenciar na opinião do eleitor apenas por sua linguagem corporal ao falar sobre os candidatos.

Embora o *Twitter* seja a forma mais simples de obtermos conteúdo para análise de opinião dos usuários, devido à disponibilidade de suas APIs - e esta é uma das principais razões para a escolha desta plataforma como fonte de dados -, a eficiência de suas previsões são questionáveis, principalmente pela reprodutibilidade, representatividade e pelo viés das amostras; pois estas são geradas a partir de palavras-chave escolhidas previamente por quem está analisando. Não é possível coletar novamente os dados analisados utilizando novas palavras-chave e observar o mesmo período eleitoral.

Este cenário nos motiva a buscar novas fontes de informação para complementar os métodos de previsões, a fim de poder reproduzir o estudo, variar as palavras-chave utilizadas e tornar o processo mais rápido e com menor custo que os métodos tradicionais.

1.3 Hipóteses

Este estudo se propõe a analisar as seguintes hipóteses:

- É possível, em eleição brasileira, prever-se com precisão semelhante aos previsores convencionais (por *surveys*) quem são os primeiros dois candidatos e em que ordem usando-se a polaridade de comentários associados, *likes* e *dislikes* em notícias

politicas de jornais online no primeiro turno das eleições.

- Com o mesmo modelo de previsão, conseguimos prever o candidato eleito no segundo turno.

1.4 Metodologia de pesquisa

A pesquisa apresentada nesta dissertação é de natureza quantitativa. Para conduzir esta pesquisa, os seguintes passos foram executados:

- Revisão bibliográfica sobre modelos de previsão de resultados de eleição, utilizando a opinião de usuários coletada a partir de redes sociais;
- Desenvolvimento de um modelo de previsão de resultado de eleições, utilizando como fonte de informação os comentários de usuários sobre notícias de cunho político de jornais *online*;
- Realização de experimentos para analisar o modelo apresentado;
- Apresentação de resultados e análise quantitativa dos mesmos com a intenção de verificar a eficiência do modelo proposto desenvolvido.

Capítulo 2

Fundamentação teórica

2.1 Sistema Eleitoral Brasileiro

O processo eleitoral é a escolha coletiva que envolve a agregação de preferências individuais em uma única opção por meio de voto, onde o eleitor materializa a representação de sua escolha individual dentre um conjunto de opções de candidatos disponíveis. A votação é uma característica indispensável de uma democracia e impacta, de forma direta ou indireta, em todos aqueles que vivem em países democráticos. Para a maioria das pessoas, a experiência da votação é a única significativa de participação política [30].

No Brasil, o regime político está fundamentado na democracia (onde o povo determina quem serão os seus governantes) e no sistema presidencialista (composto por três poderes: Executivo, exercido pelo Presidente; Legislativo, exercido pelo parlamento; e Judiciário, que garante o cumprimento da Constituição Federal e aplica as leis, julgando determinada situação e as pessoas nela envolvidas) [1].

A constituição brasileira rege a obrigatoriedade do voto eleitoral para todos os cidadãos, exceto para analfabetos, menores de 18 anos ou para idosos com mais de 70 anos. Eleitores entre 16 e 18 anos possuem voto facultativo, não sendo obrigados a votar. Desta forma, o eleitor precisa expressar a sua opinião no dia da eleição ou justificar o motivo de sua ausência. O sistema eleitoral é baseado no voto direto e secreto, ou seja, o eleitor vota diretamente no candidato que representa sua ideologia, de maneira sigilosa; pois seu voto não pode ser divulgado a terceiros.

As principais funções que elegemos são: Presidente da República, Governador do estado e Prefeito da cidade ou município. Para cada um destes processos, a eleição pode

[1] http://www.brasil.gov.br/governo/2010/09/processo-eleitoral

ser disputada em 2 turnos, desde que nenhum candidato possua mais do que 50% dos votos. Caso contrário, o candidato é definido já no primeiro turno das eleições, que acontece sempre no primeiro domingo do mês de outubro. O segundo turno, quando houver, geralmente é realizado no último domingo do mesmo mês, apenas nas eleições para Presidente, governador e prefeito, em municípios com mais de 200 mil eleitores. Além disso, deve haver mais de dois candidatos no 1º turno de votação e nenhum deles ter conquistado a maioria absoluta dos votos válidos (50% mais um).

Para registrar os votos dos eleitores, o processo eleitoral brasileiro utiliza a urna eletrônica (introduzida no Brasil no ano de 1996), cujas vantagens são os vários mecanismos de segurança que impedem adulterações e garantem o sigilo do voto. A impossibilidade de identificação do eleitor, em conjunto com a inexistência de ligação da urna a qualquer dispositivo de rede, tornam a urna eletrônica indispensável para evitar violações do processo de votação.

Inicialmente, estas urnas foram utilizadas em todo o Estado do Rio de Janeiro, nas capitais dos demais estados e nos municípios com mais de 200 mil eleitores, totalizando 57 cidades no país. Um terço dos quase 100 milhões de eleitores votou por meio das urnas eletrônicas nas eleições municipais. De forma gradual, nas eleições do ano de 1998, dois terços do eleitorado votaram eletronicamente e, a partir do ano 2000, a urna eletrônica foi adotada por todo o país [2].

Com o objetivo de aumentar ainda mais a segurança do processo eleitoral, a partir do ano de 2008, passou a ser adotado em algumas localidades brasileiras o sistema biométrico, que identifica o eleitor através de suas impressões digitais. Desde então, a Justiça Eleitoral vem de forma gradual realizando o recadastramento biométrico de todos os eleitores brasileiros.

A apuração e divulgação dos resultados das eleições brasileiras acontecem no mesmo dia em que é realizada a votação. Por esta agilidade e segurança, o Brasil se tornou referência mundial em eleições.

Segundo estatísticas do TSE (Tribunal Superior Eleitoral) do ano de 2016, o Brasil conta com cerca de 276 milhões de eleitores distribuídos de acordo com a figura 2.1.

[2] http://www.tse.jus.br/hotsites/catalogopublicacoes/pdf/urna_eletronica/livretournaprograma-educativo_web.pdf

Dezembro - 2016		
Abrangência	Quantidade	%
CENTRO-OESTE	10.562.837	7,221
EXTERIOR	424.800	0,290
NORDESTE	39.326.717	26,885
NORTE	11.367.860	7,772
SUDESTE	63.396.418	43,341
SUL	21.196.388	14,491
	146.275.020	
	146.275.020	

Figura 2.1: Censo do ano de 2016.
Fonte: `http://www.tse.jus.br`

2.2 As pesquisas eleitorais

As pesquisas de opinião pública estão cada vez mais presente no cotidiano da população, abordando assuntos mais variados para inferir a preferência ou julgamento das pessoas a respeito do tema abordado.

Pesquisas eleitorais são levantamentos amostrais conduzidos por institutos de pesquisa que procuram investigar a opinião pública, captando opiniões ou preferências individuas a respeito dos candidatos, utilizando questionários, formulários e entrevistas, em uma tentativa de previsão sobre qual candidato será eleito. Estes questionários são preparados de forma a tentar minimizar qualquer influência na opinião, mostrando, por exemplo, em seu formulário, a lista de candidatos organizados em formato circular, não priorizando, assim, nenhum dos candidatos, como ilustrado na figura 2.2

Estas pesquisas geram grande impacto nas decisões de voto do eleitor, através até do fenômeno *Bandwagon*[1], e seus resultados influenciam diretamente nas estratégias de campanhas dos partidos. O candidato busca intensificar em sua campanha de *marketing* ou minimizar tais fatos, caso esteja à frente ou abaixo das pesquisas eleitorais, respectivamente [14].

Embora muito comum, o levantamento de opinião pública é um assunto ainda muito polêmico e importante, que pode ser criticamente estudado sob vários aspectos. Por exemplo, em relação a ciências sociais, Pierre Bourdieu [4], em seu texto "A Opinião Pública não Existe", faz críticas às sondagens de opinião, onde aborda três principais

[1] *Bandwagon* é o termo que indica a tendência do eleitor em dar seu voto ao candidato que aparece como favorito nas pesquisas. Por outro lado, dá-se o nome de *Underdog* ao efeito oposto, onde exista a tendência em se apoiar o candidato mais fraco.

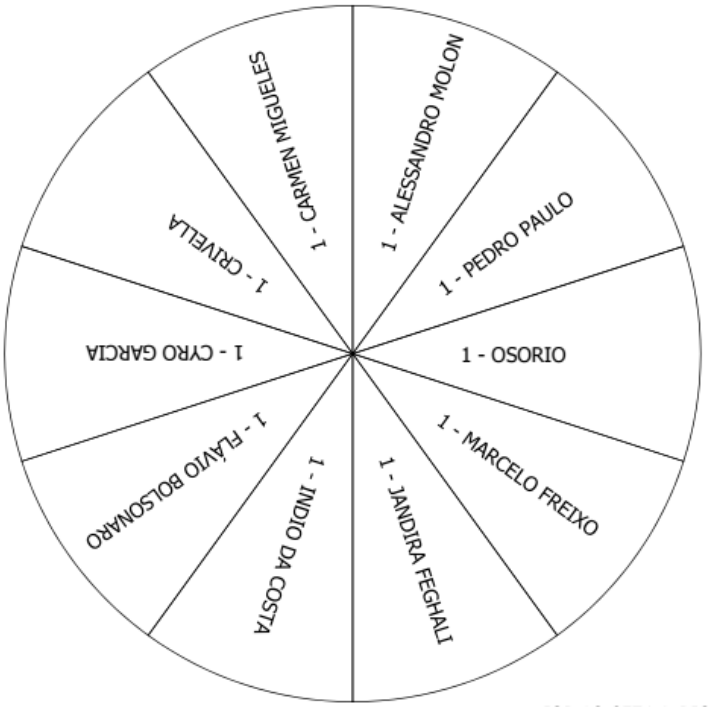

Figura 2.2: Fonte: Formulário do IBOPE nas eleições municipais de 2016 na cidade do Rio de Janeiro.

questões[2].

Devido a esta grande influência sobre o eleitorado, as entidades e empresas que quiserem realizar pesquisas de opinião pública, por lei, devem registrar cada pesquisa até cinco dias antes da divulgação de cada resultado e os dados registrados ficam à disposição de qualquer interessado pelo prazo de 30 dias[3].

2.3 Métodos tradicionais de previsão de eleições

No Brasil, os principais institutos de pesquisa são o IBOPE (criado em 1942) e o Datafolha (criado em 1983). Este institutos têm em comum a utilização do método de amostragem por quotas.

Para realizar uma pesquisa nacional de intenção de votos utilizando o método de amostragem por cotas, devemos seguir os seguintes conceitos:

[2]Pierre Bourdieu discute três postulados sobre pesquisas de opinião: a) Qualquer pesquisa de opinião supõe que todo mundo pode ter uma opinião; ou, colocando de outra maneira, que a produção de uma opinião está ao alcance de todos; b) Supõe-se que todas as opiniões têm valor; c) Pelo simples fato de se colocar a mesma questão a todo mundo, está implícita a hipótese de que há um consenso sobre os problemas, isto é, que há um acordo sobre as questões que merecem ser colocadas

[3]http://www.tre-sp.jus.br/imprensa/noticias-tre-sp/2016/Janeiro/pesquisas-de-opiniao-sobre-as-eleicoes-2016-devem-ser-registradas-em-site-da-justica-eleitoral

1. A amostra é uma fatia do Universo de todos os eleitores, de acordo com o número de eleitores de cada região;

2. A seleção da amostra de pessoas que serão entrevistadas deverão corresponder proporcionalmente à composição da população de eleitores, de acordo com algumas características, tais como: sexo, idade e classe social de cada região;

3. O próximo passo, após a definição da amostra, é preparar o questionário que será apresentado aos entrevistados. O pré-requisito para participar da pesquisa é residir na zona previamente estabelecida pelo plano da pesquisa e fazer parte do grupo que possui as características definidas;

4. Ao utilizarmos a amostragem por cota, existe um erro amostral conhecido e calculado em função do tamanho da amostra e dos resultados obtidos na pesquisa. Quando, por exemplo, consideramos que a estimativa de erro de uma determinada pesquisa é de dois pontos percentuais, devemos entender que um resultado de 30% pode apresentar uma variação de dois pontos, para mais ou para menos, e deve ser lido como um intervalo de 28% a 32%. Quanto maior a homogeneidade da população pesquisada, menor será o erro amostral e vice-versa.

2.4 Métodos de previsão que utilizam redes sociais

Opiniões nas redes sociais ajudam na tarefa de compreender e explicar diversos fenômenos sociais complexos, bem como também prevê-los. Como, atualmente, os atuais avanços tecnológicos permitem o armazenamento e recuperação de enorme quantidade de dados eficientemente, abriu-se uma enorme oportunidade de desenvolvermos metodologias para extração de informações e criação de conhecimento a partir de fontes de dados distintas. Estudos [33, 3, 32, 44, 19] exemplificam métodos para mapear a preferência do usuários e criam modelos de previsão de eleição [34] descritos no capítulo 3.

Dentre as principais vantagens em relação aos métodos tradicionais, podemos listar as seguintes:

- Menor custo;
- Análise das opiniões coletadas e analisadas em tempo real;
- Opinião do usuário coletada sem a interferência do entrevistador.

Em relação às principais desvantagens, podemos listar as seguintes:

- Os usuários não são escolhidos randomicamente assim como é feito com os métodos tradicionais de pesquisa de opinião. Os usuários costumam seguir a fonte de informação que sejam coerentes com seu modo de ver o mundo [36] e interagem apenas com a informação que lhes interessa;
- As opiniões do usuários podem ser simuladas por robôs mal intencionados que podem ser de difícil detecção, pois estão cada vez mais sofisticados;
- Difícil caracterização da amostra de acordo com o perfil da população.

Para definir a polaridade em termos de conteúdo positivo, negativo ou neutro, utilizamos as técnicas de análise de sentimento (AS) também conhecido como mineração de opinião [20], que diz respeito à avaliação automática de opiniões, avaliações, sentimentos, entre outros, expressos em forma de textos e comentários a respeito de opiniões políticas, religiosas ou mesmo sobre marcas, produtos e serviços, a fim de criar conhecimento estruturado que possa ser utilizado por um sistema de apoio ou tomador de decisão.

A análise de sentimento tem atraído atenção da comunidade e pesquisadores, tais como aprendizado de máquina e processamento de linguagem natural e a figura 2.3 ilustra o crescimento da procura no serviço Google pel termo "Sentiment Analysis" ao longo do tempo.

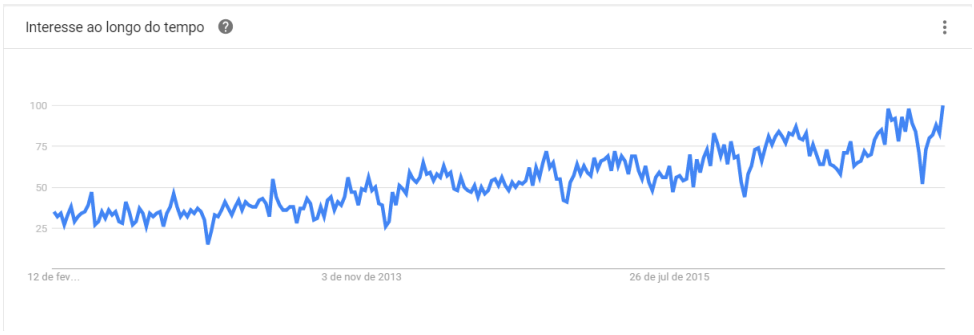

Figura 2.3: Pesquisa do termo "Sentiment Analysis" no Google Trends.

Os principais objetivos da mineração de opinião são identificar documentos que contém fatos e opiniões e classificá-los quanto à sua polaridade entre positivo, negativo ou neutro [39].

A classificação de opinião pode ser realizada em termos de documento, onde visa determinar se este como um todo contém uma opinião positiva ou negativa quanto ao

aspecto geral da entidade em questão [21]. Quando a classificação for realizada em termos de sentença, parte do suposto que que cada sentença pode conter apenas uma opinião e que as entidades e aspectos mencionados nas sentenças são conhecidos [21]. Nesse caso, o objetivo é identificar se uma sentença contém um fato ou uma opinião positiva ou negativa.

Estratégias para a classificação de opiniões em termos de documentos e sentenças assumem, geralmente, que as opiniões expressas referem-se a apenas uma entidade e que sentenças contêm apenas uma opinião. No entanto, é possível que os aspectos associados às entidades avaliadas não sejam julgados da mesma forma, quanto às suas polaridades.

A polaridade representa o grau de positividade e negatividade de um texto. Alguns métodos tratam a polaridade como um resultado discreto binário (positivo ou negativo) ou ternário (positivo, negativo ou neutro). Por exemplo, a frase "Como você é gentil" é positiva e a frase "O meu time foi derrotado" é negativa, já a frase "Meu carro é alugado" não possui polaridade e, normalmente, é classificada como neutra.

A força do sentimento representa a intensidade de um sentimento ou da polaridade. Normalmente, é um ponto flutuante entre (-1 e 1). Há trabalhos que, por exemplo, medem a força de sentimentos nos títulos das notícias [29], capaz de separar eficientemente para o usuário notícias boas de notícias ruins.

A classificação de documentos pode ser realizada através de técnicas supervisionadas e não supervisionadas.

1. Técnicas supervisionadas

   A abordagem supervisionada emprega o termo supervisionado pelo fato de exigir uma etapa de treinamento de um modelo com amostras previamente classificadas. O procedimento para realizar a aprendizagem de máquina compreende quatro etapas principais:

   (a) Obtenção de dados rotulados que serão utilizados para treino e para testes;

   (b) Definição das *features* ou características que permitam a distinção entre os dados;

   (c) Treinamento de um modelo computacional com um algoritmo de aprendizagem;

   (d) Aplicação do modelo.

2. Técnicas não supervisionadas

As técnicas não supervisionadas, diferentemente das supervisionadas, não necessitam de sentenças previamente rotuladas e treinos para a criação de um modelo. Esta é uma das suas principais vantagens, uma vez que desta forma não mantém aplicação restrita ao contexto para o qual foram treinados. As técnicas que mais se destacam nesta abordagem são as léxicas, baseadas em um dicionário léxico de sentimento, uma espécie de dicionário de palavras, que ao invés de possuir como conteúdo o significado de cada palavra, possui em seu lugar um significado quantitativo onde pode ser um número entre -1 a 1, e -1 indicando um valor sentimental mais negativo e 1 o valor mais positivo. Uma outra forma é as palavras possuírem um valor qualitativo como, por exemplo, positivo, negativo, feliz ou triste. Estas abordagens léxicas partem do princípio que palavras individuais possuem o que é chamado de polaridade prévia, ou seja, uma orientação semântica independente de contexto e que pode ser expressada com um valor numérico ou classe [37].

Capítulo 3

Trabalhos relacionados: Previsão de eleições e mineração de opiniões utilizando redes sociais.

O *slogan* original do *Twitter* - *What are you doing ?* - fez um excelente trabalho; pois incentivou os usuários à compartilhar seus pensamentos e atividades com seus amigos e seguidores. Hoje, o *Twitter* é considerado uma das mais importantes ferramentas de disseminação de opiniões. Apesar disso, ainda permanece o fato de que a grande maioria dos usuários da *Internet*, sem mencionar as pessoas em geral, não estão utilizando *Twitter* [15].

Devido a este enorme volume de informações, a busca por métodos alternativos de previsão de eleição tem sido motivo de estudos que analisam a opinião de eleitores expressadas através das mídias sociais. Estes estão baseados no uso do *Twitter*, devido à disponibilidade de seus dados, ao contrário do *Facebook*, que limita o acesso somente à sua rede de contatos e às páginas públicas com informações agrupadas, tais como a quantidade de seguidores. Contudo, Sapiras et al. [33] descrevem em seu estudo a experiência de mineração de opiniões em relação ao aspecto dos comentários realizados em jornais *online* sobre notícias eleitorais, onde mostrou ser possível identificar, classificar a polaridade e sumarizar a opinião de leitores de um jornal sobre os aspectos saúde e educação, relacionados ao candidato à eleição municipal.

Em seu estudo sobre previsão de eleição utilizando redes sociais, Tumasjan et al. [41, 40] coletaram os *tweets* durante aproximadamente 5 semanas, utilizando como palavra-chave o nome dos candidatos e, em sua conclusão, verificou que a simples menção ao nome destes candidatos pode significar voto a favor e refletir diretamente no resultado da eleição. Estes pesquisadores alegaram que o método descrito não só era possível como

também bastante simples.

Este método envolve os seguintes passos :

1. Definir a eleição;
2. Selecionar palavras-chave que representam os candidatos ou partidos políticos participantes da eleição em questão;
3. Coletar *tweets* durante um período de tempo precedente à eleição que contenha as palavras-chave ou *hashtags*;
4. Contabilizar o número de *tweets* que mencione cada candidato ou partido;
5. Utilizar o número contabilizado no passo anterior para calcular o percentual de votos de cada candidato.

Apesar do resultado positivo encontrado, este pode ter sido afetado diretamente pela escolha das palavras-chave realizada pelos pesquisadores. Uma diferente combinação de palavras no momento da coleta poderia ter apresentado um resultado diferente. Estes fatos tornam o método proposto difícil de ser reproduzido para a mesma eleição com diferente combinação de palavras.

Neste capítulo, organizamos e apresentamos os estudos através das seguintes seções:

- Modelos utilizados para prever o resultado ou tendências das eleições;
- Detecção de técnicas de manipulação de opiniões;
- Tratamento dos consumidores passivos de informação;
- Coleta de dados.

3.1 Modelos utilizados para prever o resultado ou tendências das eleições

Miranda et al. [24] criaram um modelo de previsão para eleição municipal de 6 municípios do Brasil e compararam com o resultado oficial das instituições de pesquisa e votos tradicionais IBOPE, Datafolha e a simples menção do nome do candidato. Consideraram que se um usuário comentasse sobre um candidato da cidade X, ele votaria na cidade X.

Desenvolveram uma metodologia que caracterizou o usuário de acordo com os atributos demográficos, dentre eles, gênero, idade e classe social. Filtraram *tweets* de conteúdo editorial e *spammers*, utilizaram análise de sentimento para classificar os *tweets* e computou o percentual de cada candidato, levando em consideração que cada comentário positivo correspondia a um voto e cada comentário negativo era distribuindo percentualmente entre os demais candidatos. No melhor caso, acertou 50% dos candidatos ao segundo turno, sem mencionar a ordem entre o primeiro e o segundo, e 66% dos vencedores da disputa do segundo turno da eleição.

Hagar [17] utilizou uma abordagem de regressão linear para verificar a relação entre as variáveis independentes "Twitter Use", "Followees", "Followers", "Candidate Tweets", "Candidate Replies", "Favourites", "Retweets", "Mentions", "Voter Replies", "Sex", "Incumbency", "Age", para verificar a relação entre os votos recebidos pelos candidatos chamados *Incumbents* e *Challengers* nas eleições municipais do Canadá. Verificou que, quanto mais o candidato era mencionado no *Twitter* sua popularidade também crescia, apesar da amostra relativamente pequena que correspondia a 1 mês antes das eleições. Sua proposta mostrou a relação de cada uma dessas variáveis com os votos, mas não detalhou os candidatos e o modelo de previsão utilizando as variáveis indicadas.

Dwi e Hauff [13] debateram o fato de que na Indonésia, diferentemente dos países desenvolvidos, os métodos tradicionais de pesquisas eleitorais não possuem um bom desempenho. Executaram o experimento de análise de eleição variando entre estratificar os usuário em termos de gênero e geolocalização, mudaram o período de coleta, incluíram análise de sentimento, computaram como voto a quantidade de usuários ou quantidade de *tweets*, testaram palavras-chave adicionais para identificar o candidato e removeram *spammers*. Com todas estas combinações, acertaram o candidato vencedor na melhor combinação das estratégias em 74% dos casos.

Gayo-Avello [15] filtrou *tweets* de acordo com a geolocalização, especialmente para uma eleição de interesse mundial como a do Estados Unidos, pois pessoas do mundo todo que não necessariamente eleitores compartilham a sua própria opinião e este fato iria distorcer consideravelmente a previsão. Com isso, usuários originários do país em questão são incluídos no modelo de previsão e os demais são ignorados. Detalhou um estudo que superestimou a vitória de Obama em 2008 nas eleições presidenciais do Estados Unidos. Gayo-Avello [16] também contesta as previsões baseadas no *Twitter*, argumentando que os pesquisadores possuem a tendência de não publicarem resultados ruins, o que pode gerar uma falsa sensação de sucesso na comparação dos métodos com o resultado da pes-

quisa. Relata, ainda, que a maioria das pesquisas são realizadas pós-eleição e que o fato de alguém postar uma mensagem no *Twitter* significa intenção de voto. Com base nas premissas de Gayo-Avello, Almeida [1] desenvolveu uma nova metodologia para definir amostras estratificadas, utilizando atributos demográficos para previsão de eleição também utilizando análise de sentimento em *tweets* para contabilizar votos, criando amostras geradas de acordo com os dados demográficos inferidos para prever o resultado da eleição. Apesar de demonstrar sucesso em suas previsões, relatou que, para criar boas amostras estratificadas, necessitaria de um número bastante elevado de *tweets* para se igualar a real distribuição.

Bovet et al. [5], utilizando *tweets*, previram a vitória de Hillary Clinton sobre Donald Trump combinando processamento de linguagem natural e classificação de aprendizado de máquina para inferir a opinião do usuário a respeito dos candidatos da eleição presidencial dos Estados Unidos de 2016. A eleição foi vencida por Donald Trump, o percentual de votos foi maior para a candidata Hillary Clinton. Caso a análise levasse em consideração o processo de apuração distrita, talvez pudesse ter sido diferente e indicado o verdadeiro vencedor. O resultado encontrado esteve de acordo com o índice *New York Times National Polling Average* que agrega a informação e vários institutos de pesquisa tradicionais renomados.

Burnap et al. [6] utilizaram análise de sentimento em *tweets* não para prever os votos dos candidatos, mas sim dos partidos nas eleições gerais do Reino Unido em 2015. Apesar do resultado do partido vencedor ter ficado próximo da previsão, este não pode ser generalizado, porque foi pressuposto que os eleitores estavam igualmente distribuídos e isto pode ter sido a causa na distorção dos resultados; pois os partidos possuem grande influências regionais onde a maior dificuldade foi a definição da amostra.

Saleiro et al. [31] compararam a previsão baseada em *Twitter* com pesquisas tradicionais nas eleições portuguesas de 2016 utilizando um modelo de regressão baseado em funções agregadas em análise de sentimento em um *corpus* de 1500 *tweets* anotados manualmente por 3 estudantes de ciências políticas. Suas variáveis independentes eram computadas mensalmente para acompanhar as previsões realizadas pelos institutos tradicionais de pesquisas.

Não apenas o *Twitter* tem sido utilizado como base para previsões eleitorais, o *Facebook* também é uma fonte rica em opiniões de usuários, mas apresenta uma importante limitação para a coleta de dados, devido à restrição de acesso a conteúdos não autorizados. Previsões utilizando *Facebook* são geralmente baseadas em informações abertas,

Tabela 3.1: Estudos sobre previsão de eleição

Artigo	Publicação	País	Ano da Eleição	Tipo de eleição	Coleta	Fonte	Método	Resultado
[24]	2015	Brasil	2012	Municipal	2 meses	Twitter	Segmentação dos usuários. Contagem de tweets. Análise de sentimento	50% dos candidatos ao segundo turno. 66% no segundo turno
[17]	2015	Canadá	2014	Municipal	1 mês	Twitter	Contagem de tweets. Menções e hashtags.	Encontrou uma relação positiva entre a menção ao nome do candiato e o percentual de voto
[13]	2015	Indonésia	2014	Presidencial	3 meses	Twitter	Contagem de tweet. Análise de sentimento. Variação do tamanho da amostra.	Acertou o candidato vencedor 74% das vezes.
[15]	2011	USA	2008	Presidencial	6 meses	Twitter	Contagem de tweet Análise de sentimento	MAE 11,63 %
[41]	2010	Alemanha	2009	Federal	5 semanas	Twitter	Contagem de tweets Menções e hashtags	MAE 1,7 %
[5]	2016	USA	2016	Presidencial	4 meses	Twitter	Processamento de linguagem natural e classificação de tweets	Clinto 55,5% e Trump 44,5%
[28]	2015	USA	2012	Presidencial	2 meses	Facebook	Segmentou usuários seguidores para avaliar o engajamento político	N/A
[6]	2016	Reino Unido	2015	Presidencial	1 mês	Twitter	Contagem de tweet Análise de sentimento	MAE 2,3%
[31]	2016	Portugal	2011-2014	Resgate de Portugal	mais de 2 anos	Twitter	Modelo de regressão utilizando funções agregadas de análise de sentimento	MAE 0.63%
[12]	2015	Suécia	2014	Parlamentar	8 meses	Twitter	link mining	N/A
[3]	2015	Índia	2014	Parlamentar	5 meses	Facebook	Número de likes	86.6% de acurácia

que a plataforma oferece, tais como a quantidade de seguidores de um determinado candidato. Para comprovar a relação entre seguir um candidato e estar engajado com sua campanha, Pennington [28] realizou um experimento onde foram escolhidos 135 alunos, e estes foram divididos entre dois grupos durante o estudo. O grupo presidencial, seguindo Romney/Obama; ou o grupo de controle, que não seguia nenhum candidato. Ao longo deste período, foi mensurado o engajamento do participante de acordo com interesse, participação e discussão. O estudo mostrou que o engajamento do eleitor não foi afetado durante o estudo, o que indica que seguir um candidato no *Facebook* não necessariamente corresponde a estar engajado.

Dokoohaki et al. [12] propuseram uma abordagem de *link mining* para avaliar a interação de candidatos e os eleitores, criando um grafo para representar esta interação e comparando o percentual de votos com o grau dos vértices que representam a relação entre os candidatos e as pessoas que seguem os candidatos. Mostrou que existe uma importante correlação entre votos, mas não criou um modelo de previsão.

A tabela 3.1 resume estudos apresentados com o objetivo de prever eleição ou encontrar relacionamento entre votos e participação de usuários em redes sociais.

3.2 Detecção de técnicas de manipulação de opiniões

Devido ao fato de cada vez mais as pessoas compartilharem as suas opiniões e os estudos se mostrarem promissores para previsão de eleição, ao mesmo tempo grupos mal intencionados se aproveitam destes métodos para tentar manipular a opinião das pessoas publicando informações muitas vezes mal intencionadas, criando artifícios para burlar a análise através de robôs que simulam o comportamento de um eleitor e pode influenciar negativamente no resultado das previsões.

Além da utilização do *Twitter* para expressar suas opiniões sobre as campanhas políticas, a grande maioria das pessoas que possuem acesso à *Internet* utilizam também aplicações de busca para se informar. Estas ferramentas de buscas, tais como Google.com, Bing.com são consideradas confiáveis pelas pessoas e geralmente utilizam apenas a primeira página do resultado para se instruir sobre o tema pesquisado, acreditando que estão recebendo informações verdadeiras. Apesar da evolução das ferramentas de buscas, que visam sempre entregar o conteúdo mais relevante para o usuário em termos de importância e confiabilidade, *spammers* utilizam todo o tipo de truques chamados de *black hat* [1] na tentativa de burlar as regras e promover conteúdos para influenciar os eleitores que, por sua vez, acreditam que o conteúdo seja relevante; pois estar na primeira página é um grande indicador de reputação e popularidade. Neste sentido, Metaxas e Mustafaraj [27] detectaram e descreveram como um grupo político criou contas automáticas para realizar ações para desacreditar um candidato, utilizando *Twitter*. E estes *tweets* são rapidamente disponibilizados nas ferramentas de busca.

Castilho at al. [9] criaram um modelo de classificação para determinar a credibilidade de uma informação a partir de um conjunto de *tweets* e os classificou em crível ou não crível, baseado em características extraídas, tais como comportamento de "re-tweet", o conteúdo do *post* e a citação de fontes externas.

Mendoza at al. [23] mostraram que a propagação de *tweets* verdadeiros diferem de *tweets* que correspondem a rumores em situações de crise; pois este último tende a ser muito mais questionado pela comunidade. Seu resultado apresentou que é possível detectar rumores utilizando técnicas de agregação.

[1] Black Hat é um termo utilizado para se referir às pessoas ou técnicas que visam atingir um objetivo sem a autorização do órgão, empresa ou pessoa responsável. Esse objetivo pode ser a entrada em um sistema protegido, monetização por meios não autorizados ou o acesso às informações confidenciais.

3.3 Interações do usuário com o conteúdo.

Não podemos negar que a grande maioria dos usuários de redes sociais são consumidores passivos de informação. Como exemplo, mais de 75% dos *tweets* no *Twitter* são produzidos por menos de 20% dos usuários [10]. Podemos perceber, através dos vídeos dos vídeos mais populares do *YouTube* [2], que estes possuem menos do que 1% dos comentários em relação à sua audiência [2] e também muito menos do que a quantidade de "likes"e "dislikes"destes mesmos vídeos. Neste sentido, Venkataraman et al [42] argumentam que a mineração de opinião deveria considerar as participações passivas e ativas nas redes sociais e inferem a quantidade de usuários afetados por uma informação utilizando o percentual de carga dos servidores responsáveis pelo conteúdo, mas não inferem a preferência do usuário. Mustafaraj et al. [26] apontam em seu estudo que a real maioria costuma se manifestar apenas depois do evento concluído. Defendem que o tipo de conteúdo gerado por estes dois tipos de usuários são essencialmente diferentes, e que a minoria cria seus *tweets* utilizando mais *hashtags*, *links*, menções e "retweetando"uma taxa de duas vezes mais do que a maioria das pessoas.

3.4 Coleta de dados

Twitter é uma enorme rede de de compartilhamento de mensagens curtas e é conhecido no mundo acadêmico por disponibilizar seus dados de forma aberta. Para acessar os bilhões de *tweets* gerados, podemos utilizar a "Streaming API", que fornece uma amostra de cerca de 1% dos dados gerados que estiverem de acordo com uma determinada palavra-chave informada. Este serviço tem auxiliado como base para inúmeros estudos de comportamento utilizando apenas esta pequena amostragem da informação. Um estudo realizado por Morstatter [25] proporciona uma comparação em termos estatísticos dos dados fornecidos pela "Streaming API"em relação à versão paga chamada *Firehose*, onde todos os *tweets* são disponibilizados e não apenas uma amostragem. Em ambas as versões, os dados podem ser coletados por palavra-chave ou a respeito de um determinado usuário [3].

Para coletar informações do *Facebook*, podemos utilizar a *Facebook Graph API*. Entretanto, o acesso às informações se mostra bastante limitado e dificulta a realização de pesquisas sobre as mensagens compartilhadas pelos usuários. A opção é coletar da-

[2] https://www.youtube.com/feed/trending
[3] https://dev.twitter.com/rest/public

dos abertos, como quantidades de seguidores e mensagens de usuários classificados como "Friends", de quem realiza a pesquisa.

Capítulo 4

Estudo piloto

A partir dos estudos realizados sobre modelos de previsão de eleição na rede social *Twitter*, onde os *tweets* publicados pelos usuários possuem as características de tamanho do texto reduzido semelhantes aos comentários realizados por usuários em jornais *online*, concluímos que faria sentido realizar um estudo piloto que mostrasse a viabilidade e a eficiência de criar um modelo de previsão de eleição baseado nesses comentários de jornais *online*, com a vantagem de podermos separar os comentários que estejam contextualizados apenas em notícias de cunho político e com a possibilidade de reprodução do experimento em qualquer tempo; pois as notícias e seus respectivos comentários continuam disponíveis online e não precisamos definir antecipadamente palavras-chave que delimitem o conjunto de comentários capturados, ao contrário de análises de *tweets*. Este trabalho foi baseado em Sapiras et al. [33], mas com o intuito de criar um modelo de previsão de eleições. Para coletar os comentários dos jornais, utilizamos o *framework* descrito no apêndice C

Consideramos diferente municípios para as bases de treinamento e teste com o intuito de criar um modelo genérico que fosse independente das características dos municípios.

Neste capítulo, estão as seguintes etapas :

- Escolha dos jornais e notícias.
  Nesta seção, descrevemos o processo de escolha dos jornais.
- Escolha dos municípios e candidatos.
  Nesta seção, apresentamos o processo de escolha dos municípios e candidatos.
- Captura dos comentários, *likes* e *dislikes*.
  Nesta seção, mostramos as informações relevantes em um comentário.

- Escolha dos atributos. Nesta seção, definimos os atributos utilizados para criar o modelo de previsão.
- Filtro de análise de sentimento. Nesta seção, relatamos a escolha da base de palavras para definir a polaridade dos comentários.
- Geração do modelo. Nesta seção, finalmente, apresentamos o processo de geração do modelo gerado a partir da base de treinamento.

4.1 A escolha dos jornais e notícias

Neste estudo, utilizamos como fonte de informação os comentários realizados sobre notícias políticas de renomados jornais *online* do Brasil de conteúdo editorial, pois partimos do princípio que estes jornais são isentos de direcionamento partidário e são fontes de informação consideradas verdadeiras. Para realizar o estudo, selecionamos as notícias que estavam em um intervalo de tempo de até 4 meses antes do dia da eleição.

A escolha dos jornais como fonte de informação não foi apenas uma contrapartida ao *Twitter*, foi também uma maneira de coletar a opinião em um ambiente onde os eleitores compartilham a sua opinião em forma de “likes” e “dislikes”, sem necessariamente realizar um comentário. A informação de "dislike"não existe na plataforma do *Twitter*, assim como também não existe no *Facebook*.

Escolhemos 4 jornais *online* para diversificar a representatividade do público alvo onde as matérias são publicadas. Desta forma, minimizamos o viés que poderíamos ter, caso o direcionamento do conteúdo estivesse focando em apenas uma classe social ou região do país.

Os jornais escolhidos foram G1 [1], Folha de São Paulo [2], Gazeta do Povo [3] e Extra Online [3]. A quantidade de notícias está descrita na tabela 4.1.

O jornal G1, que não possui direcionamento partidário, representa o *site* de maior abrangência e audiência nacional e, por este motivo, foi escolhido para representar a maior parte dos municípios escolhidos.

[1] http://g1.globo.com/
[2] http://www.folha.uol.com.br/
[3] http://extra.globo.com/
[3] http://www.gazetadopovo.com.br/

Tabela 4.1: Quantidade de notícias por jornais.

Jornal Online	Quantidade de notícias
Extra	688
Folha de São Paulo	331
G1	2930
Gazeta do Povo	273

Tabela 4.2: Base de treinamento.

Município	Notícias	Comentários	População
Belo Horizonte	289	1789	2.513.451
Campos dos Goytacazes	168	611	487.186
Fortaleza	154	3480	2.609.716
Guarulhos	93	653	1.337.087
Osasco	228	2696	696.382
Manaus	68	270	2.938.092
Nova Iguaçú	145	1920	797.435
Porto Alegre	222	1476	1.481.019
Salvador	133	1048	2.938.092
Santos	104	5046	434.359

4.2 Escolha dos municípios e candidatos

Devido ao vasto território Brasileiro e os 5570 do municípios do Brasil, existe uma dificuldade de estes municípios possuírem jornais *online* que apenas os representem. Nos problemas que encontramos, os jornais ou não possuíam a funcionalidade de comentários ou a quantidade e comentários era irrelevante. Desta forma, criamos uma lista de 14 municípios relevantes em âmbito nacional, dentre os quais escolhemos aleatoriamente 10 para a geração do modelo e outros 4 separamos para a fase de experimento. Escolhemos 5 dos principais candidatos para cada prefeitura, com o objetivo de maximizar a chance de este candidato escolhido ter relevância para ser comentado.

Os municípios escolhidos para as bases de treinamento e teste, suas notícias relacionadas e a quantidade de comentários estão detalhados na tabela 4.2 e 4.3 respectivamente, onde a fonte de informação sobre população é o IBGE [1].

4.3 Captura dos comentários, *likes* e *dislikes*

Partindo da premissa observada que comentários anônimos não são permitidos em jornais *online*, e sendo esta premissa seguida por todos os principais jornais do Brasil, podemos

[1] http://www.cidades.ibge.gov.br

Tabela 4.3: Base de teste.

Município/Cidade	Notícias	Comentários	População
Rio de Janeiro	460	10.196	6.498.837
Recife	364	1.890	1.625.583
Salvador	131	914	2.938.092
São Paulo	564	14.374	12.038.175

considerar que cada comentário corresponde à opinião de um usuário previamente conhecido. Como a constituição brasileira rege a obrigatoriedade do voto eleitoral para todos os cidadãos, exceto nas situações descritas na seção 2.1, este usuário também pode ser considerado um eleitor.

Para cada comentário, como, por exemplo, o ilustrado na figura 4.1, devemos coletar as seguintes informações :

- Nome do usuário.
  O nome do usuário é um atributo que não pode ser repetido para diferentes usuários, ou seja, só existe um único usuário para cada nome. Este comportamento foi encontrado em todos os jornais utilizados.

- Quantidade de *likes* de cada comentário.
  A quantidade de *likes* representa todos os diferentes usuários que possuem cadastro no jornal em questão e que concordam com a opinião descrita.

- Quantidade de *dislikes* de cada comentário.
  A quantidade de *dislikes* representa todos os diferentes usuários que possuem cadastro no jornal em questão e que discordam da opinião descrita.

- Data do comentário.
  A data que o comentário foi realizado, e para que este seja considerado, precisa ser anterior à data da eleição.

4.4 Escolha dos atributos para geração do modelo

Os atributos escolhidos foram baseados no rastro de intenções de voto que os usuários dos jornais *online* deixam ao expor a sua opinião em forma de comentários positivos, negativos ou neutros, bem como *likes* e *dislikes*, onde concordam ou discordam da opinião dos outros usuários.

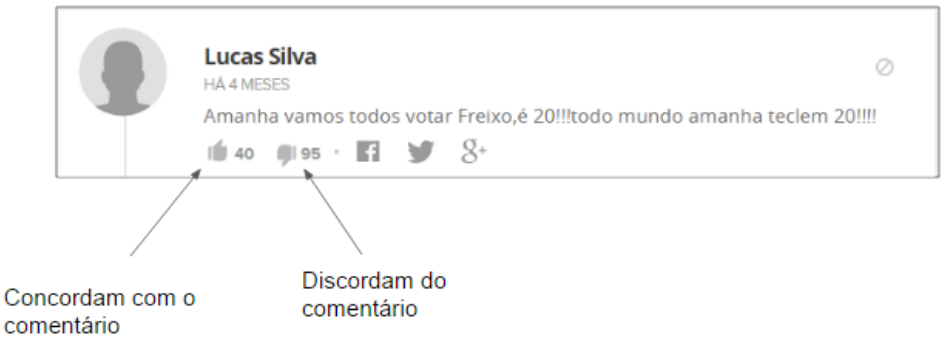

Figura 4.1: Exemplo de comentário

Para a definição dos atributos, temos as seguintes regras :

Seja X o conjunto de candidatos de uma determinada eleição,

Seja C o conjunto de comentários,

Seja Sc a função que avalia a polaridade do sentimento em relação a um comentário que contenha referência ao candidato, considerando cada termo do comentário, tal que o resultado desta função seja P para positivo, N para negativo e Z para neutro.

Sendo k o total de comentários e j o total de candidatos a uma eleição, temos:

CP soma dos usuários que realizaram comentário positivos sobre um determinado candidato de forma que

$$CP(x) = \{\sum_{i=1,}^{K} 1 | Sc(c,x) = P \wedge c \in C \wedge x \in X\},$$

CN soma dos usuários que realizaram comentários negativos sobre um determinado candidato de forma que

$$CN(x) = \{\sum_{i=1,}^{K} 1 | Sc(c,x) = N \wedge c \in C \wedge x \in X\},$$

CZ soma dos usuários que realizaram comentários neutros sobre um determinado candidato de forma que

$$CZ(x) = \{\sum_{i=1,}^{K} 1 | Sc(c,x) = Z \wedge c \in C \wedge x \in X\},$$

Como o mesmo usuário pode expressar *like* e *dislikes* em mais de um comentário, não

podemos simplesmente somar a quantidade de *likes* de todos os comentários e considerar que são eleitores diferentes em potencial. Assim, para utilizar como um possível apoio de diferentes usuários, levamos em consideração o coeficiente de retorno chamado "returning visitor", que indica do total de visitantes de um jornal *online* qual o percentual que se refere a usuários que estão retornando de visitas anteriores.

Seja Lc a função que contabiliza a quantidade de *likes* em um comentário, Dc a função que contabiliza a quantidade de *dislikes* em um comentário e α a constante que indica o percentual de usuários que retornam ao *site*, temos :

Seja MPU a quantidade de usuários que deram *like* para comentários positivos a determinado candidato, de forma que:

$$MPU(x) = \{\max Lc(c,x) + (\sum_{i=1,}^{K} Lc(c,x)) * (1-\alpha) | Sc(c,x) = P \wedge c \in C \wedge x \in X \ \}$$

Seja MNU a quantidade de usuários que deram like para comentários negativos a determinado candidato de forma que

$$MNU(x) = \{\max Lc(c,x) + (\sum_{i=1,}^{K} Lc(c,x)) * (1-\alpha) | Sc(c,x) = N \wedge c \in C \wedge x \in X \ \}$$

Seja MPD a quantidade de usuários que deram dislike para comentários positivos a determinado candidato de forma que

$$MPD(x) = \{\max Dc(c,x) + (\sum_{i=1,}^{K} Dc(c,x)) * (1-\alpha) | Sc(c,x) = P \wedge c \in C \wedge x \in X \ \}$$

Seja MND a quantidade de usuários que deram dislike para comentários negativos a determinado candidato de forma que

$$MND(x) = \{\max Dc(c,x) + (\sum_{i=1,}^{K} Dc(c,x)) * (1-\alpha) | Sc(c,x) = N \wedge c \in C \wedge x \in X \ \}$$

Consideramos como intenção de voto cada umas das funções CP(x) , CN(x) , CZ(x) ,

MNU(x) , MPD(x) e MND(x) e estas foram as escolhidas criação do modelo de regressão

$$CP_p(x) = (\frac{CP(x)}{\sum_{i=1,}^{j} CP(i))}) * 100$$

$$CN_p(x) = (\frac{CN(x)}{\sum_{i=1,}^{j} CN(i))}) * 100$$

$$CZ_p(x) = (\frac{CZ(x)}{\sum_{i=1,}^{j} CZ(i))}) * 100$$

$$MNU_p(x) = (\frac{MNU(x)}{\sum_{i=1,}^{j} MNU(i))}) * 100$$

$$MPD_p(x) = (\frac{MPD(x)}{\sum_{i=1,}^{j} MPD(i))}) * 100$$

$$MND_p(x) = (\frac{MND(x)}{\sum_{i=1,}^{j} MND(i))}) * 100$$

Para a criação do modelo de regressão, utilizamos α , a constante que indica o percentual de usuários que retornam ao *site*, com o valor de 70%. Este valor de "returning visitor"é o valor comum para jornais *online*.

Também incluímos o tamanho da população ao conjunto de variáveis independentes.

4.5 Filtro de análise de sentimento.

Estudos realizados utilizando análise de sentimentos geralmente são baseados na língua inglesa e, dentre os poucos disponíveis na língua portuguesa, escolhemos a base de palavras gerada no estudo Sentilex-PT [35], para realizar a verificação da polaridade do texto de cada comentário no nível de documento. Esta ferramenta foi escolhida por se tratar de um léxico de sentimento especificamente concebido para a análise de sentimento e opinião sobre entidades humanas em textos redigidos em português, que é o caso de comentários de jornais. Atualmente, constituído por 7.014 lemas e 82.347 formas flexionadas.

Para determinar a polaridade de cada comentário, verificamos individualmente cada termo para identificar se este possui polaridade negativa ou positiva de acordo com a base de palavras classificadas. A polaridade resultante do comentário será positiva se a maioria

dos termos forem positivos, negativa se a maioria dos termos forem negativos ou neutra se tivermos a mesma quantidade de termos negativos e positivos.

4.6 Geração do modelo

Para identificarmos a menção ao candidato nos comentários, utilizamos as tabelas 4.4 e 4.5 para as bases de treinamento e teste respectivamente, onde nos casos de nomes com mais facilidade de erro de grafia adicionamos sinônimos. Evitamos alterar muito o nome do candidato para minimizar a influência no resultado.

Utilizamos a ferramenta de mineração de dados *Weka* [18], para aplicar o método de Regressão Linear sobre os dados coletados a respeito dos candidatos, usamos a estratégia de testes 10-Folds cross-validation e o método M5 para seleção de atributos, onde este remove os atributos com menor coeficiente de correlação até que não tenhamos mais melhorias no coeficiente de erro. Repetimos o teste retirando um município a cada execução para verificarmos se algum deles geraria distorções no modelo ou uma grande diferença de coeficiente de correlação, como ilustrado na tabela 4.4. Como não encontramos nenhuma diferença significativa retirando qualquer das cidades, mantemos o modelo com todas elas, que é a execução representada por #1, onde encontramos o maior coeficiente de correlação. Mesmo retirando Manaus, execução #4, por sua baixa quantidade de comentários, o modelo não sofreu alterações significativas e ainda continuou mantendo as mesma variáveis selecionadas.

Tabela 4.4: Nome dos candidatos para a base de treinamento.

Município	Candidato	Palavras-chave adicionais
Belo Horizonte	João Leite	
Belo Horizonte	Kalil	calil
Belo Horizonte	pacheco	paxeco
Belo Horizonte	Reginaldo Lopes	
Belo Horizonte	Délio Malheiros	
Campos dos Goytacazes	Rafael Diniz	Raphael Dinis
Campos dos Goytacazes	Chicão	Xicão
Campos dos Goytacazes	Caio Vianna	Viana
Campos dos Goytacazes	Nildo	
Campos dos Goytacazes	Pudim	
Fortaleza	Roberto Cláudio	
Fortaleza	Capitão Wagner	
Fortaleza	Luizianne Lins	Luiziane
Fortaleza	Heitor Ferrer	
Fortaleza	Ronaldo Martins	
Guarulhos	Guti	
Guarulhos	Eli Correa	Correia
Guarulhos	Eloi Pieta	
Guarulhos	Jorge Wilson Xerife	
Guarulhos	Fausto Miguel Martello	Martelo
Osasco	Rogério Lins	
Osasco	Jorge Lapas	
Osasco	Cláudio Piteri	
Osasco	Oswaldo Vergínio	
Osasco	Valmir Prascidelli	Prascideli
Manaus	Artur Neto	Arthur
Manaus	Marcelo Ramos	
Manaus	Silas Câmara	
Manaus	José Ricardo	
Manaus	Serafim Correa	
Nova Iguaçú	Rogério Lisboa	
Nova Iguaçú	Nelson Bornier	
Nova Iguaçú	Rosângela Gomes	
Nova Iguaçú	Delegado Carlos	
Nova Iguaçú	Leci	
Porto Alegre	Nelson Marchezan	
Porto Alegre	Sebastião Melo	
Porto Alegre	Raul Pont	
Porto Alegre	Maurício Dziedricki	
Porto Alegre	Luciana Genro	
Salvador	Acm Neto	
Salvador	Alice Portugal	
Salvador	Pastor Sagento Isídio	
Salvador	Cláudio	
Salvador	Fábio Nogueira	
Santos	Paulo	
Santos	Vitral	
Santos	Schiff	Schif
Santos	Bosco	
Santos	Edgar	

Tabela 4.5: Nome dos candidatos para a base de testes.

Município	Candidato	Palavras-chave adicionais
Rio de Janeiro	Crivella	Crivela
Rio de Janeiro	Freixo	
Rio de Janeiro	Pedro Paulo	
Rio de Janeiro	Bolsonaro	Bolso
Rio de Janeiro	Índio da Costa	
Recife	Geraldo Júlio	
Recife	Jõao	
Recife	Daniel Coelho	
Recife	Priscila Krause	
Recife	Edilson Silva	
Salvador	ACM neto	
Salvador	Alice Portugal	
Salvador	Pastor Sargento Isidorio	
Salvador	Cláudio	
Salvador	Fábio Nogueira	
São Paulo	João Doria	
São Paulo	Fernando Haddad	adad hadad
São Paulo	Celso Russomanno	russomano
São Paulo	Marta	
São Paulo	Luiza Erundina	

Tabela 4.6: Execuções do processo de geração do modelo.

Execução	Coeficiente de Correlação	Erro médio absoluto
#1	0.8686	8.2033
#2	0.7183	8.7277
#3	0.7189	9.3913
#4	0.7258	9.4013
#5	0.7616	8.6211
#6	0.703	8.2496
#7	0.7461	8.5715
#8	0.6883	9.845
#9	0.8686	8.2118
#10	0.6707	9.7235
#11	0.8313	6.9857

Capítulo 5

Modelo

Como descrito no capítulo 4, o comentário de usuários em jornais *online* se mostra valioso para prevermos a intenção de voto dos eleitores, assim como a interação dos demais usuários que não necessariamente comentam.

O modelo proposto utiliza como base o método de regressão linear com o objetivo de facilitar o entendimento da relação das variáveis independentes, bem como fornecer um resultado que apresente o percentual de votos de cada candidato para que possamos verificar a ordem de classificação dos mesmos.

O diferencial do modelo proposto em relação aos estudos relacionados apresentados no capítulo 3 são:

1. Os comentários são coletados utilizando as notícias que mencionam o nome dos candidatos. Desta forma, não coletamos apenas comentários que mencionem uma palavra-chave específica;
2. Os comentários coletados são referentes a notícias políticas e segmentados por região;
3. Utilizamos a opinião indireta dos eleitores realizadas em forma de *Likes* e *Dislikes*;
4. Podemos analisar o passado sem estar limitado aos filtros por palavra-chave no momento que o usuário realiza o comentário;

Neste capítulo, apresentamos o modelo criado na seção 4 e os atributos do comentário utilizado para a criação do previsor e suas caraterísticas.

5.1 Descrição do modelo

O modelo proposto está baseado no rastro de intenções de voto que os usuários dos jornais *online* deixam ao expor a sua opinião em forma de comentários positivos, negativos ou neutros, bem como *likes* em comentários estritamente negativos, ou seja, que concordam com as opiniões negativas sobre o candidato.

A partir dos jornais escolhidos, filtramos as notícias relacionadas aos candidatos e, a partir dessas notícias, coletamos todos os comentários realizados. De posse de todos os comentários coletados sobre determinada eleição, utilizamos o modelo proposto para prever os candidatos eleitos para o segundo turno. Este fluxo está descrito na figura 5.1.

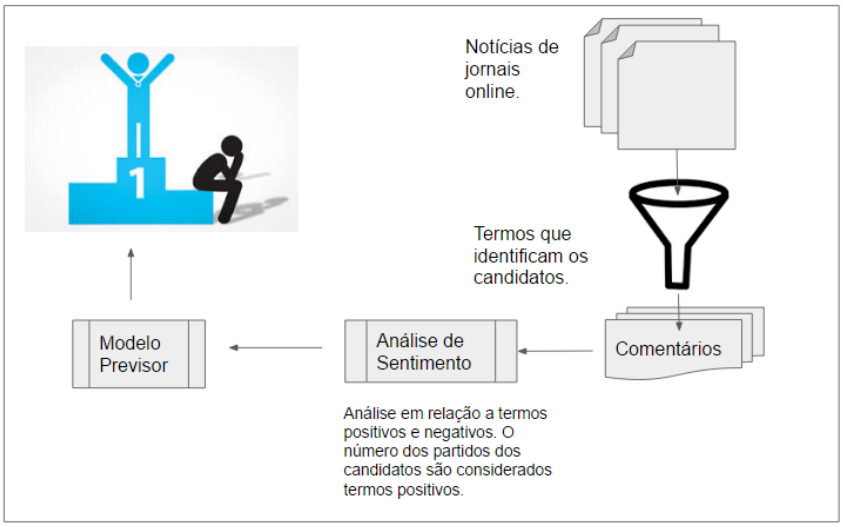

Figura 5.1: Fluxo de execução o modelo.

O modelo é composto pelas seguintes regras:

Seja X o conjunto de candidatos de uma determinada eleição, C o conjunto de comentários, e Sc a função que avalia a polaridade do sentimento em relação a um comentário que contenha referência ao candidato, considerando todo o documento, tal que o resultado desta função seja P para positivo, N para negativo e Z para neutro.

Sendo k o total de comentários e j o total de candidatos a uma eleição, temos:

CP soma dos usuários que realizaram comentário positivos sobre um determinado candidato de forma que

$$CP(x) = \{\sum_{i=1,}^{K} 1 | Sc(c,x) = P \wedge c \in C \wedge x \in X\},$$

CZ soma dos usuários que realizaram comentários neutros sobre um determinado candidato de forma que

$$CZ(x) = \{\sum_{i=1,}^{K} 1 | Sc(c,x) = Z \wedge c \in C \wedge x \in X\},$$

Como descrito na seção 4, MNU(x) significa a quantidade estimada de usuários diferentes que expressaram a sua opinião em forma de *likes* em comentários considerados negativos, levando em consideração o coeficiente de retorno chamado "returning visitor", que indica do total de visitantes de um jornal *online* qual o percentual que se refere a usuários que estão retornando de visitas anteriores. Onde MNU(x) está descrito da seguinte maneira:

Seja Lc a função que contabiliza a quantidade de *likes* em um comentário e α a constante que indica o percentual de usuários que retornam ao *site*, que consideramos ser 70% para os jornais escolhidos, temos:

Seja MNU a quantidade de usuários que deram likes para comentários negativos a determinado candidato, de forma que:

$$MNU(x) = \{\max Lc(c,x) + (\sum_{i=1,}^{K} Lc(c,x)) * (1-\alpha) | Sc(c,x) = N \wedge c \in C \wedge x \in X\ \}$$

Consideramos como intenção de voto cada umas das funções CP(x) , CZ(x) , MNU(x) para encontramos o percentual de cada uma delas para cada candidato de forma que:

$$CP_p(x) = (\frac{CP(x)}{\sum_{i=1,}^{j} CP(i))}) * 100$$

$$CZ_p(x) = (\frac{CZ(x)}{\sum_{i=1,}^{j} CZ(i))}) * 100$$

$$MNU_p(x) = (\frac{MNU(x)}{\sum_{i=1,}^{j} MNU(i))}) * 100$$

De acordo com o estudo piloto, percentual de voto VOTO(x) de cada candidato é determinado pela equação 5.1

$$VOTO(x) = 0.5392 * CP_p(x) + 0.4942 * CZ_p(x) - 0.2854 * MNU_p(x) + 5.0375 \quad (5.1)$$

Capítulo 6

Experimento

Para verificar a eficiência do modelo proposto, utilizamos os dados relativos às notícias sobre candidatos de 4 cidades/município, que foram escolhidas aleatoriamente dentre as 14 predefinidas na seção estudo piloto 4 como base de teste para verificar a eficiência em relação à previsão de percentual de votos e à hipótese de que conseguimos prever os candidatos elegíveis ao segundo turno do processo de eleição.

6.1 Descrição das cidades e os candidatos

As cidades selecionadas para testes foram Rio de Janeiro, Recife, Salvador e São Paulo. As características dessas cidades estão descritas na tabela 4.3. Os comentários foram coletados até o dia 10/01/2016, dia da eleição.

A lista com o nome dos candidatos utilizados para encontramos os comentários relacionados está descrita na tabela 4.5.

6.2 Resultados do experimento

Utilizando a tabela 4.5 como entrada de palavras-chave para localizar os comentários coletados referente aos candidatos e após a execução do modelo, encontramos os resultados listados na tabela 6.1 cujo resultado será discutido no capítulo 7.

Tabela 6.1: Resultado do experimento.

Eleição	Candidato	Real	Previsão
Rio de Janeiro	Crivella	32.62	44.794
Rio de Janeiro	Freixo	21.44	33.026
Rio de Janeiro	Pedro Paulo	18.93	10.869
Rio de Janeiro	Bolsonaro	16.44	4.785
Rio de Janeiro	Índio	10.55	6.488
Recife	Geraldo Júlio	49.72	32.064
Recife	João	23.94	23.076
Recife	Daniel Coeho	18.73	21.203
Recife	Priscila Krause	5.47	13.722
Recife	Edilson Silva	2.11	9.9
Salvador	Acm Neto	74.24	57.471
Salvador	Alice Portugal	14.6	11.543
Salvador	Sargento Isidório	8.64	6.567
Salvador	Cláudio	1.46	13.322
Salvador	Fábio Nogueira	1.04	11.081
São Paulo	João Dória	54.96	39.157
São Paulo	Fernando Haddad	17.22	22.181
São Paulo	Celso Russomanno	14.06	12.019
São Paulo	Marta	10.45	13.079
São Paulo	Erundina	3.28	13.535

Capítulo 7

Discussão dos resultados

Por meio dos resultados obtidos, podemos verificar que, apesar do foco principal dos estudos em relação à coleta de opinião em redes sociais estarem direcionados à utilização do *Twitter* e *Facebook*, existe um grande potencial na análise de comentários de jornais, no qual o usuário está compartilhando a sua opinião diretamente em notícias onde a fonte é reconhecidamente verdadeira.

De acordo com o modelo 5.1, podemos verificar que, quanto maior CP(x), comentários positivos, e maior o CZ(x), comentários neutros, maior será o percentual de votos do candidato. Comentários negativos diminuem a participação percentual do candidato à medida que os usuários concordem com este em forma de "likes". Este fato corrobora com a percepção intuitiva de que quanto mais comentários positivos, maior a chance do candidato, sendo a variável independente CP(x) a que possui o maior coeficiente.

Neste capítulo, apresentamos a avaliação dos resultados, a comparação com diferentes modelos.

7.1 Risco de validade interna

Devido ao fato de agregarmos os comentários de cada candidato em um período de 4 meses, não levamos em consideração fatos relevantes que podem ter ocorrido durante este período que possa ter alterado a opinião do eleitor.

Como consideramos 5 candidatos por municípios e os dados de todos os candidatos são agregados e dependentes dos seus concorrentes, existe uma importante dependência entre os candidatos. Ou seja, para que possamos inferir o voto de um candidato, precisamos conhecer os atributos dos demais concorrentes na eleição do mesmo município.

7.2 Limitações do tamanho da base

As conclusões estão sendo tiradas com uma quantidade de municípios muito pequenas em relação à totalidade de municípios do Brasil. No Brasil, existem 5.561 e este estudo leva em consideração apenas 14, pois foram considerados os relevantes em relação a notícias em jornais *online* que permitem a interação do usuário em forma de comentário.

7.3 Avaliação e interpretação dos resultados

Por meio da tabela 6.3, podemos verificar que acertamos 90% dos casos em que os candidatos foram elegíveis ao segundo turno.

Para compararmos a acurácia com demais algoritmos, incluindo o RNA e a menção de palavras, foi necessário discretizar o resultado em termos de PRIMEIRO, SEGUNDO e OUTROS. A tabela 7.1 ilustra a comparação entre os algoritmos.

O IBOPE apresentou os melhores resultados, acertando 100% dos casos. O modelo proposto, com um custo financeiro e de logística relativamente muito menor do que o IBOPE, obteve 90% de acurácia em relação à colocação dos dois primeiros candidatos. O custo de uma pesquisa no IBOPE para o ano de 2016 foi em torno de R$80.000.00 [38] por pesquisa.

Em relação ao RNA, obtivemos os mesmos 90%. O diferencial do modelo proposto é que levamos em consideração que sempre existirá um candidato vencedor. O modelo de classificação RNA não atende à premissa de que teremos necessariamente um candidato em primeiro lugar, um candidato em segundo lugar e três candidatos classificados como "outros".

Na comparação com o modelo de menção ao nome do candidato, obtivemos uma acurácia superior; pois levamos em consideração o impacto negativo para o candidato em relação aos "likes"em comentários negativos a seu respeito.

7.4 Análise do segundo turno

Para verificar a eficiência do modelo proposto, coletamos comentários de notícias de jornais durante o segundo turno das eleições até o dia da eleição, para as cidades do Rio de Janeiro, nossa cidade natal, e Curitiba. A soma dos percentuais de voto podem não resultar em

Tabela 7.1: Comparação dos modelos de previsão.

Eleição	Candidato	Colocação	IBOPE	RNA	Menção	Modelo
Rio de Janeiro	Crivella	PRIMEIRO	✓PRIMEIRO	✓PRIMEIRO	✓PRIMEIRO	✓PRIMEIRO
Rio de Janeiro	Freixo	SEGUNDO	✓SEGUNDO	PRIMEIRO	OUTROS	✓SEGUNDO
Rio de Janeiro	Pedro Paulo	OUTROS	OUTROS	OUTROS	SEGUNDO	OUTROS
Rio de Janeiro	Bolsonaro	OUTROS	OUTROS	OUTROS	OUTROS	OUTROS
Rio de Janeiro	Índio	OUTROS	OUTROS	OUTROS	OUTROS	OUTROS
Recife	Geraldo Júlio	PRIMEIRO	✓PRIMEIRO	✓PRIMEIRO	SEGUNDO	✓PRIMEIRO
Recife	João	SEGUNDO	✓SEGUNDO	✓SEGUNDO	PRIMEIRO	✓SEGUNDO
Recife	Daniel Coeho	OUTROS	OUTROS	OUTROS	OUTROS	OUTROS
Recife	Priscila Krause	OUTROS	OUTROS	OUTROS	OUTROS	OUTROS
Recife	Edilson Silva	OUTROS	OUTROS	OUTROS	OUTROS	OUTROS
Salvador	Acm Neto	PRIMEIRO	✓PRIMEIRO	✓PRIMEIRO	✓PRIMEIRO	✓PRIMEIRO
Salvador	Alice Portugal	SEGUNDO	✓SEGUNDO	OUTROS	✓SEGUNDO	OUTROS
Salvador	Sargento Isidório	OUTROS	OUTROS	OUTROS	OUTROS	OUTROS
Salvador	Cláudio	OUTROS	OUTROS	OUTROS	OUTROS	SEGUNDO
Salvador	Fábio Nogueira	OUTROS	OUTROS	OUTROS	OUTROS	OUTROS
São Paulo	João Dória	PRIMEIRO	✓PRIMEIRO	✓PRIMEIRO	✓PRIMEIRO	✓PRIMEIRO
São Paulo	Fernando Haddad	SEGUNDO	✓SEGUNDO	✓SEGUNDO	✓SEGUNDO	✓SEGUNDO
São Paulo	Celso Russomanno	OUTROS	OUTROS	OUTROS	OUTROS	OUTROS
São Paulo	Marta	OUTROS	OUTROS	OUTROS	OUTROS	OUTROS
São Paulo	Erundina	OUTROS	OUTROS	OUTROS	OUTROS	OUTROS
		Acurácia	100%	90%	80%	90%

Tabela 7.2: Resultados do segundo turno.

Município	Candidato	Real	Previsão	erro	Posição
Rio de Janeiro	Crivella	59.36	51.4	-7.96	✓PRIMEIRO
Rio de Janeiro	Freixo	40.64	33.47	-7.172	✓SEGUNDO
Curitiba	Rafael Greca	53.25	56.09	2.847	✓PRIMEIRO
Curitiba	Ney Leprevost	46.75	28.77	-17.979	✓SEGUNDO

100% dos votos; pois estes não foram normalizados. Neste exemplo, acertamos todas as posições dos candidatos em ambas as eleições para segundo turno. O resultado está ilustrado na tabela 7.2.

7.5 Análise de termos negativos

Uma forma de utilizar o modelo proposto é utilizarmos somente termos pejorativos sobre os candidatos utilizados durante o período da eleição, para verificar, segundo o modelo, qual impacto estes termos exercem sobre o resultado. Para este experimento, escolhemos os dados coletados durante o segundo turno das eleições do Rio de Janeiro, onde conhecemos bem estes termos.

Para identificar o candidato Crivella, utilizamos os termos Bispo, Macedo e Dizimo. Em contra partida, para o candidato Freixo, usamos os termos Frouxo e Esquerdista.

Tabela 7.3: Análise de termos negativos.

Municipio	Candidato	Real	Previsão	erro	Posição
Rio de Janeiro	Crivella	59.36	72.029	12.669	✓PRIMEIRO
Rio de Janeiro	Freixo	40.64	12.839	-27.801	✓SEGUNDO

O resultado de previsão do modelo ilustrado na tabela 7.3 indica uma grande vantagem para o Candidato Crivella, o que pode trazer à discussão se os termos podem possuir duplo sentido. Não podemos garantir que o termo bispo possa ser considerado negativo para todos os eleitores. Para eleitores evangélicos, esse termo pode ser considerado um elogio.

Capítulo 8

Conclusões e Trabalhos futuros

Nesta dissertação, foi proposto um modelo para previsão de eleição utilizando a opinião dos eleitores expressada em forma de comentários em mídias sociais de conteúdo editorial, de maneira mais rápida e barata do que os tradicionais métodos de pesquisa e opinião. Os resultados obtidos comprovam as hipóteses de que é possível prever os dois candidatos mais bem colocados na eleição municipal, bem como prever o vencedor do segundo turno.

A fonte de informação utilizada foi uma coleção de comentários extraídos de importantes jornais *online* do Brasil, sem direcionamento político definido, onde o conteúdo é criado por jornalistas especializados e os comentários são abertos para qualquer usuário previamente autenticado. Coletamos notícias de cunho político para filtrar opiniões relacionadas aos candidatos ao cargo de prefeito, a fim de prever os resultados das eleições de 2016.

Tendo em vista que estratégias de previsão de eleição no *Twitter*, em geral, utilizam um conjunto arbitrário de palavras-chave para definir o universo de comentários analisados - que por sua vez pode influenciar o estudo, devido ao viés desta escolha -, um dos objetivos deste trabalho foi propor um método para coletar a opinião de eleitores em notícias políticas, visando maximizar a chance de esta opinião estar relacionada ao processo de eleição.

Os experimentos computacionais realizados tiveram como objetivo verificar o desempenho do modelo proposto, que leva em consideração a polaridade do comentário em termos de sentimento e a aprovação e reprovação dos demais eleitores ao mesmo comentário. O desempenho foi comparado com o resultado do IBOPE, com o modelo RNA e de menções ao nome do candidato. Os resultados obtidos mostraram que a estratégia proposta possibilitou prever os dois candidatos com o maior percentual de votos em 90%

dos municípios utilizados na análise do modelo.

8.1 Contribuições

Dentre as principais contribuições apresentadas nesta dissertação, podemos enumerar:

- Criação de robôs para coleta de comentários dos principais jornais do Brasil;
- Utilização de uma fonte de dados alternativa de opinião dos usuários para inferir preferência de candidatos;
- Base de dados de comentários coletados para a realização de futuras análises;
- Modelo de previsão de eleição para identificar os candidatos elegíveis ao segundo turno das eleições municipais, utilizando como fonte de dados comentários de jornais *online* e a polaridade dos comentários em termos de sentimento e bem como "likes"realizados nos comentários negativos.

8.2 Limitações

Dentre as principais limitações encontradas para a realização deste estudo, podemos enumerar:

- A baixa representatividade de notícias nos principais jornais *online* e o fato de muitos jornais de regiões menores não possuírem a funcionalidade de comentários, impossibilitando a análise, devido à vasta quantidade de municípios do Brasil;
- As pessoas escolherem o que, onde e como comentar ou compartilhar nas mídias sociais. Desta forma, estamos analisando apenas pessoas que produzem conteúdo, ao contrário da grande maioria que consome muito mais do que produz;
- A possibilidade de poder gerar uma grande quantidade de falso positivos, por causa de candidatos que não possuem codinomes que os diferenciem de nomes comumente encontrados tais como Marcelo e Silva;
- A não existência de perguntas feitas diretamente ao eleitor. Diferentemente de um processo tradicional de pesquisa de opinião, as sentenças dos comentários coletadas não foram analisadas semanticamente para mapear o entendimento do que foi publicado;

- As sentenças negativas serem detectadas com mais precisão do que as sentenças positivas, pois as opiniões geradas de maneira informal podem estar carregadas de ironias e palavras de duplo sentido e que são difíceis de detectarmos apenas com métodos automatizados de análise de sentimentos. De acordo com carvalho2009clues, mais de 35% das opiniões consideradas positivas estão relacionadas à ironia;

- As pesquisas eleitorais inferirem sobre a intenção de votos do eleitor e terem o papel de mostrar como a opinião está se formando, mas não garantirem o número exato do resultado;

- O fato de estarmos assumindo que todas as opiniões são verdadeiras e que representam a intenção de voto do eleitor. Expressar a opinião em mídias sociais não garante nenhum compromisso com a real intenção e o fato de o usuário utilizar pseudônimos o protege de qualquer vínculo com sua própria opinião.

8.3 Trabalhos futuros

Em relação à análise de sentimentos, utilizamos uma estratégia baseada em dicionários [35], sem levar em consideração a adição de novos termos, além do número do partido que o candidato representa. Termos estes que poderiam classificar melhor os comentários; pois existe questões de regionalização da língua portuguesa bem como palavras que descrevem o candidato, mas que só seriam conhecidas dentro de um determinado contexto sobre as particularidades dos candidatos, tais como: o termos "bispo", para se dirigir negativamente ao candidato Crivella nas eleições da prefeitura do Rio de Janeiro; ou, então, o termo "frouxo", para se opor ao candidato Marcelo Freixo nesta mesma eleição. Além disso, sentenças negativas são detectadas com mais precisão do que as sentenças positivas; pois as opiniões geradas de maneira informal podem estar carregadas de ironias e palavras de duplo sentido e que são difíceis de detectarmos com métodos automatizados de análise de sentimentos. De acordo com [8] , 35% das opiniões consideradas positivas estão relacionadas à ironia. Nesta direção, uma possibilidade de trabalho futuro seria adicionar os termos regionais que identificam os candidatos bem como aplicar outras técnicas de análise de sentimento, como a estratégia estatística e evolutiva [7] para selecionar palavras relacionadas a este contexto específico e, também, a diferenciação de ironia e *emoticons* [22] nos comentários realizados. Ainda neste sentido, este estudo não considerou a existência da citação de mais de um candidato no mesmo comentário. Uma contribuição futura poderia identificar as entidades do texto não estruturado e realizar

esta diferenciação de contexto.

Apesar do resultado ter se mostrado eficiente na previsão dos candidatos elegíveis ao segundo turno das eleições municipais, devemos considerar a existência de importantes redes de usuários em mídias sociais, em que a possibilidade de trabalho futuro seria a criação de um previsor que levasse em consideração não apenas comentários em notícias de jornais como fonte de dados, mas ao mesmo tempo *tweets* relacionados e informações estatísticas retiradas do *Facebook*, como a quantidade de seguidores na página do candidatos durante o período da eleição.

Devido à grande quantidade de comentários que não foram utilizados no processo de análise, trabalhos futuros poderiam considerar demais comentários do mesmo eleitor, cuja polaridade tenha sido previamente definida, para mapear a opinião de demais eleitores que apoiem ou rejeitem estes comentários.

Referências

[1] Almeida, J. M.; Pappa, G. L., et al. Twitter population sample bias and its impact on predictive outcomes: a case study on elections. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015* (2015), ACM, pp. 1254–1261.

[2] Anand, S.; Venkataraman, M.; Subbalakshmi, K.; Chandramouli, R. Spatio-temporal analysis of passive consumption in internet media. *IEEE Transactions on Knowledge and Data Engineering 27*, 10 (2015), 2839–2850.

[3] Barclay, F. P.; Pichandy, C.; Venkat, A.; Sudhakaran, S. India 2014: Facebook 'like'as a predictor of election outcomes. *Asian Journal of Political Science 23*, 2 (2015), 134–160.

[4] BOURDIEU, P. A opinião pública não existe em crítica metodológica, investigação social e enquete operária, 5' edição, michel thiollent (autor e org).

[5] Bovet, A.; Morone, F.; Makse, H. A. Predicting election trends with twitter: Hillary clinton versus donald trump. *arXiv preprint arXiv:1610.01587* (2016).

[6] Burnap, P.; Gibson, R.; Sloan, L.; Southern, R.; Williams, M. 140 characters to victory?: Using twitter to predict the uk 2015 general election. *Electoral Studies 41* (2016), 230–233.

[7] CARVALHO, J. D. S. Uma estratégia estatística e evolutiva para mineração de opiniões em tweets.

[8] Carvalho, P.; Sarmento, L.; Silva, M. J.; De Oliveira, E. Clues for detecting irony in user-generated contents: oh...!! it's so easy;-. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion* (2009), ACM, pp. 53–56.

[9] Choi, H.; Varian, H. Predicting the present with google trends. *Economic Record 88*, s1 (2012), 2–9.

[10] Chu, Z.; Gianvecchio, S.; Wang, H.; Jajodia, S. Who is tweeting on twitter: human, bot, or cyborg? In *Proceedings of the 26th annual computer security applications conference* (2010), ACM, pp. 21–30.

[11] DE JUSTIÇA, T. As pesquisas eleitorais e a democracia: necessidade de novas exigências técnico-legais.

[12] Dokoohaki, N.; Zikou, F.; Gillblad, D.; Matskin, M. Predicting swedish elections with twitter: A case for stochastic link structure analysis. In *Proceedings*

of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (2015), ACM, pp. 1269–1276.

[13] Dwi Prasetyo, N.; Hauff, C. Twitter-based election prediction in the developing world. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (2015), ACM, pp. 149–158.

[14] Ferraz, C., et al. Crítica metodológica às pesquisas eleitorais no brasil.

[15] Gayo-Avello, D. Don't turn social media into another'literary digest'poll. *Communications of the ACM 54*, 10 (2011), 121–128.

[16] Gayo-Avello, D. No, you cannot predict elections with twitter. *IEEE Internet Computing 16*, 6 (2012), 91–94.

[17] Hagar, D. # vote4me: the impact of twitter on municipal campaign success. In *Proceedings of the 2015 International Conference on Social Media & Society* (2015), ACM, p. 19.

[18] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The weka data mining software: an update. *ACM SIGKDD explorations newsletter 11*, 1 (2009), 10–18.

[19] Khatua, A.; Khatua, A.; Ghosh, K.; Chaki, N. Can# twitter_trends predict election results? evidence from 2014 indian general election. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on* (2015), IEEE, pp. 1676–1685.

[20] Liu, B. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition*. Chapman and Hall/CRC, 2010, pp. 627–666.

[21] Liu, B. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies 5*, 1 (2012), 1–167.

[22] Liu, K.-L.; Li, W.-J.; Guo, M. Emoticon smoothed language models for twitter sentiment analysis. In *AAAI* (2012).

[23] Mendoza, M.; Poblete, B.; Castillo, C. Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics* (2010), ACM, pp. 71–79.

[24] Miranda Filho, R.; Almeida, J. M.; Pappa, G. L. Twitter population sample bias and its impact on predictive outcomes: a case study on elections. In *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on* (2015), IEEE, pp. 1254–1261.

[25] Morstatter, F.; Pfeffer, J.; Liu, H.; Carley, K. M. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *arXiv preprint arXiv:1306.5204* (2013).

[26] Mustafaraj, E.; Finn, S.; Whitlock, C.; Metaxas, P. T. Vocal minority versus silent majority: Discovering the opionions of the long tail. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (2011), IEEE, pp. 103–110.

[27] Mustafaraj, E.; Metaxas, P. T. From obscurity to prominence in minutes: Political speech and real-time search.

[28] Pennington, N.; Winfrey, K. L.; Warner, B. R.; Kearney, M. W. Liking obama and romney (on facebook): An experimental evaluation of political engagement and efficacy during the 2012 general election. *Computers in Human Behavior 44* (2015), 279–283.

[29] Reis, J.; Gonçalves, P.; Vaz de Melo, P.; Prates, R.; Benevenuto, F. Magnet news: You choose the polarity of what you read. *Proceedings of ICWSM* (2014).

[30] Riker, W. H.; Ordeshook, P. C. A theory of the calculus of voting. *American political science review 62*, 01 (1968), 25–42.

[31] Saleiro, P.; Gomes, L.; Soares, C. Sentiment aggregate functions for political opinion polling using microblog streams. In *Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering* (2016), ACM, pp. 44–50.

[32] Sameki, M.; Gentil, M.; Mays, K. K.; Guo, L.; Betke, M. Dynamic allocation of crowd contributions for sentiment analysis during the 2016 us presidential election. *arXiv preprint arXiv:1608.08953* (2016).

[33] Sápiras, L. A.; Becker, K. Mineraç ao da opiniao sobre aspectos de candidatos a eleiç oes em comentários de notıcias. SBBD.

[34] Shi, L.; Agarwal, N.; Agrawal, A.; Garg, R.; Spoelstra, J. Predicting us primary elections with twitter. *URL: http://snap. stanford. edu/social2012/papers/shi. pdf* (2012).

[35] Silva, M. J.; Carvalho, P.; Sarmento, L. Building a sentiment lexicon for social judgement mining. In *International Conference on Computational Processing of the Portuguese Language* (2012), Springer, pp. 218–228.

[36] Sunstein, C. R. Republic. com, princeton. *Telhami, Shibley: 2010 Arab Public Opinion Poll (conducted by the University of* (2001).

[37] Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-based methods for sentiment analysis. *Computational linguistics 37*, 2 (2011), 267–307.

[38] Trebilcox-Ruiz, P. Consulta às pesquisas registradas, 2017. Disponível em `http://www.tse.jus.br/eleicoes/eleicoes-2016/pesquisas-eleitorais/consulta-as-pesquisas-registradas`.

[39] Tsytsarau, M.; Palpanas, T. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery 24*, 3 (2012), 478–514.

[40] Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; Welpe, I. M. Election forecasts with twitter: How 140 characters reflect the political landscape. *Social Science Computer Review* (2010), 0894439310386557.

[41] Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; Welpe, I. M. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM 10* (2010), 178–185.

[42] Venkataraman, M.; Subbalakshmi, K.; Chandramouli, R. Measuring and quantifying the silent majority on the internet. In *Sarnoff Symposium (SARNOFF), 2012 35th IEEE* (2012), IEEE, pp. 1–5.

[43] Weiser, M. The computer for the 21st century. *Scientific american 265*, 3 (1991), 94–104.

[44] Wong, F. M. F.; Tan, C. W.; Sen, S.; Chiang, M. Quantifying political leaning from tweets, retweets, and retweeters. *IEEE Transactions on Knowledge and Data Engineering 28*, 8 (2016), 2158–2172.

APÊNDICE A – Utilizando Gebish e Grovvy para capturar notícias de jornais.

Configurando o Gradle Criando um projeto

1. Criar uma pasta para o projeto.

2. Abrir o prompt de comando: Acesse o diretório da pasta Execute o comando: “mkdir crawler” Execute o comando: “gradle init –type java-library”

3. Editando o arquivo build.gradle gerado.

   Adicionar as seguintes linhas :
   apply plugin: 'groovy'
   Incluir as dependências:
   compile 'org.codehaus.groovy:groovy-all:2.4.4'
   compile "org.gebish:geb-core:0.12.2"
   compile "org.seleniumhq.selenium:selenium-firefox-driver:2.53.0"
   compile "org.seleniumhq.selenium:selenium-chrome-driver:2.53.0"
   compile "org.seleniumhq.selenium:selenium-support:2.53.0"

4. Criar pastas para o groovy e o arquivo que será executado. Criar as pastas *src/main/groovy* *src/main/groovy/scripts* Criar classe do projeto nesta pasta: *Ex.: crawlernews.groovy*

5. Editando o aquivo. Colocar na primeira linha: *package scripts;* para indicar que este é o pacote correspondente a pasta como ilustrado na figura A.1.

6. Criando a task que será executada.

   Abrir a classe build.gradle na raiz do projeto. Criar uma task no final do arquivo com ilustrado na figura A.2. *task globo (dependsOn: 'classes', type: JavaExec) { main = 'scripts.crawlernews'*

classpath = sourceSets.main.runtimeClasspath

}

7. Executando o processo.

   Para executar o processo basta executar o comando *gradle globo* a partir da pasta do projeto.

```groovy
package scripts
import geb.Browser;
import org.openqa.selenium.firefox.FirefoxDriver
import org.openqa.selenium.firefox.FirefoxProfile
import org.openqa.selenium.JavascriptExecutor;
//Abrindo o browser
def browser = new Browser(driver:getFirefoxDriver())
browser.driver.manage().window().maximize()
// Começando a captura pela página inicial do globo
String url = "http://www.globo.com"
browser.go(url)
def array = new ArrayList<String>();
// Percorrendo pelos links de matérias e adicionando em um vetor
browser.$("a", class:"hui-premium__link hui-highlight__link").each{
    println "Materia: " + it.$("p", class:"hui-premium__title").text()
    println "Link: " + it.@href
    println "Title: " + it.@title
    println "======================================================"
    array.add(it.@href)
}
// Imprimindo as notícias
println "Materias secundarias"
browser.$("a", class:"topglobocom__content-title").each{
    println "Materia: " + it.text()
    println "Link: " + it.@href
    println "Title: " + it.@title
    println "======================================================"
}
// Navegando pelas noticias e aguardando 2s entre cada uma.
for(int i = 0; i < array.size(); i++){
    browser.go(array[i])
    sleep(2000)
}
//Método que abre o firefox
def getFirefoxDriver(){
    FirefoxProfile profile = new FirefoxProfile();
    profile.setPreference("permissions.default.image", 2);
    FirefoxDriver ffDriver = new FirefoxDriver(profile);
    return ffDriver;
}

```

Figura A.1: Código fonte do arquivo crawlernews.groovy

```
/*
 * This build file was auto generated by running the Gradle 'init' task
 * by 'marcussouza' at '13/12/16 16:17' with Gradle 2.13
 *
 * This generated file contains a sample Java project to get you started.
 * For more details take a look at the Java Quickstart chapter in the Gradle
 * user guide available at https://docs.gradle.org/2.13/userguide/tutorial_java_projects.html
 */

// Apply the java plugin to add support for Java
apply plugin: 'java'
apply plugin: 'groovy'

// In this section you declare where to find the dependencies of your project
repositories {
    // Use 'jcenter' for resolving your dependencies.
    // You can declare any Maven/Ivy/file repository here.
    jcenter()
}

// In this section you declare the dependencies for your production and test code
dependencies {
    // The production code uses the SLF4J logging API at compile time
   compile 'org.slf4j:slf4j-api:1.7.21'
   compile 'org.codehaus.groovy:groovy-all:2.4.4'
   compile "org.gebish:geb-core:0.12.2"
   compile "org.seleniumhq.selenium:selenium-firefox-driver:2.53.0"
   compile "org.seleniumhq.selenium:selenium-chrome-driver:2.53.0"
   compile "org.seleniumhq.selenium:selenium-support:2.53.0"

    // Declare the dependency for your favourite test framework you want to use in your tests.
    // TestNG is also supported by the Gradle Test task. Just change the
    // testCompile dependency to testCompile 'org.testng:testng:6.8.1' and add
    // 'test.useTestNG()' to your build script.
    testCompile 'junit:junit:4.12'
}

task globo (dependsOn: 'classes', type: JavaExec) {
  main = 'scripts.crawlernews'
  classpath = sourceSets.main.runtimeClasspath
}
```

Figura A.2: Código fonte do arquivo build.gradle

APÊNDICE B – Capturando notícias diretamente da busca orgânica do Google

O código de exemplo abaixo exemplifica a carga de notícias realizada diretamente pelo site do Google [1].

```
1  package scripts
2  //Importação das bibliotecas necessárias
3  import geb.Browser;
4  import org.openqa.selenium.firefox.FirefoxDriver
5  import org.openqa.selenium.firefox.FirefoxProfile
6
7  //Abrindo o browser Firefox
8  def browser = new Browser(driver:getFirefoxDriver())
9
10 //Indicação do nome do candidato para encontrar as notícias relacionadas
11 String candidato = "Nelson+Marchezan+junior"
12
13 //A variável site indica o filtro que será utilizado para trazer apenas
14 //notícias relacionadas a este endereço escolhido
15 String site = "http://www.gazetadopovo.com.br/vida-publica/eleicoes/2016/"
16
17 //url da busca do google concatenando as variáveis de configuração
18 // o parâmetro num=100 indica que queremos receber 100 resultados por página
19 String url = "https://www.google.com.br/search?q=${candidato}+site:${site}&num=100"
20 browser.go(url)
21
22 //A tag HTML h3 é a que contém os links retornados pela busca
23 browser.$("h3", class:"r").$("a").each{
24     String noticia = it.@href;
25     //imprimindo todas as notícias retornadas.
26     println "noticia "+noticia;
27 }
28
29 def getFirefoxDriver(){
30     FirefoxProfile profile = new FirefoxProfile();
31     FirefoxDriver ffDriver = new FirefoxDriver(profile);
32     return ffDriver;
33 }
```

Figura B.1: Carga de notícias diretamente pelo Google.

[1] http://www.google.com.br

APÊNDICE C – Framework de captura dos comentários

Diferentemente do Twitter, em geral os jornais online não fornecem nenhuma API que facilite a coleta das notícias e comentários e por este motivo criamos um web bot utilizando o framework de automação Gebish [1], que é que por sua vez é baseado no sistema de testes de software Selenium Webdriver [2], onde programamos na linguagem Groovy a simulação da navegação do usuário para acessar as informações dos comentários das notícias. Um exemplo de utilização do Gebish em conjunto com Groovy para ler notícias de jornais esta descrito no apêndice A. Ao contrário de facilitar este procedimento, existe uma tendência de todos os jornais online de limitar o acesso as notícias apenas para usuários cadastrados. Esta abordagem de criação de um robô que simula a navegação foi escolhida para evitar bloqueios de acesso a informação.

Para o processo de captura de dados, este robô simula a navegação no web site dos jornais escolhidos capturando as notícias relacionadas utilizando como primeiro passo a busca do próprio jornal utilizando como palavra-chave o nome de cada um dos candidatos. Devido a limitação de inexistência de API, para cada um dos jornais escolhidos foi necessário criar um robô específico, onde este identifica o conteúdo através do posicionamento exato de cada elemento mapeado do conteúdo HTML em cada uma das páginas relacionadas ao conteúdo das notícias e comentários. Por exemplo, para o jornal G1 foi necessário identificar as páginas de busca, notícias e comentários como ilustrados nas figuras C.1, C.2, C.3 e C.4 que representam o mapeamento destas informações respectivamente. A exceção a esta regra foi o jornal Gazeta do Povo pois ele utiliza internamente o motor de busca Google e por isso a estratégia de coleta de notícias utilizou o processo descrito no apêndice B onde a parte do processo de busca de notícias é realizada através do serviço de busca Google [3].

[1] http://gebish.org
[2] http://www.seleniumhq.org/projects/webdriver/
[3] http://google.com

Figura C.1: Localiza as notícias relativas aos candidatos.

Figura C.2: Localização da informação de quantas páginas precisam ser coletadas.

Este framework foi criado de forma que pode abrir vários navegadores web e coletar simultaneamente as informação de vários jornais e estas informações de notícias e comentários foram armazenadas em um banco de dados relacional para que pudessem ser utilizadas na geração do modelo de previsão. O modelo de entidades e relacionamentos esta descrito na figura C.5.

Figura C.3: Localização da informação do link da notícia.

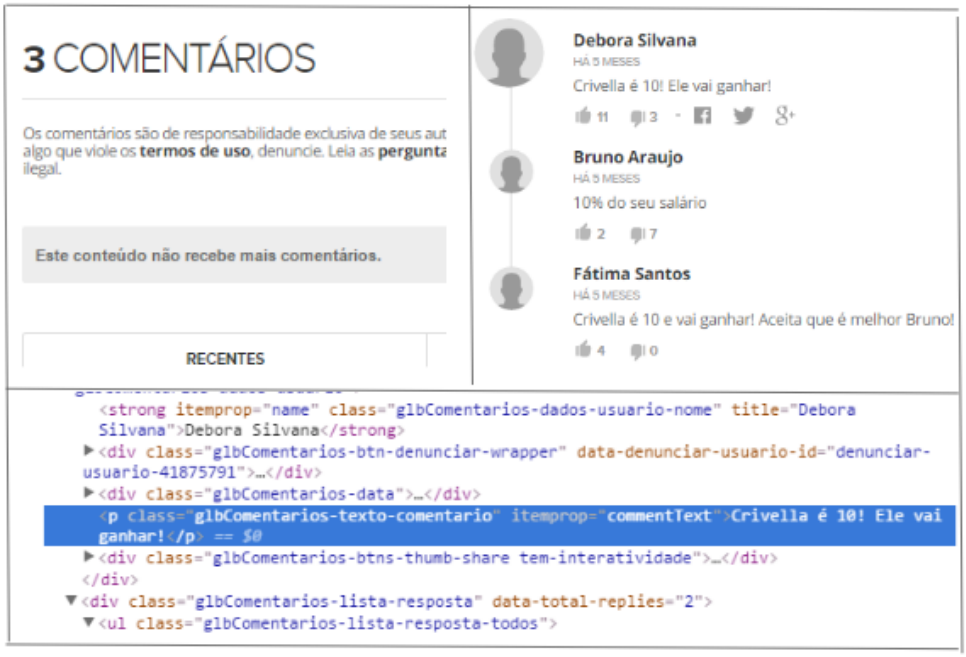

Figura C.4: Coleta dos comentários.

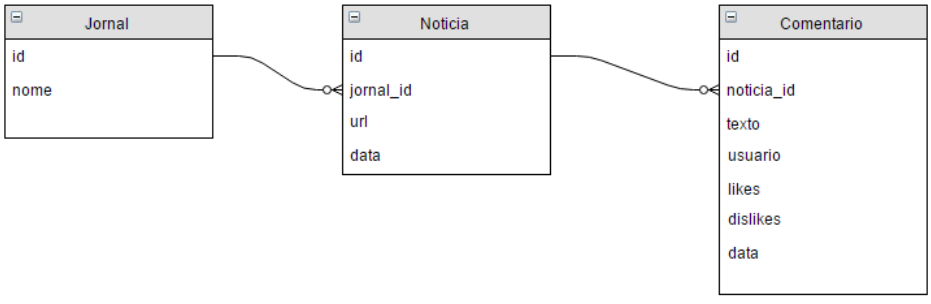

Figura C.5: Modelo de dados para armazenamento das notícias.