UNIVERSIDADE FEDERAL FLUMINENSE

WILTON DE PAULA FILHO

Utilizando hashtags e conteúdo da descrição do perfil do usuário para melhorar a classificação de tweets no cenário eleitoral

NITERÓI

2018

UNIVERSIDADE FEDERAL FLUMINENSE

WILTON DE PAULA FILHO

Utilizando hashtags e conteúdo da descrição do perfil do usuário para melhorar a classificação de tweets no cenário eleitoral

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Doutor em Computação. Área de concentração: Engenharia de Sistemas e Informação.

Orientador:
Prof. Dr. José Viterbo Filho

Co-orientadora:
Profa. Dra. Isabel Cristina Mello Rosseti

NITERÓI

2018

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

P324u Paula filho, Wilton de
Utilizando hashtags e conteúdo da descrição do perfil do usuário para melhorar a classificação de tweets no cenário eleitoral / Wilton de Paula filho ; José Viterbo Filho, orientador ; Isabel Cristina Mello Rosseti, coorientador. Niterói, 2018.
115 f. : il.

Tese (doutorado)-Universidade Federal Fluminense, Niterói, 2018.

DOI: http://dx.doi.org/10.22409/PGC.2018.d.04715232677

1. Mineração de opiniões. 2. Classificação do sentimento de tweets no cenário eleitoral. 3. Algoritmos de aprendizado de máquina. 4. Política no Twitter. 5. Produção intelectual. I. Filho, José Viterbo, orientador. II. Rosseti, Isabel Cristina Mello, coorientador. III. Universidade Federal Fluminense. Escola de Engenharia. IV. Título.

CDD -

Bibliotecária responsável: Fabiana Menezes Santos da Silva - CRB7/5274

WILTON DE PAULA FILHO

Utilizando *hashtags* e conteúdo da descrição do perfil do usuário para melhorar a classificação de *tweets* no cenário eleitoral

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Doutor em Computação. Área de concentração: Engenharia de Sistemas e Informação.

Aprovada em dezembro de 2018.

BANCA EXAMINADORA

Prof. Dr. José Viterbo Filho - Orientador, UFF

Profa. Dra. Isabel Cristina Mello Rosseti - Co-orientadora, UFF

Prof. Dra. Aline Marins Paes Carvalho, UFF

Prof. Dra. Flávia Cristina Bernardini, UFF

Prof. Dra. Adriana Santarosa Vivacqua, UFRJ

Prof. Dra. Jonice de Oliveira Sampaio, UFRJ

Prof. Dra. Solange Oliveira Rezende, USP

Niterói

2018

=

"Embora ninguém possa voltar atrás e fazer um novo começo, qualquer um pode começar agora e fazer um novo fim." (Chico Xavier)

Agradecimentos

O meu primeiro agradecimento, indubitavelmente, vai a DEUS por me proporcionar todas as condições que um ser humano precisa para conseguir chegar ao término de um doutorado em Ciência da Computação: saúde física e mental, condição financeira, paz de espírito, apoio da família, amigos, professores, etc.

Segundo, quero deixar registrado o meu agradecimento ao meu guia e modelo, Jesus Cristo, por ter acompanhado os meus passos e por ter me ajudado a concluir esse projeto tão importante em minha vida pessoal e profissional, o doutorado.

Terceiro, agradecer todo o apoio que a minha família me proporcionou. Primeiro, a minha queridíssima esposa Camilla Macedo Rocha de Paula pelo apoio incondicional e que não me deixou desistir desse ciclo tão importante em minha vida. A sua fé, o seu amor, a sua mão, o seu olhar, as suas palavras e a sua renúncia me fortaleceu em muitos momentos em que eu achava que eu ia fraquejar. Devo muito esse doutorado a você. A minha amada filha Clara Rocha de Paula, que nasceu durante o período de realização do doutorado e que apesar de tão pequenininha foi capaz de me ensinar tantas coisas, como a perseverança. A minha querida e amada mãe, Joane D'arc Melo de Paula, que teve que aprender a conviver com a minha ausência em muitos momentos em que a minha presença era frequente, e aos meus irmãos pela ajuda, força, exemplos e tolerância. A minha sogra, Rozangela Macedo, por ter se ausentado do seu lar diversas vezes para oferecer suporte a minha família. Ao meu pai, Wilton de Paula (em memória), pelos exemplos de homem de bem e por ter sido o primeiro a me incentivar a fazer doutorado.

Quarto, ao Instituto Federal do Triângulo Mineiro (IFTM) por ter me oferecido todos os recursos necessários para realização dessa capacitação.

Quinto, aos meus orientadores. A professora Rosseti, por ter aceitado a co-orientação desse trabalho e por ter me ajudado nos momentos mais difíceis do processo de doutoramento. Ao professor Viterbo, por ter aceitado a orientação desse trabalho e ter acreditado na ideia, além de toda contribuição técnica.

Por último, obrigado a todos os amigos, professores, técnicos-administrativos, etc. que contribuíram para realização deste trabalho.

Resumo

Tweets, no cenário eleitoral, têm sido utilizados em pesquisas científicas sob diversas perspectivas, por exemplo para predizer o resultado de eleições presidenciais e para investigar reações de eleitores durante eventos, como debates eleitorais. A mineração de opiniões tem sido uma das abordagens utilizadas para avaliar a opinião expressa por usuários nesse tipo de mensagem. Melhorar a acurácia do sentimento dessa categoria de *tweet* tem sido um dos desafios enfrentados pelos pesquisadores, devido alguns fatores, tais como a quantidade limitada de caracteres permitidos em um *tweet* e o uso de *hashtags* e *slogans* de campanha para expressar opiniões políticas sobre candidatos. No cenário de eleições, *hashtags* têm sido utilizadas em *tweets* com uma frequência cada vez maior. Em diversos trabalhos reportados na literatura, *hashtags* têm sido utilizadas para coletar e selecionar *tweets*, rotular mensagens, além de receberem tratamentos especiais na fase de pré-processamento ou são simplesmente descartadas das análises. A contribuição de *hashtags* na análise de sentimentos de *tweets* no cenário eleitoral tem sido pouco explorada. Neste trabalho, são consideradas duas categorias de *hashtags*, as políticas e as não-políticas, e dois novos atributos baseados em *hashtags* políticas, denominados $TPSB$ e $DPSB$, que refletem na melhoria do desempenho de algoritmos de aprendizado de máquina supervisionado, utilizados no processo de classificação do sentimento de *tweets* no cenário eleitoral. A análise experimental proposta neste trabalho, foi realizada a partir de duas amostras rotuladas manualmente contendo mensagens coletadas em períodos de eleições presidenciais, cada uma com, aproximadamente, 4.000 *tweets*. Foi avaliada a contribuição do uso das *hashtags* contidas em *tweets* e em descrições de perfis de usuários, e atributos para a melhoria da acurácia de classificadores baseados em *Naive Bayes* (NB), *Multinomial Naive Bayes* (MNB) e *Support Vector Machine* (SVM). Nos testes realizados com uma das amostras, as *hashtags* políticas foram responsáveis por aumentar, com significância estatística, as acurácias de todos os classificadores. Em outro *dataset*, as acurácias de todos os classificadores foram incrementadas e, em 66% dos casos, os aumentos obtidos foram estatisticamente significantes. Em um dos experimentos propostos, a acurácia do algoritmo NB, utilizando o formato unigrama, foi incrementada em 3,2% ao considerar *hashtags* políticas nas análises. Os resultados obtidos sugerem que *hashtags* contendo palavras fazendo referência a candidatos, a partir do primeiro nome e/ou sobrenome deles associado a outras palavras/números, *slogans* de campanha e palavras de repúdio, podem ser úteis na classificação do sentimento de *tweets* políticos e que usuários do *Twitter*, durante períodos de campanha eleitoral, expressam opinião política não somente a partir de *tweets*, mas também a partir de suas descrições de perfis.

Palavras-chave: *twitter*, *tweet*, *hashtag*, descrição do perfil do usuário, política, mineração de opinião, análise de sentimentos.

Abstract

Tweets in the election scene have been used in scientific research from a variety of perspectives, for example to predict the outcome of presidential elections and to investigate voter reactions during events such as electoral debates. The opinion mining has been one of the approaches used to evaluate the opinion expressed by the user in this type of message. Improving the accuracy of this category of tweet has been one of the challenges faced by researchers due to factors such as the limited amount of characters allowed in a tweet and the use of hashtags and campaign slogans to express political opinions about candidates. In the election scene, hashtags have been used in tweets with increasing frequency. In several papers reported in the literature, hashtags have been used to collect and select tweets, label messages, and receive special treatments in the preprocessing phase or are simply discarded from the analyzes. The contribution of hashtags in the sentiment tweets analysis in the election scene has been little explored. In this work, two categories of hashtags, the policies and the non-policies, and two new features based on political hashtags, called TPSB and DPSB, are considered, which reflect on the performance improvement of supervised machine learning algorithms used in the tweets sentiment classification in the election scene. The experimental analysis proposed in this work was performed from two manually labeled samples containing messages collected during periods of presidential elections, each with approximately 4,000 tweets. The contribution of the use of hashtags contained in tweets and descriptions of user profiles and attributes to improve the accuracy of Naive Bayes (NB), Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM) classifiers was evaluated. In the tests performed with one of the samples, political hashtags were responsible for increasing, with statistical significance, the accuracy of all classifiers. In another dataset, the accuracy of all classifiers was increased and, in 66% of cases, the increases obtained were statistically significant. In one of the proposed experiments, the accuracy of the NB algorithm, using the unigram format, was increased by 3.2% when considering political hashtags in the analyzes. The results obtained suggest that hashtags containing words referring to candidates, from their first name and / or surname associated with other words / numbers, campaign slogans and repudiation words, may be useful in classifying the political tweets sentiment and that Twitter users, during campaigning periods, express political opinion not only from tweets but also from their profile descriptions.

Keywords: twitter, tweet, hashtag, description of user profile, politics, opinion mining, sentiment analysis.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

#: *hashtag*

AB: *Adaptative Boosting*

ADA-SVM: *Adaboost com Support Vector Machine*

API: *Application Programming Interface*

AS: Análise de Sentimentos

BLR: *Bayesian Logistic Regression*

BOW: *Bag-Of-Words*

CPP1: Conjunto de Pré-processamento 1

CPP2: Conjunto de Pré-processamento 2

CPP3: Conjunto de Pré-processamento 3

CRF: *Conditional Random Field*

DH: *Description Hashtag*

DPSB: *Description Political Support Bit*

DT: *Decision Tree*

EUA: Estados Unidos da América

F1: *F1-Measure*

FBI: *Federal Bureau of Investigation*

FN: *False Negative*

FP: *False Positive*

HTTP: *Hyper Text Transfer Protocol*

LR: *Logistic Regression*

MAGA: *Make America Great Again*

MaxEnt: *Maximum Entropy*

MNB: *Multinomial Naive Bayes*

MPQA: *Multi-Perspective Question Answering*

N: Neuto

NB: *Naive Bayes*

NLTK: *Natural Language Tookit*

NLTK: *Natural Language Tookit*

PA: Pró-Aécio

PC1: Pró-Candidato 1

PC1: Pró-Candidato 2

PD: Pró-Dilma

PH: Pró-Hillary

PLN: Processamento de Linguagem Natural

PMI: *Pointwise Mutual Information*

POS *Part-Of-Speech*

PSDB: Partido da Social Democracia Brasileira

PT: Partido dos Trabalhadores

PT: Pró-Trump

RF: *Random Forrest*

RT: *Retweet*

SVM: *Support Vector Machine*

TF-IDF: *Term Frequency - Inverse Document Frequency*

TF: *Term-Frequency*

TH: *Tweet Hashtag*

TN: *True Negative*

TP: *True Positive*

TPSB: *Tweet Political Support Bit*

URL: *Uniform Resource Locator*

VP: *Voted Perceptron*

Sumário

Capítulo 1

Introdução

Eleição para escolha de um governante é uma parte importante na democracia de qualquer instância, seja ela um país, estado ou cidade. Um elemento importante numa eleição é a pesquisa de opinião pública. Em [63], os autores afirmaram que o principal objetivo desse tipo de pesquisa é fornecer informações aos cidadãos e também as partes interessadas para que esses possam fazer os ajustes apropriados, caso julguem necessário. Institutos específicos realizam, durante períodos de campanha eleitoral, por exemplo para escolha do presidente de um país, pesquisas de opinião pública com o objetivo principal de analisar as intenções de voto dos eleitores, a partir, geralmente, de um conjunto de definições, tais como a abrangência do levantamento da pesquisa (com ou sem segundo turno, com ou sem rejeição, etc.), o contexto da pesquisa (país, estado ou município), a estratificação da amostra (definição da proporção entre pessoas do sexo masculino e feminino, faixa etária, etc.), a realização de entrevistas e a compilação dos dados levantados (etapa automatizada para se calcular os resultados finais da pesquisa) [26]. No Brasil, por exemplo, um dos institutos de opinião pública, o Ibope, precisou de sete dias para coletar, analisar e divulgar as intenções de voto dos eleitores brasileiros, em relação aos dois principais presidenciáveis das eleições no ano de 2018 [87]. Além do tempo, o custo é outro fator que deve ser levado em consideração. Por exemplo, uma pesquisa no Brasil sobre intenções de voto para governador chegou próximo a 100 mil reais [83].

Após o sucesso da campanha eleitoral de Barack Obama no *Twitter*, no ano de 2008, e devido ao grande volume de informações produzidas nas diversas mídias sociais, como *Facebook* e *Twitter*, principalmente durante períodos que antecedem grandes eventos, como uma eleição para escolha do presidente de uma nação, a análise de mensagens com conteúdo político tem sido realizada como forma de ampliar as pesquisas nesse campo sob diversas perspectivas [25, 75, 95, 88, 125]. Por exemplo, a predição do resultado de eleições

é uma área que tem sido bastante explorada por pesquisadores [14, 15, 18, 19, 34, 56, 86, 88, 95, 102, 125, 138]. Segundo [102], o estudo das reações de eleitores em mídias sociais durante eventos, como debates eleitorais, tem sido uma outra área de pesquisa utilizando o *Twitter* no cenário de eleições. Já no trabalho proposto em [25], o *Twitter* foi investigado como uma ferramenta aliada ao processo de divulgação de propostas políticas.

O *Twitter*, em particular, é uma mídia social que tem crescido significativamente nos últimos anos. Entre o primeiro trimestre do ano de 2010 e o terceiro trimestre de 2017, o número de contas ativas no *Twitter* subiu de 10 para 330 milhões [117]. No segundo semestre de 2016, período em que ocorreu a eleição para a escolha do presidente da República dos Estados Unidos da América (EUA), esse país possuía o maior número de contas ativas de usuários do *Twitter* no mundo, com quase 70 milhões [116]. Esse serviço de *microbloging*, lançado em 2006, tem como principal característica permitir aos seus usuários publicarem mensagens curtas de até 280 caracteres, chamados de *tweets*. Nessas mensagens, é possível fazer menções a usuários, incorporar endereços eletrônicos de páginas *web* e utilizar *hashtags*.

No *Twitter*, várias terminologias são utilizadas nas mensagens. Um exemplo de *tweet* coletado pelos autores deste trabalho[1], a partir do uso da *Search API*[2] do *Twitter*, durante o período de campanha eleitoral para a escolha do presidente do Brasil no ano de 2014, é "RT @USER: CANDIDATO1 assegura que descontos da conta de luz permaneçam para o consumidor http://t.co/ZiZepua46e #HASHTAG". Nesse *tweet*, o símbolo "RT" representa um *tweet* retransmitido (*retweeted*) do usuário @USER (nesse caso, o símbolo @ + palavra(s) é utilizado para representar um usuário do *Twitter*), *"http"* corresponde a um endereço eletrônico, utilizado geralmente para complementar a informação de um *tweet*, e o símbolo "#" utilizado para representar uma *hashtag*, ou seja, para marcar um tópico de interesse ou um público específico.

Nos trabalhos que utilizaram o *Twitter* como fonte de dados no cenário eleitoral, a análise de sentimentos foi uma das abordagens utilizadas para classificar a opinião expressa pelo usuário no *tweet*. Por exemplo, na área de predição do resultado de eleições essa abordagem tem sido utilizada na etapa de cálculo de predição com o objetivo de classificar o sentimento contido nas mensagens, por exemplo em positivo ou negativo, sobre os candidatos [95, 102]. Para analisar o sentimento de *tweets* no cenário eleitoral,

[1] O texto original do *tweet* foi parcialmente alterado. A identificação do autor do *tweet*, o nome do candidato e o conteúdo da *hashtag* foram alterados para @USER, CANDIDATO1 e #HASHTAG, respectivamente.

[2] Conjunto de rotinas do *Twitter* responsável por retornar um conjunto relevante de *tweets*. Disponível em: https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html

duas abordagens são utilizadas com frequência, o uso de métodos baseados em aprendizado de máquina supervisionado e métodos baseados em dicionários. Os métodos que compreendem a primeira abordagem, são caracterizados por fazerem uso de algoritmos de aprendizado supervisionado, tais como *Naive Bayes* (NB), Multinomial Naive Bayes (MNB) e *Support Vector Machine* (SVM). Já a segunda abordagem, é caracterizada por fazer uso de dicionários léxicos, como o *SentiWordNet* [7] e o *General Inquirer* [119]. Num estudo realizado em [17], os autores argumentaram que a abordagem de aprendizado de máquina supervisionado produziu resultados mais precisos do que a segunda abordagem. Segundo [102], o método utilizado com mais frequência, entre os trabalhos que utilizaram a abordagem de análise de sentimentos, nos modelos propostos para predizer o resultado de eleições, foram os que fizeram uso da abordagem supervisionada.

1.1 Definição do problema

Nesse campo de pesquisa, onde informações contidas em *tweets* políticos têm sido utilizadas como a principal fonte de dados na análise de sentimentos, pesquisadores têm encontrado diversos desafios, tais como a quantidade limitada de caracteres permitidos em um *tweet*, o uso expressivo de *hashtags* contendo termos, como *slogans* de campanha e nomes de candidatos utilizados durante períodos de eleições para expressar opinião sobre candidatos [98], dificuldades em analisar textos contendo expressões que surgem durante períodos de eleições, presença de gírias, repetições de caracteres e abreviações de palavras [67, 99, 113]. No trabalho reportado em [72], os autores afirmaram que no domínio político há outra tarefa desafiadora que torna o processo de deteção do sentimento de mensagens nesse cenário ainda mais difícil, que é a necessidade de conhecer a afiliação política do candidato. Em um estudo realizado em [98], os autores chegaram a conclusão que usuários do *Twitter*, durante períodos de eleições, expressaram opinião política não somente a partir de *tweets*, mas também a partir de informações localizadas em seus perfis, tornado a tarefa de análise de sentimentos desse tipo de mensagem ainda mais desafiadora. Outra questão que tem sido investigada por pesquisadores nesse cenário, refere-se a informações falsas (*fake news*) divulgadas sobre candidatos durante períodos de campanha eleitoral, conforme reportado em [111, 4].

Em relação aos algoritmos utilizados para classificação do sentimento de *tweets* políticos, outro problema enfrentado pelos pesquisadores refere-se ao fato deles possuírem uma abordagem mais generalista, isto é, não são projetados, em sua maioria, para atenderem especificamente o cenário político e, como consequência, em muitos casos, acabam ob-

tendo um baixo desempenho [105]. Uma constatação desse fato pode ser encontrado em [29], onde segundo os autores mais de 7.000 trabalhos já foram publicados sobre análise de sentimentos. Esses esforços têm se concentrado justamente na investigação de métodos para produzirem melhores acurácias em cenários específicos (revisões de filmes, avaliações de produtos, etc).

A quantidade limitada de caracteres permitidos em *tweets* e o grande volume de informações divulgadas sobre políticos na *web*, durante períodos de campanha eleitoral, têm impulsionado usuários a expressarem opiniões políticas, a partir de *hashtags* e endereços eletrônicos, com uma frequência cada vez maiores. Esse comportamento contribui para tornar a tarefa de análise de sentimentos de *tweets* no cenário político ainda mais complexa. No caso de *hashtags*, alguns fatores contribuem para isso, como o surgimento de novos termos, criados durante períodos de campanha eleitoral, utilizados para fazer referência implítica ou explícita a candidatos, e a combinação/concatenação de várias palavras para definição de uma única *hashtag* [98]. Com relação a endereços eletrônicos, o problema consiste em identificar a opinião expressa por determinado usuário, a partir de informações localizadas externamente aos *tweets*. No trabalho reportado em [99], os autores investigaram o impacto de informações contidas em endereços eletrônicos no desempenho de algoritmos de aprendizado de máquina.

Iniciativas já foram realizadas com o objetivo de propor soluções para redução dos problemas supracitados. Entre as mais frequentes destacam-se a variação de atributos na representação dos *tweets* aos classificadores [3, 82], a utilização de diferentes algoritmos de aprendizado de máquina [59, 89], a incorporação de aspectos linguísticos [24, 33] e a variação de técnicas utilizadas na fase de pré-processamento [30, 31]. Nos trabalhos que utilizaram a abordagem de análise de sentimentos o conteúdo disponível em *tweets* públicos dos usuários foi a principal fonte de informação utilizada nas análises [34, 88, 95, 125]. Outras informações disponibilizadas nos perfis dos usuários, tais como o sexo [13], a idade [102, 103] e a localização [100], já foram utilizadas em alguns trabalhos, porém para coletar mensagens.

No cenário eleitoral, a representatividade do uso de *hashtags* em *tweets* tem se tornado cada vez mais relevante. No estudo proposto em [98], os autores investigaram a contribuição de *hashtags*, disponíveis em *tweets* políticos e em descrições de perfis de usuários, sob várias perspectivas, e concluíram que a representatividade de mensagens contendo *hashtags* é bastante expressiva (23%) no cenário de eleições. Resultado semelhante a esse foi reportado em [70], onde foi identificado que 26% das mensagens da base de dados,

utilizadas nas análises, continham pelo menos uma *hashtag*. Em [96, 97], os autores mostraram que informações contidas em descrições de perfis de usuários, coletadas durante períodos de campanha eleitoral, foram utilizadas com eficiência no processo de rotulação semiautomática de *tweets*. No domínio político, são encontrados com frequência trabalhos reportando o uso de *hashtags* principalmente para coletar e selecionar *tweets* [12, 13, 102]. Nesse cenário, outros estudos têm utilizado *hashtags* diretamente na fase de cálculo da predição, a partir da decomposição das palavras que compõem as *hashtags* [31, 113]. Já em outros trabalhos, *hashtags* são excluídas das análises, na etapa de limpeza dos dados [15, 20, 95, 102]. E, em outras pesquisas, como em [60, 79, 113], *hashtags* são utilizadas para criação automática de um *corpus* de mensagens rotuladas.

Diante dos problemas supracitados, nesta tese a seguinte questão de pesquisa é investigada: *hashtags*, disponíveis em *tweets* e em descrições de perfis de usuários, postados durante períodos de campanha eleitoral presidencial, melhoram a acurácia de algoritmos de aprendizado de máquina supervisionados, utilizados para classificar o sentimento de mensagens no cenário eleitoral?

1.2 Objetivos

Neste trabalho, *hashtags*, contidas em *tweets* políticos e em descrições de perfis de usuários, são analisadas sob diversas perpectivas, com o objetivo principal de investigar a contribuição delas na análise de sentimentos de mensagens no cenário eleitoral.

Um dos objetivos específicos consiste em realizar um estudo para investigar a relevância de *hashtags*, contidas em *tweets* e em descrições de perfis de usuários, no cenário de eleições. Um segundo objetivo, consiste em investigar a contribuição de *hashtags* contendo expressões relativas a candidatos, identificadas durante períodos de campanha eleitoral, na classificação do sentimento de *tweets* políticos. É investigado também, como atributos baseados em *hashtags* políticas podem ser levados em consideração para melhorar a acurácia de algoritmos de aprendizado de máquina supervisionado. Um quarto objetivo, consiste em investigar a contribuição de informações contidas em descrições de perfis de usuários no processo de análise de sentimentos de *tweets* no cenário eleitoral. Por último, é realizada uma investigação para verificar a contribuição individual e em conjunto de *hashtags*, contidas em *tweets* e em descrições de perfis de usuários, no processo de mineração de opiniões de *tweets* no cenário de eleições.

1.3 Metodologia

Com a finalidade de responder a questão de pesquisa proposta neste trabalho, primeiro foi realizada uma pesquisa bibliográfica nas principais bases científicas, com o objetivo de compreender o domínio investigado. Foram investigados trabalhos na literatura que haviam feito uso de *hashtags* no domínio político. Em seguida, foi proposto um estudo para avaliar a relevância de *hashtags*, contidas em *tweets* e em descrições de perfis de usuários, em cenários de eleições. Logo após, foi proposto um modelo para avaliar a contribuição de *hashtags*, contidas em *tweets* e em descrições de perfis, na melhoria do desempenho de classificadores, utilizados para analisar o sentimento de *tweets* em cenários de eleições. Por último, foi analisado o desempenho do modelo proposto, a partir de *datasets* coletados durante períodos de eleições presidenciais.

O modelo proposto nesta tese, é composto pelas seguintes etapas: coleta de dados, pré-processamento e criação e avaliação do modelo. Na fase de coleta de dados, *tweets* relacionados ao processo eleitoral são coletados e um conjunto de *datasets* é obtido. Na etapa de pré-processamento, diversas técnicas para limpeza de *tweets*, tais como remoção de números e endereços eletrônicos, são utilizadas. Neste trabalho, é realizado um experimento para investigar a contribuição das técnicas de remoção de *stopwords* e *stemming*, na análise de sentimentos das mensagens. Na última etapa, é analisada a contribuição de *hashtags*, contidas em *tweets* e em descrições de perfis de usuários, sob várias perspectivas, no processo de análise de sentimentos de *tweets* políticos.

Neste trabalho, *tweets* postados sobre os dois principais presidenciáveis do Brasil e dos EUA, nos anos de 2014 e 2016, respectivamente, foram coletados, juntamente com as informações contidas nas descrições dos perfis dos usuários. Aproximadamente 4.000 *tweets* de cada base de dados, foram selecionados aleatoriamente e rotulados em três classes distintas, por um conjunto composto por 16 indivíduos, em média. Para avaliar a contribuição de *hashtags* no cenário de eleições, foram utilizados os algoritmos de aprendizado de máquina supervisionado *Naive Bayes* (NB), *Multinomial Naive Bayes* (MNB) e *Support Vector Machine* (SVM). A medida acurácia foi utilizada para avaliar o desempenho dos classificadores. Em cada um destes experimentos, foram utilizados testes estatísticos com o objetivo de verificar se os incrementos obtidos nas acurácias dos classificadores, a partir do uso de *hashags*, foram estatisticamente significantes.

1.4 Contribuições

As principais contribuições desta tese são:

- um modelo para avaliar a influência de *hashtags*, disponíveis tanto em *tweets* quanto em descrições de perfis de usuários, na melhoria da acurácia de algoritmos de aprendizado de máquina supervisionado, utilizados para classificar o sentimento de *tweets* no cenário eleitoral; e
- análise da relevância de atributos baseados em *hashtags* políticas no processo de análise de sentimentos de *tweets* no cenário eleitoral.

1.5 Organização

Esta tese está organizada da seguinte forma:

- Capítulo 2 - Mineração de Opiniões. Neste capítulo, são apresentados os principais conceitos da área de mineração de opiniões e que são utilizados no desenvolvimento deste trabalho. Primeiro, são apresentadas algumas definições básicas dessa área. Segundo, são apresentadas duas abordagens comumente utilizadas na literatura para classificação do sentimento de mensagens. Terceiro, são apresentadas técnicas utilizadas com frequência na fase de pré-processamento. Por último, são detalhados os três algoritmos de aprendizado de máquina supervisionado utilizados nos experimentos propostos neste trabalho e as medidas utilizadas com frequência para avaliá-los;
- Capítulo 3 - *Hashtag* no cenário eleitoral. Neste capítulo, são apresentadas algumas abordagens de uso do *Twitter* no domínio político e uma relação de trabalhos que reportaram o uso de *hashtags* em cenário de eleições;
- Capítulo 4 - Materiais e Métodos. Neste capítulo, é apresentado um modelo proposto para avaliar a contribuição de *hashtags*, contidas em *tweets* e em descrições de perfis, na análise de sentimentos de mensagens no cenário eleitoral;
- Capítulo 5 - Relevância de *hashtag* em cenários de eleições. Neste capítulo, é investigada a relevância de *hashtags*, contidas em *tweets* e em descrições de perfis de usuários, a partir da análise de duas amostras presidenciais.

- Capítulo 6 - Análise Experimental. Neste capítulo, o modelo proposto é avaliado, a partir de um conjunto composto por quatro experimentos computacionais.

- Capítulo 7 - Conclusões. Neste capítulo, são apresentadas as principais conclusões obtidas nesta tese, ressaltando as contribuições obtidas, as limitações deste trabalho, bem como os encaminhamentos de trabalhos futuros.

Capítulo 2

Mineração de Opiniões

Neste capítulo, são apresentados conceitos da área de mineração de opiniões e as principais estratégias utilizadas nessa área para classificação do sentimento de mensagens curtas, como *tweets*. A Seção 2.1 apresenta conceitos dessa área e a terminologia utilizada no processo de classificação de opiniões expressas em textos. A Seção 2.2 apresenta as abordagens comumente encontradas na literatura para classificação do sentimento de mensagens. A Seção 2.3 apresenta técnicas de pré-processamento textual. A Seção 2.4 apresenta os principais algoritmos de aprendizado de máquina utilizados na análise de sentimentos. Por último, a Seção 2.5 aborda medidas utilizadas para avaliação de algoritmos de classificação.

2.1 Definições principais

Mineração de Opiniões ou Análise de Sentimentos (AS) [65] é um campo de pesquisa na área de Processamento de Linguagem Natural (PLN) que analisa as opiniões de indivíduos, avaliações, sentimentos, atitudes e emoções por meio do tratamento computacional da subjetividade no texto [105]. Segundo [124], a análise de sentimento de um texto tem, basicamente, dois objetivos: (a) analisar e diferenciar um texto entre objetivo (não expressa opinião) e subjetivo (que expressa opiniões/sentimentos) e (b) caso seja subjetivo, classificá-lo quanto a sua polaridade, por exemplo em positivo ou negativo.

Uma parte importante no processo de coleta de informações textuais consiste em descobrir o que indivíduos pensam a respeito de produtos, políticos, esportes, notícias em geral, etc., ou seja, em descobrir, por exemplo, o sentimento (positivo ou negativo) sobre tais aspectos.

A expressão do sentimento de um indivíduo sobre algo, na forma textual, consiste basicamente na identificação do (a) alvo da opinião e (b) na descoberta do sentimento sobre esse alvo [65, 124].

Entidade [65] ou tópico [124] são outras expressões encontradas na literatura para fazer referência a alvos contidos em documentos. O alvo pode se referir a uma pessoa, objeto, produto, serviço ou qualquer outro elemento por meio do qual o indivíduo expressa suas emoções (positivas ou negativas). Segundo [124], um documento textual é qualquer fragmento de texto em linguagem natural. Por exemplo, na sentença "Fulano é o melhor candidato à Presidência da República do Brasil", um sentimento positivo expresso por meio da palavra "melhor" está sendo expresso a entidade "Fulano". A área de análise de sentimentos consiste em identificar o sentimento expresso por um indivíduo, por meio de um documento, sobre uma determinada entidade.

A classificação do sentimento de uma unidade de informação (texto, parágrafo, frase, etc.) pode variar em relação à granularidade do documento analisado. Documentos com granularidades menores exigem classificações mais específicas. Em relação ao tamanho das unidades de informação, a classificação pode ser realizada em nível de:

- **Documento (*document level sentiment*):** nesse nível, a análise produz como resultado uma opinião positiva ou negativa para o documento como um todo. Segundo [67], nesse nível a classificação de sentimento assume que um documento opinativo "D", por exemplo um documento contendo revisão de um produto, expressa opiniões sobre uma única entidade conhecida "E", e essas opiniões são provenientes de um único indivíduo "H". Por esse motivo, a identificação da entidade é irrelevante e o sentimento é analisado quanto ao aspecto geral da entidade em questão [66].

- **Sentença (*sentence level sentiment*):** Um documento "D" pode ser composto por uma ou mais frases "F". Naturalmente, a mesma técnica de classificação de sentimento em nível de documento pode ser aplicada a frases individuais [67]. A tarefa de classificação de uma frase em subjetiva ou objetiva é frequentemente conhecida como classificação de subjetividade [45, 106, 132, 135, 136, 139]. O processo para classificação da polaridade de uma frase consiste em duas etapas: (1) classificação de subjetividade: determina se "F" é uma frase subjetiva ou objetiva e (2) classificação do sentimento em nível de sentença: se "F" é subjetiva, determina se expressa uma opinião positiva, negativa ou neutra. Segundo [67], muitas pesquisas têm assumido a suposição que uma frase expressa uma única opinião a respeito de um único indivíduo "H". Essa suposição é apropriada apenas para frases simples com uma

única opinião, por exemplo em "Fulano foi o melhor governador de Minas Gerais". O sentimento positivo ("melhor") foi atribuído à entidade "Fulano". Porém, em frases mais complexas, podem aparecer diversas opiniões sobre várias entidades, por exemplo em "Candidato1 foi o pior governador de MG e a Candidata2 foi a melhor presidente e mais eficaz chefe da casa civil que o Brasil já teve". Nesse exemplo, a palavra "pior" é utilizada para expressar um sentimento negativo em relação a entidade "Candidato1" e as palavras "melhor" e "eficaz" são utilizadas para expressar um sentimento positivo em relação a entidade "Candidato2".

- **Entidades e Aspectos (*aspect level*):** A classificação do sentimento de textos em nível de documento ou sentença é útil em muitos casos, porém não é capaz de oferecer detalhes em cenários específicos. Um documento ou uma frase contendo informações sobre determinada entidade em particular pode expressar uma opinão positiva em determinado momento e negativa em outro. Por exemplo, na frase "O candidato X recuperou várias rodovias no estado de Minas Gerais, porém acabou com a educação", a entidade "X" é julgada positivamente por meio da sentença "recuperou várias rodovias" e negativamente por meio da expressão "acabou com a educação". Um documento contendo uma opinião positiva sobre uma entidade em particular não significa que o autor tenha opiniões positivas sobre todos os aspectos da entidade. Segundo [67], um documento contendo uma opinião negativa não significa que o autor não goste de tudo. Em um documento de opiniões típico, o autor escreve aspectos positivos e negativos da entidade, embora o sentimento geral sobre a entidade possa ser positivo ou negativo. A classificação do sentimento nesse caso é possível a nível de aspecto [67].

Neste trabalho, o alvo a ser considerado nas mensagens será composto pelo conjunto contendo os nomes dos presidenciáveis, assim como em outros trabalhos [31, 95]. A abordagem de classificação do sentimento de *tweets* políticos utilizada nesta tese será realizada em nível de sentença e será assumido que cada frase ou conjunto de frases presentes em determinado *tweet* conterá apenas uma única opinião sobre certa entidade. Na fase de mineração de opiniões dos *tweets*, é utilizada a abordagem ternária para treinar e executar os classificadores, assim como em outros trabalhos [1, 2, 37, 59, 60, 89, 92, 105].

2.2 Métodos para mineração de opiniões

Os métodos para mineração de opiniões de textos podem ser classificados em duas classes principais: (a) métodos baseados em aprendizado de máquina (abordagem supervisionada) e (b) métodos léxicos ou baseados em dicionários.

2.2.1 Métodos baseados em aprendizado de máquina

Os métodos que compreendem essa categoria são caracterizados por fazerem uso de algoritmos supervisionados e de um *corpus* de treinamento. Esses algoritmos utilizam sumariamente duas fases:

- **Fase de Treinamento:** nessa fase, um algoritmo de aprendizado de máquina supervisionado, por exemplo *Naive Bayes* (NB) [55], *Support Vector Machine* (SVM) [46] e *Multinomial Naive Bayes* (MNB) [76], é utilizado para treinar um modelo de classificação utilizando um conjunto de atributos extraídos de documentos da base de treinamento (*corpus* de treinamento). Os dados de treinamento consistem em pares de entrada de dados e a saída desejada (classe ou rótulo) correspondente [67].
- **Fase de Inferência:** nessa fase, o modelo de classificação obtido na etapa anterior é utilizado para classificar o sentimento de novos exemplos, isto é, distintos daqueles que foram utilizados na composição do *corpus* da fase de treinamento [107].

O trabalho prógono a utilizar algoritmos de aprendizado de máquina supervisionado na análise de sentimento de texto foi apresentado em [94]. Os autores desse trabalho utilizaram os algoritmos SVM, NB e *Maximum Entropy* (MaxEnt) para calcular a polaridade (positiva/negativa) de opiniões de usuários expressas nas avaliações de filmes retirados da *Web* [94]. Nesse trabalho, a abordagem de aprendizado de máquina obteve um desempenho melhor ao do *baseline* utilizado. Dos três algoritmos utilizados, o SVM foi o que apresentou a melhor acurácia (82,9%), utilizando o formato unigrama na representação do atributo. Outros formatos também foram testados, como unigramas e bigramas e somente bigramas.

Na abordagem de aprendizado de máquina supervisionado, a quantidade e a variação dos atributos utilizados para representação das informações têm sido um fator responsável pela melhoria do desempenho de algoritmos. Por isso, muitos pesquisadores têm investigado a melhor combinação em cenários específicos. Em [82], os pesquisadores examinaram o desempenho de atributos obtidos de fontes diversas e atributos baseados em

um tópico, ou seja, foi considerado se uma frase citava o tópico em discussão. Além disso, foram utilizados como atributos valores de *PMI* (*Pointwise Mutual Information*) [126]. A mesma base de dados utilizada em [94] foi usada em [82], com o objetivo de comparação dos resultados. Em [82], os autores conseguiram melhorar o desempenho do classificador SVM para 86% de acurácia, um acréscimo de 3,1% em relação ao melhor resultado obtido em [94]. A partir do estudo proposto em [82], conclui-se que a variação de atributos melhoram efetivamente o desempenho de algoritmos de aprendizado de máquina.

Outro trabalho onde é possível concluir que a variação de atributos é capaz de melhorar a acurácia de algoritmos de aprendizado de máquina pode ser encontrado em [3]. Nesse trabalho, os autores investigaram os benefícios em utilizar *hashtags* para determinar a polaridade de *tweets* no domínio político. Para classificar a polaridade do sentimento das mensagens, os autores utilizaram três atributos distintos: número de *hashtags* positivas e negativas, número de palavras positivas e negativas e unigrama. Utilizando os algoritmos NB, SVM, *Logistic Regression* (LR) e *Random Forrest* (RF), os autores concluíram que atributos baseados em *hashtags* foram capazes de melhorar as acurácias dos cinco algoritmos. Em um dos experimentos propostos, onde *hashtags* foram utilizadas para formação do atributo, os autores conseguiram obter 96% de acurácia utilizando os algoritmos NB, SVM e LR. Nesse mesmo experimento, porém utilizando o formato unigrama para formação do atributo, as acurácias obtidas pelos mesmos algoritmos foram de 90,5%, 91,5% e 94,5%, respectivamente.

Outras pesquisas foram realizadas com o objetivo de investigar a melhoria da qualidade do processo de classificação do sentimento de textos não somente a partir da variação de atributos, mas também a partir da utilização de diferentes algoritmos de aprendizado de máquina e na incorporação de aspectos linguísticos para capturar o sentimento expresso nos textos [24, 33, 82]. Resultados apresentados em [33], mostraram, por exemplo que a melhor configuração obtida é a partir da utilização do algoritmo SVM combinando n-grama e aspectos linguísticos. Embora tenha sido mostrado em [33] e [82] que aspectos linguísticos favorecem a classificação, em [24] os melhores resultados foram obtidos apenas a partir do uso de n-gramas.

Em vários outros trabalhos o uso da abordagem de aprendizado de máquina, a partir da utilização de algoritmos supervisionados, tem sido investigado. Em [92], foram utilizados nos experimentos os algoritmos MNB, SVM e *Conditional Random Field* (CRF). Em [2], além do MNB foi utilizado nas análises o classificador *Decision Tree* (DT). No trabalho proposto em [112] foram utilizados os algoritmos de classificação NB, DT e

SVM. Já em [93], os autores utilizaram cinco algoritmos nos experimentos: *SVM*, *NB*, *Voted Perceptron* (VP), *Adaptative Boosting* (AB) e *Bayesian Logistic Regression* (BLR). Nos resultados reportados em [1, 9, 53], o classificador SVM foi utilizado em todos os experimentos. Em [59], foram utilizados os algoritmos *SVM* e *NB*, enquanto em [89] foram utilizados os algoritmos *SVM* e *MNB*.

No cenário político, o processo de classificação do sentimento de textos a nível de sentença, utilizando algoritmos de aprendizado de máquina, tem sido investigado por pesquisadores que se propuseram a criar modelos para predizer o resultado de eleições presidenciais, a partir da análise de informações contidas em *tweets* publicados durante períodos de campanha eleitoral [10, 14, 15, 19, 31, 71, 77, 86, 95].

Em [10], os autores propuseram um modelo de sentimento político, a fim de capturar as intenções de voto de usuários do *Twitter* que haviam postado mensagens durante o período de campanha eleitoral para as eleições gerais da Irlanda. Os autores utilizaram a técnica de tokenização das mensagens em unigramas e preservaram *emoticons* e pontuações não convencionais (Ex: "!!!!"). Termos, como menções a usuários e *URL's*, foram removidos na etapa de pré-processamento. Como resultado final, o classificador *Adaboost* MNB foi o que apresentou a melhor acurácia (65,09%), utilizando a técnica de validação cruzada *k-fold* (k=10), seguido dos algoritmos *SVM* (64,82%), *ADA-SVM* (64,28%) e MNB (62,94%). Nesse trabalho, não foram discutidas as melhores combinações para a escolha dos atributos, e também não foi analisada a contribuição do uso de técnicas de análise de recursos linguísticos.

Em [95], o classificador *Naive Bayes* foi utilizado para classificar a polaridade de *tweets* postados durante o período de realização das eleições presidenciais ocorridas no Brasil no ano de 2014. Utilizando esse algoritmo, o modelo não conseguiu prever corretamente o resultado das eleições. Nesse trabalho, não foi realizado um estudo comparativo com outros algoritmos supervisionados, tão pouco foram exploradas outras variações na representação dos atributos, além do formato unigrama utilizado, e também não foi investigada a contribuição de outras técnicas como *Part-Of-Speech* (POS) no processo de melhoria da classificação do sentimento das mensagens.

Em [31], diversas técnicas de pré-processamento foram aplicadas aos *tweets* antes de analisar o sentimento das mensagens, tais como normalização de palavras (Ex: hahaha foi normalizado para haha), mapeamento de símbolos *unicode* e *emoticons* para palavras e a decomposição de *hashtags*. Esse último procedimento foi realizado a partir da utilização de algoritmo de programação dinâmica [110], baseado nos dados *one-gram* do Google

para segmentar as *hashtags* em suas palavras constituintes. Usando esse algoritmo, por exemplo as *hashtags* "#votewisely" e "#teamgej" são mapeadas para as características unigrama "vote" e "wisely" e "team" e "gej", respectivamente.

O uso da abordagem de aprendizado de máquina para análise do sentimento de texto em cenários específicos, como no domínio político, é um campo que ainda carece de investigações. Assim como em outros cenários, no domínio político os pesquisadores têm investigado como obter melhores resultados na classificação do sentimento de mensagens, a partir de uma série de iniciativas, como variar o conjunto de classificadores, investigar formas variadas de representação dos atributos e testar técnicas de pré-processamento. Uma constatação desse fato é reportado em [15], onde segundo os autores mais de 7.000 trabalhos já foram publicados sobre análise de sentimentos. Segundos os autores, esses esforços têm se concentrado justamente na investigação de métodos para produzirem melhores acurácias em cenários específicos (revisões de filmes, avaliações de produtos, etc.).

2.2.2 Métodos baseados em dicionários

A classificação do sentimento de texto utilizando métodos baseados em dicionário, também conhecida como abordagem Léxica, é caracterizada pelo cálculo semântico das palavras que compõem o texto, a partir do uso de um dicionário de sentimento. Orientação semântica é uma medida de subjetividade que geralmente captura um fator de avaliação (positivo ou negativo) e a potência ou força (grau em que a palavra, frase ou documento em questão é positivo ou negativo) em relação a um assunto, pessoa ou ideia [90].

A polaridade do sentimento de um texto, utilizando essa abordagem, é calculada a partir de informações contidas em um ou mais dicionários léxicos. Em [113], os autores combinaram vários dicionários léxicos, a fim de analisar a polaridade do sentimento de *tweets* postados durante o período de realização da copa do mundo de futebol de campo, ocorrida no Brasil no ano de 2014.

Em um dicionário típico, além das palavras que o compõe há outras informações como a orientação semântica de cada palavra. Dessa forma, o sentimento geral de um texto pode ser calculado pela média dos valores correspondentes ao sentimento individual de cada palavra do texto extraído do dicionário.

Dicionários podem ser construídos manualmente [119, 123] ou automaticamente, a partir da utilização de palavras sementes (*seed words*) utilizadas para expandir uma lista

de palavras já existente [44, 126, 127]. O processo de construção de dicionários automáticos consiste, basicamente, na utilização de um conjunto de *seed words*, cujas orientações semânticas (positiva ou negativa) são definidas inicialmente de forma manual, para geração de uma lista maior de palavras. Nesse caso, utiliza-se um algoritmo para pesquisar em dicionários *online* os sinônimos e antônimos de todas as *seeds*. As novas palavras identificadas são armazenadas na lista de *seeds*. A interação continua até não haver mais palavras a serem pesquisadas [66]. Um exemplo de uso desse tipo de abordagem foi encontrado em [49].

Na literatura, vários dicionários léxicos, tais como *General Inquirer* [119], *SentiWordNet* [7], *Senti-Strength* [122], MPQA *Subjectivity Lexicon* [134], *Multi-Perspective Question Answering* (MPQA), *Opinion Corpus* [134], Bing Liu's *opinion lexicon* [66], NRC-*emoticon* [81], NRC-*hashtag* [80] e *Opinion-Finder* [134], têm sido utilizados na classificação do sentimento de textos.

No domínio político, o processo de classificação do sentimento de textos, a nível de sentença, utilizando a abordagem léxica, também tem sido investigado por pesquisadores que se propuseram a criar modelos para predizer o resultado de eleições presidenciais, a partir da análise das informações contidas em *tweets* publicados durante períodos de campanha eleitoral [14, 18, 20, 36, 86, 88, 138].

Em [88], os autores utilizaram o dicionário *OpinionFinder*, a fim de classificar a polaridade (positivo ou negativo) de mensagens postadas no *Twitter*, entre os anos de 2008 a 2009, sobre o norte-americano Barack Obama, em dois contextos distintos. No primeiro ano, para avaliar as intenções de voto dos eleitores durante a campanha eleitoral presidencial, e no segundo para avaliar a aprovação do governo de Obama em 2009. Nesse trabalho, um *tweet* foi definido como positivo se o total de palavras positivas fosse superior ao número de palavras negativas, e vice-versa. Segundo os autores, considerando uma janela de tempo de 15 dias o modelo proposto foi capaz de apresentar uma correlação (r) superior a 79%, em relação a pesquisas de opinião pública realizadas por institutos convencionais.

Em [86], os autores compararam o sentimento de mensagens que prevaleceram antes e depois das eleições presidenciais ocorridas nos EUA e na França no ano de 2012. O método proposto para analisar a aprovação pública dos usuários ("eleitores") do *Twitter* sobre os presidenciáveis nos EUA, consitiu em analisar as séries temporais dos dois candidatos utilizados nas análises, Barack Obama e Mitt Romney. As séries foram construídas a partir da análise do sentimento das mensagens utilizando as abordagens de aprendizado

de máquina e dicionários léxicos. O sentimento de um *tweet* foi calculado a partir de três *scores*. O primeiro (*score.polarity*), utilizava o dicionário léxico proposto por [10]. O segundo (*score.sentiment*), usava o léxico de subjetividade de Janyce Wiebes [11] para treinar o classificador NB, e o último (*score.afinn*), utilizava a lista de palavras AFINN [12]. Por exemplo, uma mensagem era classificada com o sentimento positivo, quando o resultado de dois ou mais *scores* calculavam a mesma polaridade, e neutro quando os três eram diferentes. Segundo os autores, resultados satisfatórios foram encontrados utilizando essa abordagem.

Alguns dicionários léxicos, como o *WordNet*, mesmo possuindo uma quantidade grande de palavras (> 155.280) [128], apresentam certas características que podem restringir o uso, como suporte a apenas único idioma e ausência de palavras para representar gírias e expressões comuns presentes a certos cenários, como o político, por exemplo. No *Twitter*, onde a quantidade de caracteres é restrita (até 280), e a presença de erros de ortografia e o uso de *URL's/hashtags* é frequente [98], soluções baseadas em dicionários podem não ser tão eficientes.

Em [59], os autores propuseram um estudo para comparar o desempenho das abordagens de aprendizado de máquina e léxico para classificação do sentimento de mensagens publicadas no *Twitter*. Algumas técnicas foram avaliadas, bem como formas de combiná-las. Os resultados mostraram que o método de aprendizado de máquina utilizando os algoritmos SVM e NB superaram os métodos baseados em abordagem léxica. No estudo realizado em [17], os autores argumentaram que a abordagem de aprendizado de máquina supervisionado foi capaz de produzir resultados mais precisos do que a abordagem baseada em dicionários. Segundo [102], o método utilizado com mais frequência entre os pesquisadores foram os que utilizavam a abordagem supervisionada. Neste trabalho, a abordagem de aprendizado de máquina supervisionada será utilizada nos experimentos propostos, assim como em outros trabalhos encontrados na literatura no domínio político [3, 10, 14, 31, 71, 77, 86, 95, 102].

2.3 Técnicas de pré-processamento

Nesta seção, são apresentadas técnicas de pré-processamento geralmente utilizadas no processo de classificação da polaridade do sentimento de textos.

Técnicas de pré-processamento são utilizadas por pesquisadores com o objetivo de diminuir o ruído textual da base de dados [102]. No *Twitter*, símbolos de pontuação, nú-

meros, *URL's*, *hashtags*, estruturas específicas do *Twitter*, tais como menções a usuários e símbolo de *retweet*, pronomes, artigos definidos e indefinidos, entre outros são incorporados com frequência nos *tweets* pelos usuários. Para alguns autores, essas informações têm sido consideradas indesejadas nas mensagens e por isso têm sido eliminadas na fase de limpeza dos dados [95, 102].

Alguns trabalhos já reportaram benefícios ao considerar a fase de pré-processamento no processo de classificação do sentimento de texto. Em [41], foi realizado um estudo sobre o papel do pré-processamento na mineração de opiniões, onde foi constatado a sua efetividade no desempenho da classificação. Outro trabalho que destaca a importância da utilização dessa etapa foi reportado em [1], onde foi apresentado pelos autores uma redução de mais de 38% do tamanho da base de dados, além do aumento no desempenho.

Apesar de haver trabalhos reportando benefícios, a partir do uso de técnicas de pré-processamento no processo de mineração de opiniões, há pesquisas que não fazem uso de nenhuma delas ou utilizam um conjunto limitado. Nos trabalhos onde elas são utilizadas não são apresentados estudos que comprovem a efetividade do uso delas no domínio em que são empregadas. Em [102], os autores fizeram um levantamento bibliográfico sobre trabalhos que se propuseram a predizer o resultado de eleições presidenciais, senatoriais e gerais, utilizando o *Twitter* como fonte de informação, e chegaram a conclusão que dos mais de 20 trabalhos investigados, sobre eleições realizadas em diversos países, como EUA, Alemanha, Singapura, Irlanda, Nova Zelândia, França, Itália, Venezuela, Equador e Nigéria, menos de 25% haviam filtrado os dados antes de prosseguirem para a etapa do cálculo da predição.

Dentre os métodos de pré-processamento geralmente utilizados na mineração de opiniões, destacam-se: tokenização, remoção de *URL's*, remoção de números e caracteres especiais, representação dos atributos para os classificadores, remoção de estruturas específicas do *Twitter*, remoção de *stopwords* e *stemming*. Essas técnicas são detalhadas nas próximas seções.

O *tweet* abaixo [1] coletado, pela *Search Application Programming Interface* API[2] do *Twitter* durante o período de campanha eleitoral presidencial no Brasil no ano de 2014, é utilizado para exemplificar o efeito da aplicação de cada técnica:

[1] *Tweet* coletado pelos autores deste trabalho. O texto original do *tweet* foi parcialmente alterado. A identificação do autor do *tweet*, o nome do candidato e a *hashtag* foram alterados para @USER, CANDIDATO1 e HASHTAG, respectivamente.

[2] Interface de Programação de Aplicativos do *Twitter* utilizada para coletar *tweets* públicos. Disponível em: https://developer.twitter.com/en/docs/tweets/search/overview/standard

Exemplo: *"<333 RT @USER: Voto CANDIDATO1 porque não penso só em mim! #HASHTAG http://t.co/AEFKJBmplG"*

2.3.1 Tokenização

Essa técnica consiste em decompor sentenças (frases) de um texto em partes menores, que podem ser termos ou palavras, referenciados como *tokens*. O processo de fragmentação é geralmente feito, a partir do uso de caracteres delimitadores, que podem ser uma vírgula, traço, etc. O tokenizador padrão de um *tweet* divide-o em palavras utilizando como delimitador o espaço em branco. Com o emprego dessa técnica, o texto passa a ser representado por um conjunto de palavras cuja nomeclatura utilizada é *Bag-Of-Words* (BOW). Em [10, 77, 80, 102], essa mesma técnica foi utilizada na fase de pré-processamento para limpeza das mensagens coletadas do *Twitter*.

Ao aplicar essa técnica de pré-processamento ao *tweet* usado como exemplo, resulta em:

Exemplo pré-processado: *{<333, RT, @USER:, Voto, CANDIDATO1, porque, não, penso, só, em, mim!, #HASHTAG, http://t.co/AEFKJBmplG}*

2.3.2 Remoção de *URL's*

URL's correspondem a endereços eletrônicos de páginas *web* incorporados aos *tweets*. Geralmente identificadas nas mensagens por "http://", elas têm se tornado uma alternativa para complementar as informações contidas em *tweets*, devido a quantidade limitada de caracteres permitidos nas mensagens. Em [99], os autores utilizaram três bases de dados distintas para examinar o conteúdo apontado por endereços eletrônicos, bem como o potencial impacto deles no desempenho de algoritmos de aprendizado de máquina utilizados na mineração de opiniões. Já em outros trabalhos, como em [10, 108], as *URL's* foram removidas dos *tweets* na fase de pré-processamento.

Ao aplicar essa técnica de pré-processamento ao *tweet* do exemplo, resulta em:

Exemplo pré-processado: *"<333 RT @USER: Voto CANDIDATO1 porque não penso só em mim! #HASHTAG"*

2.3.3 Remoção de números e caracteres especiais

Essa técnica consiste na remoção de números e caracteres especiais contidos em texto. Caracteres especiais consistem no conjunto composto pelos sinais de pontuação (ponto final, vírgula, ponto de interrogação, parênteses, etc.) e ortográficos (til, trema, apóstrofo, acento agudo, etc.). O emprego dessa técnica na fase de pré-processamento foi reportado em alguns trabalhos, como em [8, 80, 108].

Ao aplicar essa técnica de pré-processamento ao *tweet* do exemplo, resulta em:

Exemplo pré-processado: *"RT @USER Voto CANDIDATO1 porque não penso só em mim #HASHTAG http://t.co/AEFKJBmplG"*

2.3.4 Remoção de estruturas específicas do *Twitter*

As estruturas específicas do *Twitter* consistem basicamente nos símbolos "RT", "@" e "#", utilizados para representar *retweets*, fazer menções a usuários do *Twitter* e *hashtags*, respectivamente. Ao aplicar essa técnica de pré-processamento ao *tweet* do exemplo, resulta em:

Exemplo pré-processado: *"<333 Voto CANDIDATO1 porque não penso só em mim! http://t.co/AEFKJBmplG"*

2.3.5 Representação dos atributos para os classificadoress

Para realizar o treinamento de algoritmos de aprendizado de máquina, os *tweets* da base de treinamento precisam ser representados como atributos. Nessa representação, são contabilizadas a frequência (*Term-Frequency* (TF)), o inverso da frequência (*Term-Frequency - Inverse Document Frequency* (TF-IDF)) [73] ou presença/ausência dos *tokens* nos *tweets*. Para a utilização da frequência de um *token* (TF) como atributo, é considerado o número de vezes que ele aparece no *tweet*. Na segunda representação, é levada em consideração a importância de um *token* tanto no *tweet* quanto na coleção de mensagens, ou seja, o valor *TF-IDF* de um *token* aumenta proporcionalmente à medida que aumenta o número de ocorrências dele na mensagem, porém equilibrado pela frequência dele no *corpus*. Já na terceira representação, utiliza-se um valor binário (1/0) para indicar (presença/ausência) do termo no *tweet*.

Uma representação dos *tweets* comumente utilizada é a partir de n-gramas [10, 23, 80]. N-gramas são sequências contíguas de n *tokens* em um texto ou *tweet*. Quando o valor

de n é igual a 1 ($n = 1$) são chamados de unigramas, quando $n = 2$ de bigramas, quando $n = 3$ de trigramas, e assim sucessivamente. A Tabela 2.1 apresenta o *tweet* do exemplo tokenizado com $n = 1$ e $n = 2$.

Tabela 2.1: Representação do atributo nos formatos unigrama e bigrama para o *tweet* do exemplo

N-gram	Exemplo pré-processado
$n = 1$	<333, RT, @USER:, Voto, CANDIDATO1, porque, não, penso, só, em, mim!, #HASHTAG, http://t.co/AEFKJBmplG
$n = 2$	<333 RT, RT @USER:, @USER: Voto, Voto CANDIDATO1, CANDIDATO1 porque, porque não, não penso, penso só, só em, em mim!, mim! #HASHTAG, #HASHTAG http://t.co/AEFKJBmplG

2.3.6 Remoção de *stopwords*

Stopwords ou *stop words* consiste num conjunto de palavras consideradas mais gerais e com menos significado [67]. Essas palavras podem compreender o conjunto de preposições (Exs.: nas, com, de, etc), artigos definidos e indefinidos (Exs.: o, a, um, uma, etc.), conjunções (Exs.: mas, e, por, nas, etc.), etc.. Essa técnica consiste em remover de textos, como *tweets*, palavras que não contribuem para a semântica da mensagem no domínio analisado, a partir do uso de uma lista de *stopwords*. Uma variedade de trabalhos reportando o uso dessa técnica é comumente encontrado na literatura [1, 50, 59, 60, 68, 92, 93].

A Tabela 2.2 apresenta o resultado da aplicação da técnica de remoção de *stopwords*[3] do *tweet* usado como exemplo.

Tabela 2.2: Remoção de *stopwords* do *tweet* do exemplo

Tweet pré-processado	Tokens removidos
<333 RT @USER: Voto CANDIDATO1 porque penso mim! #HASHTAG http://t.co/AEFKJBmplG	não, só, em

2.3.7 *Stemming*

Stemming [51] é uma técnica que reduz palavras flexionadas (ou às vezes derivadas) para a sua forma base ou raiz [67]. Por exemplo, as palavras "assistir", "assistindo", "assistiu" são representadas como "assist"[4], tornando uma única palavra. Dessa forma, palavras

[3]Lista de *stopwords* utilizada: http://snowball.tartarus.org/algorithms/portuguese/stop.txt

[4]Foi utilizado o algoritmo de *stemming* Snowball . Disponível em: https://www.nltk.org/_modules/nltk/stem/snowball.html

com formas variantes podem ser representadas como o mesmo recurso. Diversos trabalhos reportando o uso dessa técnica na fase de pré-processamento são encontrados na literatura [8, 21, 27, 28, 53, 62, 64, 68, 74, 131].

Com a redução de palavras a um mesmo radical, obtém-se um único atributo para a representação de todas elas. Segundo [52, 91], essa redução de atributos pode melhorar o desempenho da classificação e diminuir o grau de esparsidade da base de dados. Esse índice mede o percentual de elementos nulos em uma base de dados. Uma maneira de se calculá-lo consiste na relação entre o volume de elementos nulos sobre o total de elementos. Reduzir o grau de esparsidade de uma base de dados significa ganhos em termos de consumo de memória e tempo de processamento da base de dados.

Os dois algoritmos de *stemming* investigados nesta tese são descritos a seguir:

- ***Lovins***: proposto por Julie Beth Lovins em 1968 [69], *Lovins* é considerado o pioneiro entre os algoritmos dessa categoria. Ele é composto por 294 sufixos, 29 condições e 34 regras de trasnformação e o seu funcionamento consiste basicamente em duas etapas. Na primeira, procura-se o sufixo da palavra na tabela de sufixos. Caso o sufixo da palavra seja encontrado, respeitando-se o conjunto de condições, ele é removido da palavra. No segundo passo, a palavra sem o sufixo é convertida, a partir de um conjunto de regras de transformação. Uma das vantagens desse algoritmo é a sua rapidez no processo de conversão, além de poder lidar com muitos plurais irregulares, como "*tooth*" e "*teeth*", "*matrix*" e "*matrices*", etc. [54].

- ***Snowball***: também chamado de *Porter*, esse algoritmo foi proposto no ano de 1980 e consiste no conceito de que sufixos são constituídos de sufixos menores e mais simples. *Snowball* é composto por cinco passos com diferentes regras de remoção dos sufixos, de tal forma que quando uma regra é aceita, o sufixo é removido e o próximo passo é executado com um novo conjunto de regras. Uma das vantagens desse algoritmo é a possibilidade de estender o seu *framework* com outros conjuntos de sufixos e idiomas [54]. A implementação desse algoritmo é facilmente encontrada em diversas plataformas, como a *Natural Language Tookit* (NLTK)[5]. Nessa plataforma, o algoritmo *Snowball* está implementado em mais de 10 idiomas diferentes, tais como o inglês, alemão, francês, português e espanhol.

[5]https://www.nltk.org/_modules/nltk/stem/snowball.html

2.4 Algoritmos de aprendizado de máquina

Nesta seção, são apresentados os algoritmos de aprendizado de máquina supervisionado utilizados nos experimentos propostos nesta tese e encontrados com mais frequência na mineração de opiniões de *tweets* políticos [3, 10, 14, 31, 71, 77, 86, 95, 102].

2.4.1 *Naive Bayes* (NB)

Naive Bayes [55] é um classificador probabilístico supervisionado que aplica o teorema de *Bayes* com suposições (ingênuas) de que os atributos são independentes entre si. No teorema de *Bayes*, a probabilidade de um documento d pertencer a uma classe c pode ser calculada usando a Equação 2.1 a seguir.

$$P(c|d) = \frac{P(d|c) \times P(c)}{P(d)} \tag{2.1}$$

$P(c)$ representa a probabilidade da classe c acontecer e é calculada pela relação entre a quantidade de documentos que pertencem à classe c e o total de documentos na base de dados. $P(d|c)$ é calculada como o produto das probabilidades de ocorrência de cada valor de atributo t_i pertencente a classe c na base de treinamento. A probabilidade de $P(d)$ é constante para todas as classes, e por isso somente as duas primeiras probabilidades precisam ser calculadas. Assume-se que cada documento é representado como um vetor $d = (t1, t2, ..., t_{|d|})$, onde t_i é o i-ésimo atributo e $|d|$ é o número total de atributos extraídos da base.

Considerando a hipótese de independência condicional entre os atributos, dado o valor da classe, $P(d|c)$ é calculada como o produto das probabilidades de ocorrência de cada valor de atributo t_i pertencente a classe c na base de treinamento. Essa probalidade é calculada utilizando a Equação 2.2.

$$P(d|c) = P((t_1, t_2, ..., t_{|d|})|c) = \prod_{i=1}^{|d|} P(t_i|c) \tag{2.2}$$

Para rotulação de uma nova instância, cada classe é calculada utilizando a Equação 2.1 e a que obtiver o maior valor de probabilidade será a classe da nova instância.

O algoritmo *Naive Bayes* apresenta bom desempenho quando há um número grande de atributos. Outra característica desse classificador é que ele possui uma implementação

com baixa complexidade, além de consumir baixo tempo de processamento.

2.4.2 *Multinomial Naive Bayes* (MNB)

Multinomial Naive Bayes [76] é um algoritmo que também utiliza o Teorema de *Bayes*, assim como o classificador *Naive Bayes*. Para o MNB, os atributos são considerados independentes uns dos outros, considerando que o valor da classe é dado. Diferentemente do NB, esse algoritmo considera, além dos atributos, a frequência com que eles ocorrem. Dessa forma, a diferença entre o cálculo realizado entre NB e MNB está no cômputo das probabilidades condicionais $P(d|c)$. Como o algoritmo MNB utiliza a frequência de ocorrência dos atributos, as probabilidades seguem uma distribuição multinomial, conforme apresentado na Equação 2.3.

$$P(d|c) = N! \times \prod_{i=1}^{k} \frac{(P_i)^{n_i}}{n_i!} \tag{2.3}$$

A quantidade de vezes que o atributo i aparece no documento é representado pelos valores $n_1, n_2, ..., n_k$ e as probabilidades de ocorrência do atributo i na classe c são representadas por $P_1, P_2, ...P_k$. N representa a soma do número de vezes que o atributo i aparece no documento ($N = n_1 + n_2 + ... + nk$). P_i é estimado pelo cálculo da frequência relativa do atributo i no conteúdo de todos os documentos pertencentes à classe c [137]. Com o valor obtido pelo cálculo dessas frequências, é possível calcular $P(d|c)$ e, posteriormente, por meio da Equação 2.1, a probabilidade de $P(c|d)$.

2.4.3 *Support Vector Machine* (SVM)

O algoritmo *Support Vector Machine* [129] padrão recebe como entrada um conjunto de dados e estima, para cada entrada dada, qual de duas possíveis classes a entrada faz parte, o que o caracteriza como um classificador linear binário não-probabilístico.

A partir de um conjunto de exemplos da base de treinamento, cujas as classes de cada exemplo já são pré-definidas, o algoritmo SVM é treinado e um modelo é construído para classificação de novos exemplos. Um modelo obtido pelo SVM é uma representação de exemplos como sendo pontos no espaço, onde os exemplos de cada categoria são divididos por um espaço claro que seja são distante quanto possível. Esse espaço consiste em um hiperplano de separação e uma margem entre as duas classes. As margens são obtidas considerando dois hiperplanos paralelos ao hiperplano inicial. Esses hiperplanos paralelos

são afastados do hiperplano inicial em direção aos exemplos dos dois grupos, até encontrá-los [130]. Esses exemplos são conhecidos como os vetores suporte e na Figura 2.1 estão representados pelos exemplos localizados dentro de círculos posicionados sobre as linhas tracejadas.

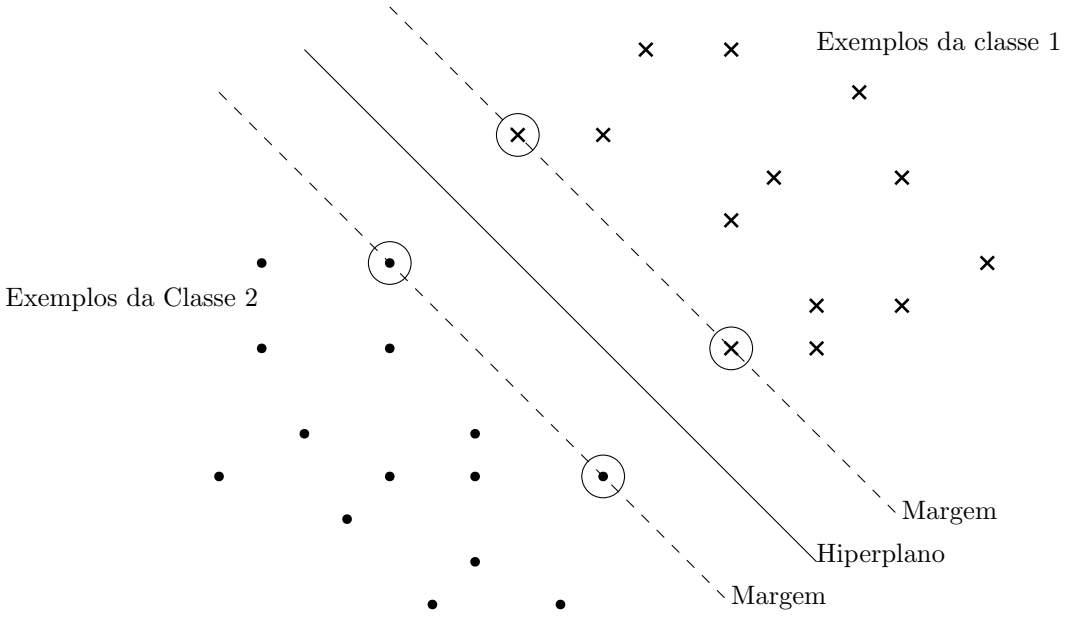

Figura 2.1: Hiperplano qualquer criado pelo classificador SVM para separar exemplos de duas classes

2.5 Medidas de avaliação dos classificadores

As medidas de avaliação de algoritmos de aprendizado de máquina supervisionado - Acurácia, Precisão (*Precision*), Sensitividade (*Recall*) e F1-*Measure* - frequentemente utilizadas na mineração de opiniões de texto, são apresentadas nesta seção.

Algoritmos de aprendizado de máquina supervisionado são capazes de induzir uma hipótese (classificador) descrita na forma de regras de classificação ou árvores de decisão. Uma tabela conhecida como matriz de confusão é utilizada para analisar o desempenho obtido por um classificador. A partir dela, é possível analisar o número de acertos e erros do classificador, além de ser possível extrair algumas medidas de desempenho, tais como Acurácia, Revocação, Precisão e F1-*Measure* [118]. A Tabela 2.3 corresponde a matriz de confusão de uma classificação binária (positivo/negativo).

Apesar da matriz de confusão mostrada representar uma classificação binária, ela também pode ser utilizada para representar múltiplas classes [42]. A partir da matriz de confusão de um classificador binário obtem-se quatro possíveis resultados:

Tabela 2.3: Matriz de confusão de classificadores binários

		Classe Predita	
		positivo	negativo
Classe Real	positivo	TP	FN
	negativo	FP	TN

- **TP:** *True Positives* ou Verdadeiros Positivos correspondem ao total de exemplos pertencente a classe positivo e que foram preditos corretamente pelo classificador como sendo positivo
- **TN:** *True Negatives* ou Verdadeiros Negativos correspondem ao total de exemplos pertencente a classe negativo e que foram preditos corretamente pelo classificador como sendo negativo
- **FP:** *False Positives* ou Falsos Positivos correspondem ao total de exemplos pertencente a classe negativo e que foram preditos erroneamente pelo classificador como sendo positivo
- **FN:** *False Negatives* ou Falso Negativos correspondem ao total de exemplos pertencente a classe positivo e que foram preditos pelo erroneamente pelo classificador como sendo negativo

A partir desses quatro resultados (TP, TN, FP e FN) obtêm-se as medidas comumente utilizadas para avaliação do desempenho de classificadores, conforme mostrado na Tabela 2.4.

Tabela 2.4: Medidas de classificação utilizadas por classificadores

Medida	Definição matemática	Avaliação
Acurácia	$\frac{TP+TN}{TP+TN+FP+FN}$	Proporção (probabilidade) dos atuais exemplos (classe: positivo e negativo) classificados corretamente
Precision	$\frac{TP}{TP+FP}$	Proporção (probabilidade) dos atuais exemplos identificados como sendo da classe positivo, corretamente classificados
Recall	$\frac{TP}{TP+FN}$	Proporção (probabilidade) dos atuais exemplos da classe positivo identificados corretamente
F1-*Measure*	$2\times\frac{Precision\times Recall}{Precision+Recall}$	Média harmônica entre as medidas *Precision* e *Recall*

As informações apresentadas neste capítulo são conceitos importantes da área de análise de sentimentos e são utilizadas na investigação da contribuição de *hashtags* na melhoria da qualidade do processo de mineração de opiniões de textos curtos extraídos do *Twitter*, durante períodos de campanha eleitoral. No próximo capítulo, são apresentados os trabalhos encontrados na literatura que investigaram o uso de *hashtags* no domínio político.

Capítulo 3

Hashtag no cenário eleitoral

Neste capítulo, é apresentado um conjunto de trabalhos que foram analisados com o objetivo de delimitar o problema investigado nesta tese. Na Seção 3.1, são apresentadas algumas abordagens de uso do *Twitter* no cenário de eleições. Na Seção 3.2, são relacionados trabalhos encontrados na literatura reportando o uso de *hashtags* no domínio investigado neste trabalho.

3.1 *Twitter* no cenário eleitoral

A popularidade do *Twitter* e o sucesso da campanha eleitoral presidencial de Barack Obama nessa rede, no ano de 2008, impulsionaram muitas pesquisas no cenário eleitoral. A predição do resultado de eleições, utilizando o *Twitter* como fonte de informação, é um exemplo de área que tem sido bastante explorada pelos pesquisadores. Em junho de 2017, os autores deste trabalho realizaram uma pesquisa na base de dados acadêmica do Google[1] com o objetivo de identificar o crescente interesse nessa área. Utilizando como critério de inclusão artigos cientícos, dissertações de mestrado e teses de doutorado, sem qualquer restrição de idioma, entre os anos de 2008 a 2015, e o conjunto de palavras-chave (*twitter OR tweet OR tweets*) + (*prediction OR predict OR predicting OR forecasting OR forecast OR forecasts*) + (*presidencial OR senate*), chegou-se a conclusão que, na janela temporal pesquisada, foram publicados 511 e 2560 trabalhos nos anos de 2008 a 2015, respectivamente, isto é um crescimento de 500%.

O *Twitter* também tem sido investigado pelos pesquisadores para uso em outras áreas no cenário de eleições. Em [25], os autores analisaram o uso do *Twitter* como uma ferra-

[1] https://scholar.google.com.br

menta aliada ao processo de divulgação de propostas políticas. Em [108], uma variedade de aspectos na modelagem do debate, utilizando o *Twitter*, além do aspecto político individual, já foi analisada. Em [12], os autores analisaram como as informações públicas dos usuários do *Twitter*, por meio da captura e análise da *hashtag* "#ausvoters", na eleição australiana no ano de 2010, foram usadas para descrever padrões de atividades em mídias sociais.

Segundo [102], o estudo das reações de eleitores em mídias sociais durante eventos, como debates eleitorais, tem sido uma outra área de pesquisa utilizando o *Twitter* nesse domínio. Por exemplo, em [75] os autores examinaram mensagens do *Twitter* durante o debate para a corrida presidencial dos EUA em novembro de 2011, e [25] utilizou a abordagem de análise de sentimentos para classificar a polaridade de mensagens postadas durante os debates presidenciais no EUA no ano de 2008.

No *Twitter*, *hashstag* é um recurso utilizado nas mensagens para marcar um tópico de interesse ou um público alvo e no cenário eleitoral ela têm sido incorporada nas mensagens de forma bastante expressiva [98]. Na próxima seção, é apresentada uma relação de trabalhos encontrados na literatura que reportaram o uso de *hashtags* em cenário de eleições.

3.2 Abordagens de uso de *hashtag*

Hashtags têm sido utilizadas em muitos estudos de análise de sentimentos que utilizam o *Twitter* como fonte de dados. No cenário político, *hashtags* têm sido utilizadas desde processos mais simples, como coleta de dados, até situações mais complexas, como rotulação automática de mensagens. Nas próximas seções, são apresentadas as diferentes formas de uso de *hashtags* disponíveis em *tweets* políticos, identificados nos trabalhos encontrados na literatura.

3.2.1 *Hashtag* e coleta de dados

Em [12], a *hashtag* *#ausvotes* foi utilizada para obtenção de *tweets* relacionados a eleição federal na Austrália no ano de 2010. Essa *hashtag* foi utilizada para coletar 415.009 *tweets* postados durante 38 dias (35 dias antes do dia da votação e 3 dias após). Além de ter sido utilizada na etapa de coleta, ela foi utilizada também na identificação de diversas métricas estatísticas, como a quantidade de menções a candidatos nos *tweets*, número de *tweets* publicados por hora, percentual de mensagens reencaminhadas (*retweets*) e tam-

bém na identificação de temas de maior relevância discutidos durante o período eleitoral. Segundo os autores, *hashtags* com características semelhantes a que foi proposta no trabalho, possibilitam a obtenção de informações importantes, a partir do rastreamento de atividades tanto individuais quanto gerais sobre acontecimentos realizados no *Twitter*.

Em [102], os autores utilizaram a estratégia de combinar nomes de candidatos e de seus vice-candidatos com *hashtags* de campanha para coletar *tweets* relacionados à eleição presidencial na Indonésia, realizada no ano de 2014. Segundo os autores, as palavras-chave utilizadas para compor a lista de *hashtags* foram obtidas manualmente a partir do *Twitter Trending Topics*,[2] e eram caracterizadas por acrônimos (Ex: *#JKW4P* e *#JKWJK*) e por combinações de palavras sem abreviações (Ex: *#DukungPrabowoHatta* e *#IndonesiaHebat*). Segundo os autores, as palavras deveriam estar relacionadas à eleição para serem selecionadas. Como resultado, 11 *hashtags* foram incorporadas ao conjunto de palavras-chave para coleta dos *tweets*. Nesse trabalho, os autores investigaram a seguinte hipótese: "utilizar mais palavras-chave descreve melhor a situação no *Twitter* e, consequentemente, melhora a acurácia da predição". Com base nos resultados obtidos, eles concluíram que a quantidade de palavras-chave influencia positivamente na melhoria da acurácia e que entre três províncias analisadas, três *hashtags* tiveram uma influência maior na comprovação dessa hipótese.

Em [6], os autores escolheram as *hashtags* *#USElections2012*, *#USElections* e *#Elections2012* como palavras-chave para coletar mensagens relacionadas às eleições presidenciais dos EUA no ano de 2012, e as *hashtags* *#BJP*, *#Congress*, *#KJP* e *#KarnatakaElections* para as eleições do estado de Karnataka na Índia no ano de 2013. Em [16], os *tweets* coletados sobre as eleições presidenciais na Colômbia, no ano de 2014, foram filtrados por quatro *hashtags* relacionadas ao processo eleitoral (*#Elecciones2014*, *#ColombiaElige*, *#EleccionesColombia* e *#ColombiaDecide*), nomes completos dos candidatos ou menções a usuários identificando algum dos presidenciáveis. No trabalho reportado em [22], as *hashtags* políticas escolhidas pelos autores, #p2 e #tcot, foram utilizadas não somente para seleção de *tweets* políticos, mas também para identificação de outras *hashtags* políticas relevantes. Em [10], *tweets* relevantes ao domínio político foram identificados e utilizados nas análises, a partir de uma pesquisa realizada pelo nome do partido do candidato e de suas abreviações, juntamente com a *hashtag* escolhida pelos autores, #gel1.

[2] Termos mais populares publicados no *Twitter* numa janela de tempo específica

3.2.2 *Hashtag* e pré-processamento

Conforme apresentado no capítulo anterior, diversas técnicas são utilizadas na fase de pré-processamento para limpeza de textos. Em alguns estudos, *hashtags* são simplesmente removidas dos *tweets* nessa fase [15, 20, 95, 125]. Em outros, elas são mantidas nas mensagens e utilizadas nas análises sem realizar qualquer tipo de investigação da contribuição delas. Já em algumas pesquisas, elas recebem tratamentos especiais.

Em alguns trabalhos, *hashtags* são substituídas por *tags* específicas (*placeholders*), por exemplo por *HASHTAG* [30], com o objetivo de padronizar *tokens* de mesmo significado semântico. Essa estratégia é encontrada em muitos trabalhos que utilizam o *Twitter* como fonte de dados [38, 39, 43, 62, 61, 68, 80, 93, 104, 112]. Em [16], os autores investigaram o potencial do *Twitter* na investigação da inferência das intenções de votos dos usuários na eleição presidencial na Colômbia, realizada no ano de 2014, e uma das técnicas utilizadas na fase de pré-processamento foi a substituição de *hashtags* por *placeholders*.

Uma outra abordagem de uso de *hashtags* na fase de pré-processamento consiste em suprimir o símbolo "#" das *hashtags* e utilizar as palavras que as compõem como parte do texto do *tweet*. Em [31], o símbolo "#" foi suprimido de todas as *hashtags* identificadas nas mensagens e as palavras contidas nelas foram decompostas em unigramas. Nesse trabalho, os autores utilizaram um algoritmo de programação dinâmica baseado nos dados *one-gram* do Google [110] para segmentar as *hashtags*. Por exemplo, utilizando esse algoritmo as *hashtags* *#votewisely* e *#teamgej* foram mapeadas nos atributos unigrama *vote* e *wisely*, e *team* e *gej*, respectivamente. Em [3], os autores também utilizaram a abordagem de segmentação de palavras contidas em *hashtags*, a partir da identificação de letras maiúsculas. Por exemplo, o *tweet* "*Ahok becomes a suspect, #JailAhok*" após ter sido pré-processado foi representado por "*Ahok becomes a suspect, Jail Ahok*". Em [5], os pesquisadores também optaram por remover o símbolo "#" das *hashtags*, porém para um grupo específico delas. Na abordagem utilizada nesse estudo, *hashtags* contendo alguma das palavras-chave utilizadas na etapa de coleta foram mantidas nas mensagens sem o símbolo "#".

3.2.3 *Hashtag* e análise de sentimentos

Além de serem utilizadas para coletar mensagens de cunho político e receberem tratamentos especiais no processo de limpeza de *tweets*, a contribuição de *hashtags* diretamente no processo de análise de sentimentos de *tweets* políticos foi investigada.

Em [3], os autores investigaram a contribuição do uso de *hashtags* na análise de sentimentos de *tweets* sobre a figura política de Basuki Tjahaja Purnama, na Indonésia no ano de 2016. Nesse trabalho, os autores analisaram o desempenho dos algoritmos NB, SVM, LR e RF, a partir do uso de três atributos: número de *hashtags* positivas e negativas ($SentiHT$), número de palavras positivas e negativas ($SentiLex$) e unigrama (utilizado como *baseline* nesse trabalho). O atributo $SentiHT$ foi definido a partir da contagem de *hashtags* positivas e negativas, presentes em um *tweet*. Num primeiro momento foi realizada a classificação manual da polaridade de todas as *hashtags* da base de dados em positivo, negativo, neutro e irrelevante. Dois *datasets* rotulados manualmente por um grupo de três pessoas, contendo 200 e 400 amostras cada um, foram utilizados nas análises. O primeiro *dataset* (A) foi coletado a partir de palavras-chave compostas por *hashtags* e o segundo (B) por *hashtags*, e o pelo nome do político. Para o *dataset* A, as acurácias obtidas pelos algoritmos NB, SVM e LR, utilizando o modelo $SentiHT$, foram iguais a 95%, superiores aos valores obtidos pelo *baseline*, cujos valores foram iguais a 91,5%, 93% e 93%, respectivamente. Para o *dataset* (B), apenas o classificador SVM obteve a melhor acurácia ao utilizar esse atributo. Segundo os autores, o estudo proposto mostrou a viabilidade de se explorar o uso de *hashtags* na classificação da polaridade de *tweets* políticos.

A contribuição do uso de *hashtags* na análise de sentimentos de mensagens em outros domínios também foi investigado. Em [113], os autores propuseram um modelo baseado em duas fases para avaliar a eficiência do uso de *hashtags* na classificação do sentimento de mensagens postadas no *Twitter*, durante a copa do mundo de futebol no ano de 2014. Na primeira fase, foi criado um classificador baseado em duas categorias de *hashtags*: *sentiment* (possui sentimento) e *non-sentiment* (não possui sentimento), a partir da combinação de vários dicionários léxicos. Na segunda fase, *tweets* contendo os dois tipos de *hashtags* foram selecionados da base de dados e classificados pelo classificador obtido na primeira fase em positivo, negativo e neutro. A acurácia do classificador proposto foi obtido a partir dos algoritmos NB, SVM, ME e $C4.5$[3]. No processo de classificação do sentimento de mensagens contendo *hashtags* do tipo *non-sentiment*, o algoritmo $C4.5$ foi o que apresentou a melhor acurácia (86,41%), seguido do classificador baseado em *hashtag* (86,07%). Nas mensagens contendo *hashtags* do tipo *sentiment*, os melhores algoritmos foram SVM (82,85%), $C4.5$ (82,78%) e o modelo proposto (81,14%). Segundo os autores, o estudo realizado forneceu evidências empíricas de que *hashtags* são úteis para a análise do sentimento de *tweets*.

[3]Extensão do algoritmo ID3 [48]

3.2.4 *Hashtag* e rotulação de mensagens

Por último, foi investigada a contribuição do uso de *hashtags* no processo de classificação automático da polaridade de *tweets* políticos. Em [11, 121], os autores utilizaram a *hashtag* *#sarcasm* para classificar automaticamente o sentimento de um *dataset* para detecção de sarcasmo, utilizando a seguinte regra: *tweets* contendo a presença da *hashtag* *#sarcasm* são rotulados como sendo do tipo sarcasmo e, do tipo não-sarcasmo, caso contrário. Utilizando essa abordagem os autores conseguiram aumentar a base de treinamento inicial de 6.000 *tweets*.

No trabalho reportado em [60], os autores também utilizaram *hashtags* para rotulação automática de um *dataset* obtido a partir do *corpus* de Edinburgh *Twitter* [101]. Nesse trabalho, a classificação automática dos *tweets* consistiu inicialmente em identificar as *hashtags* mais frequentes encontradas no *corpus* e, em seguida, selecionar um conjunto de 15 *hashtasgs* utilizadas com mais frequência juntamente com outras *hashtags* mais frequentes identificadas no *Twitter*, dividí-la em três classes: positivo, negativo e neutro. Uma base de treinamento composta por 222.570 *tweets* foi classificada utilizando essa abordagem. Segundo os autores, *hashtags* mostraram ser úteis na coleta de dados de treinamento.

Em [3], os autores rotularam automaticamente uma amostra de 4.000 *tweets* (2.000 positivos e 2.000 negativos) sobre o ex-governador de Jacarta (Indonésia), a partir do uso de *hashtags*. O volume de *hashtags* com sentimento positivo e negativo presentes em uma mensagem, foi utilizado como abordagem para classificação da polaridade das mensagens. Um *tweet* era classificado em positivo quando havia pelo menos uma *hashtag* pertencente a essa classe e nenhuma *hashtag* contendo sentimento negativo, e vice-versa.

3.2.5 Discussão

Neste capítulo, foram apresentadas formas de uso do *Twitter* no cenário de eleições e as principais abordagens utilizadas pelos pesquisadores para explorar as potencialidades de *hashtags* nesse cenário. A Tabela 3.1 apresenta um quadro demonstrativo com as diferentes estratégias de uso de *hashtags* identificadas nos trabalhos investigados.

Conforme pode ser visto no quadro demonstrativo apresentado na Tabela 3.1, os trabalhos investigados reportaram o uso de *hashtags* no domínio político de diversas maneiras:

Tabela 3.1: Quadro demonstrativo com as diferentes abordagens de uso de *hashtags*

Referência	COL	PLA	SS#	R#	AS	ROT	Fonte de dados
Alfina et al.[3]	✓		✓		✓	✓	*tweet*
Almatrafi et al.[5]			✓				*tweet*
Anjaria e Guddeti[6]	✓						*tweet*
Bermingham e Smeaton[10]	✓						*tweet*
Bouazizi e Ohtsuki[11]						✓	*tweet*
Bruns e Burgess[12]	✓						*tweet*
Ceron et al.[15]				✓			*tweet*
Ceron e León[16]	✓	✓					*tweet*
Chung e Mustafaraj[20]				✓			*tweet*
Conover et al.[22]	✓						*tweet*
Fink et al.[31]			✓				*tweet*
Kouloumpis et al.[60]						✓	*tweet*
Paula Filho e Garcia[95]				✓			*tweet*
Prasetyo[102]	✓						*tweet*
Simeon e Hilderman [113]					✓		*tweet*
Sulis et al.[11]						✓	*tweet*
Tumasjan et al.[125]				✓			*tweet*

Onde:

- COL: *hashtags* são utilizadas sozinhas ou em conjunto com outras palavras-chave no processo de coleta de *tweets*;
- PLA: *hashtags* são substituídas por *placeholders*;
- SS#: o símbolo "#" é suprimido das *hashtags* e as palavras contidas nelas são fragmentadas e incorporadas ao restante do texto do *tweet*;
- R#: *hashtags* são removidas dos *tweets*;
- AS: *hashtags* são utilizadas diretamente no processo de análise de sentimentos dos *tweets*; e
- ROT: *hashtags* são utilizadas no processo de rotulação automática de mensagens.

Dentre as estratégias de uso de *hashtag* identificadas nos trabalhos investigados, a coleta de dados foi a mais frequente entre elas. No domínio de eleições, apenas um trabalho [3] se propôs a investigar, de fato, a contribuição de *hashtags* diretamente no processo de análise de sentimentos de *tweets*, a partir da utilização delas no processo de representação

de atributos aos classificadores. Tanto nesse trabalho quanto em um segundo estudo investigado, porém em um domínio diferente do político, os autores mostraram que *hashtags* são eficazes no processo de classificação do sentimento de *tweets*.

Dos trabalhos investigados, o *tweet* foi a única fonte de dados a partir da qual as *hashtags* foram obtidas para serem utilizadas nas análises. Em outras pesquisas realizadas no domínio político, informações localizadas nos perfis dos usuários, tais como sexo [13], idade [102, 103] e localização [100] já foram utilizadas, porém na fase de coleta de dados. Nas pesquisas realizadas em [96, 97], os autores investigaram a contribuição de informações contidas nas descrições dos perfis dos usuários no processo de rotulação semiautomática de *tweets* políticos, postados durante períodos de campanha eleitoral presidencial, e chegaram a conclusão que esse tipo de informação pode contribuir na classificação da polaridade do sentimento de *tweets* de cunho político. Um outro estudo proposto em [98] mostrou que aproximadamente 21% de todas as mensagens da base de dados haviam sido publicadas por usuários cujas descrições continham algum tipo de manifestação política sobre candidatos.

Outros trabalhos reportados na literatura, que fizeram uso da abordagem de aprendizado de máquina supervisionado na análise de sentimentos de *tweets* no cenário eleitoral [3, 10, 14, 31, 71, 77, 86, 95, 102], foram investigados com o objetivo de verificar se informações contidas em descrições de perfis de usuários já haviam sido utilizadas no processo de melhoria do desempenho de algoritmos de aprendizado de máquina. Chegou-se a conclusão que nesses trabalhos apenas as informações contidas em *tweets* haviam sido utilizadas nas análises.

Até o momento não foi encontrado nenhum trabalho de pesquisa que tenha investigado a contribuição de *hashtags*, disponíveis tanto em *tweets* quanto em descrições de perfis de usuários, na análise de sentimentos de *tweets* no cenário eleitoral. Portanto, no próximo capítulo é apresentando o modelo proposto neste trabalho para verificar se *hashtags*, contidas em *tweets* e em descrições de perfis de usuários, melhoram a acurácia de algoritmos de aprendizado de máquina supervisionado utilizados no processo de classificação do sentimento de *tweets* no cenário eleitoral.

Capítulo 4

Materiais e métodos

Neste capítulo, é apresentado o modelo proposto para verificar se *hashtags*, contidas em *tweets* e em descrições de perfis de usuários, melhoram a acurácia de algoritmos de aprendizado de máquina supervisionado utilizados no processo de classificação do sentimento de *tweets* no cenário eleitoral. As etapas do modelo são apresentadas na Figura 4.1.

Na Seção 4.1, são apresentadas as principais características da etapa de coleta de dados e um conjunto de *datasets* obtidos para serem utilizados na última etapa do modelo. Na Seção 4.2, são apresentadas as técnicas de pré-processamento utilizadas pelo modelo e, na Seção 4.3, são apresentadas as principais características da etapa de análise de dados.

4.1 Coleta de dados

Nesta etapa, mensagens publicadas no *Twitter* durante períodos de campanha eleitoral são coletadas e armazenadas em uma base de dados, juntamente com informações localizadas nas descrições dos perfis dos usuários. Além disso, um conjunto de *datasets* é obtido para ser utilizado na última etapa do modelo. A Figura 4.2, apresenta as ações realizadas na etapa de coleta de dados do modelo proposto. A seguir, é apresentado em detalhes o processo de obtenção de cada um dos *datasets*.

O *dataset TweetsAL* tem como objetivo armazenar os *tweets* que serão utilizados no processo de análise de sentimentos das mensagens. Essa amostra consiste num conjunto aleatório balanceado de *tweets* extraído da base de dados pelo módulo "Seleção de *tweets*". O critério utilizado para realizar esse balanceamento leva em consideração o número de mensagens com e sem *hashtags*. Em seguida, a amostra de *tweets* selecionada é rotulada manualmente em três classes distintas, assim como em outros trabalhos [1, 2, 37, 40,

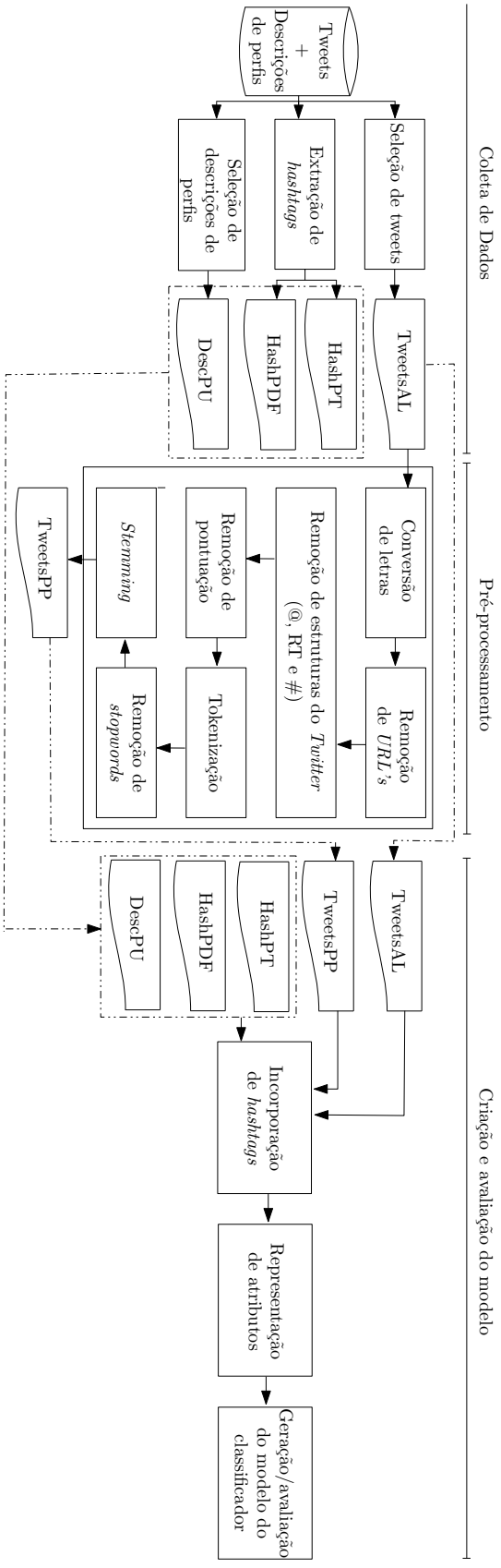

Figura 4.1: Etapas do modelo proposto

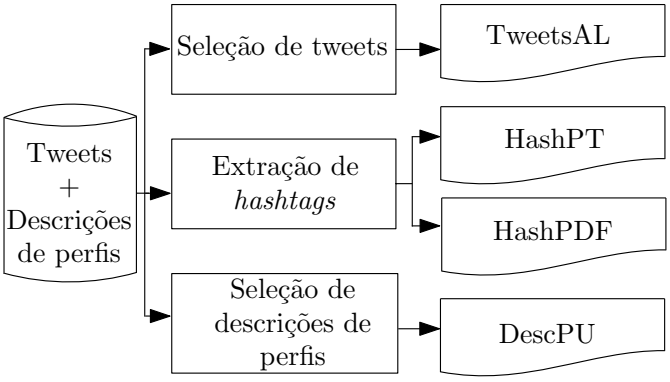

Figura 4.2: Conjunto de *datasets* obtidos na etapa de coleta de dados do modelo proposto

59, 60, 89, 92, 105]. Cada mensagem é classificada em uma das seguintes classes: Pró-Candidato1 ($PC1$), Pró-Candidato2 ($PC2$) e Neutro (N). Uma mensagem é classificada em $PC1$ ($PC2$) se houver a presença de uma ou mais menções positivas sobre o Candidato 1 (Candidato 2) e/ou a presença de uma ou mais menções negativas sobre o Candidato 2 (Candidato 1) na mesma mensagem, e Neutro, quando não existir nenhuma menção positiva/negativa sobre nenhum dos candidatos.

O segundo *dataset*, $HashPT$, consiste num grupo de *hashtags* obtidas a partir dos *tweets*, contendo palavras e/ou expressões utilizadas pelos usuários durante períodos de campanha eleitoral para fazer referência a candidatos. Esse grupo de *hashtags* será referenciado neste trabalho pela nomeclatura "*hashtags* políticas" e o conjunto de *hashtags* contendo qualquer outra palavra não relativa a candidatos será referenciada pelo termo "*hashtags* não-políticas". O conjunto de todas as *hashtags* políticas identificadas nos *tweets* da base de dados é obtido a partir do módulo "Extração de *hashtags*". Na última etapa do modelo, esse *dataset* é utilizado com o objetivo de verificar a contribuição de *hashtags* localizadas em *tweets* na melhoria do desempenho de algoritmos de aprendizado de máquina supervisionado, utilizados para classificar o sentimento das mensagens armazenados em $TweetsAL$. Além disso, as *hashtags* armazenadas nesse *dataset* são utilizadas na etapa de análise de dados do modelo proposto com o objetivo de verificar a contribuição de um novo atributo, baseado em *hashtags* políticas, na melhoria do desempenho dos classificadores.

O terceiro *dataset*, $HashPDF$, consiste no grupo de *hashtags* políticas identificadas nas descrições dos perfis dos usuários, obtidas a partir do módulo "Extração de *hashtags*", para ser utilizado na última etapa do modelo com o objetivo de verificar a contribuição desse tipo de informação na classificação do sentimento dos *tweets* armazenados em $TweetsAL$ e também para verificar a contribuição de um atributo baseado nesse tipo de

informação na melhoria do desempenho dos classificadores.

O último *dataset*, *DescPU*, corresponde ao conjunto das descrições dos perfis dos usuários responsáveis pelas postagens das mensagens armazenadas em *TweetsAL*. O critério utilizado pelo modelo para composição do *dataset DescPU* consiste em armazenar a última atualização da descrição do perfil contendo pelo menos uma palavra e/ou expressão utilizada pelo usuário, para fazer referência a algum dos candidatos.

4.2 Pré-processamento

Nesta etapa, informações indesejadas encontradas nas mensagens armazenadas no *dataset* *TweestAL* são eliminadas, a partir do conjunto de técnicas apresentadas na Figura 4.3.

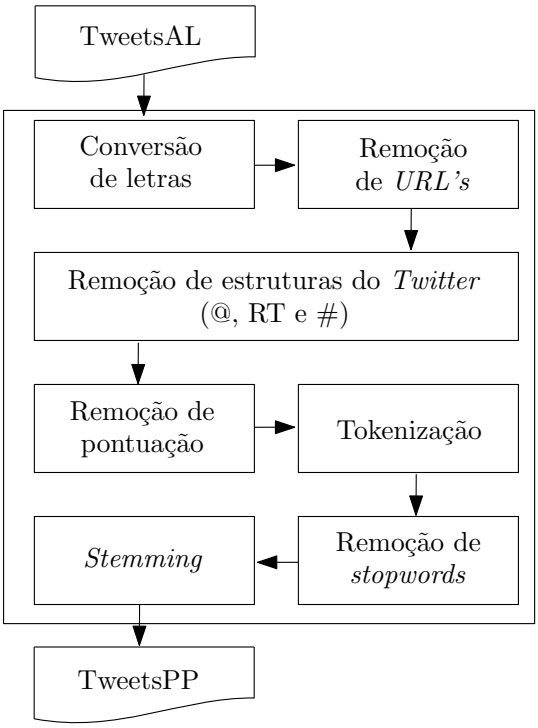

Figura 4.3: Técnicas utilizadas na etapa de pré-processamento do modelo proposto

No modelo proposto, assim como em [10, 108], as *URL's* são removidas dos *tweets* durante a fase de pré-processamento. A abordagem de converter letras maiúsculas em minúsculas, reportada em outras pesquisas, tais como [30, 60, 68, 93], também é utilizada no modelo dessa tese. Assim como em [3, 21], as estruturas do *Twitter* compostas por menções a usuários (@), símbolos de *retweet* (RT) e *hashtags* (#), e símbolos de pontuação, são removidos das mensagens. Por último, são aplicadas as técnicas de tokenização, remoção de *stopwords* e *stemming*.

O uso da técnica de remoção de *stopwords* em cenários distintos ao político é encontrado com frequência na literatura [1, 59, 60, 68, 92, 93]. No domínio político, por exemplo na área de predição do resultado de eleições presidenciais, o uso dessa técnica é reportada com menos frequência. Em [86, 108], os autores utilizaram essa técnica na fase de pré-processamento antes da mineração de opiniões das mensagens de cunho político. Nesses dois últimos trabalhos, essa técnica foi utilizada sem a investigação de sua eficácia no desempenho dos métodos utilizados. É comumente encontrado na literatura trabalhos no domínio político sem mencionar o uso dessa técnica na fase de limpeza dos dados [23, 36, 88, 95, 109, 114, 125]. Por esse motivo, no modelo proposto a contribuição dessa técnica é avaliada antes do seu uso.

Diversos trabalhos sobre mineração de opiniões de *tweets* não-políticos, encontrados na literatura, como em [8, 21, 27, 28, 53, 62, 64, 68, 74, 131], reportam o uso da técnica de *stemming* na fase de pré-processamento. Poucos trabalhos no domínio político, utilizando o *Twitter* como fonte de dados, reportam o seu uso, como em [20]. Em outras pesquisas, como em [77, 88], ela foi apenas mencionada pelos autores, mas não foi utilizada. Por esse motivo, o modelo proposto também avalia a contribuição dessa teécnica antes do seu uso.

Nesta etapa do modelo, as *hashtags* são suprimidas dos *tweets*, porém, na etapa de análise de dados elas são incorporadas novamente as mensagens com a finalidade de verificar a contribuição delas na melhoria do desempenho dos classificadores. O resultado final produzido nesta etapa do modelo consiste na obtenção do *dataset* $TweetsPP$, utilizado na última etapa do modelo.

4.3 Criação e avaliação do modelo

Nesta etapa, os *tweets* pré-processados, armazenados em $TweetsPP$, são utilizados juntamente com as informações armazenadas nos outros *datasets* ($TweetsAL$, $HashPT$, $HashPDF$ e $DescPU$) para avaliar a contribuição de *hashtags*, disponíveis tanto nos *tweets* quanto nas descrições dos perfis dos usuários, na melhoria do desempenho dos classificadores.

A contribuição de *hashtags* no processo de classificação do sentimento das mensagens é avaliada sob quatro diferentes perspectivas: (a) análise da contribuição de *hashtags* políticas e não-políticas contidas em *tweets*, (b) avaliação da contribuição de *hashtags* políticas e não-políticas contidas em descrições de perfis de usuários, (c) avaliação em

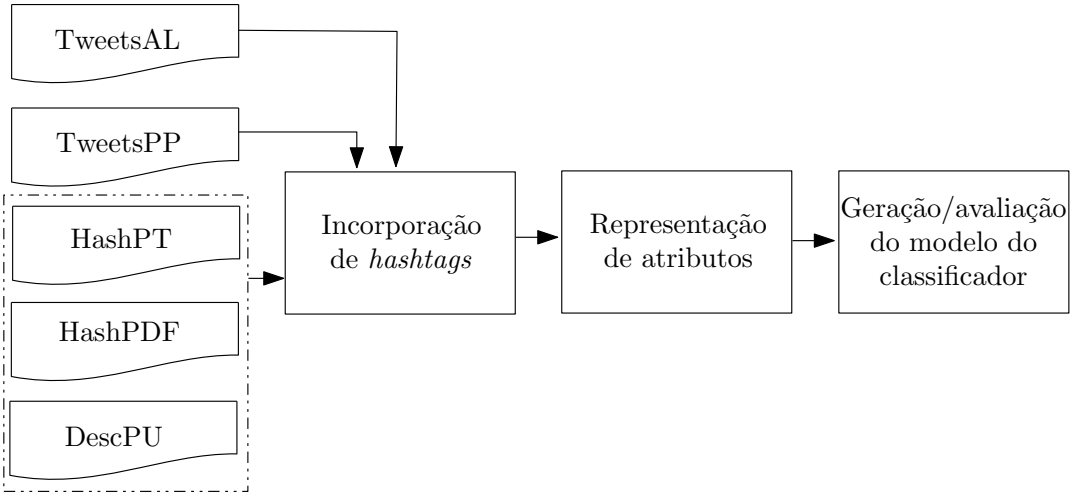

Figura 4.4: Etapas da criação e avaliação do modelo proposto

conjunto da contribuição de *hashtags* contidas em *tweets* e em descrições de perfis de usuários e (d) avaliação da contribuição de dois atributos baseados em *hashtags* políticas. A Figura 4.4 apresenta os módulos responsáveis por realizar essas avaliações. Na Seção 4.3.1, é apresentada a abordagem utilizada pelo modelo para incorporação de *hashtags* aos *tweets*. Na Seção 4.3.2, os *tweets* são representados em formatos de atributos. Na Seção 4.3.3, é realizada a análise de sentimento dos *tweets*.

4.3.1 Incorporação de *hashtags*

Nesta etapa, *hashtags* contidas em *tweets* ($HashPT$) e em descrições de perfis ($HashPDF$) são incorporadas aos *tweets* armazenados em $TweetsPP$, a partir de três estratégias, conforme descrito a seguir.

A primeira estratégia de incorporação de *hashtags* é realizada a nível de *tweets*, ou seja, as *hashtags* excluídas das mensagens na etapa de pré-processamento são novamente incorporadas a elas (posicionadas no mesmo lugar onde se encontravam antes da fase de limpeza de dados). Na etapa de análise de sentimentos, a contribuição de cada categoria de *hashtag*, política e não-política, contida em *tweets*, é investigada pelo modelo. Por isso, nesta etapa, três cenários são propostos para configuração dos *tweets*: (T1) *hashtags* não-políticas são incorporadas aos *tweets*, (T2) *hashtags* políticas são incorporadas aos *tweets* e (T3) *hashtags* políticas e não-políticas são incorporadas aos *tweets*. Nesta operação, são utilizados os seguintes *datasets*: $TweetsPP$, $TweetsAL$ e $HashPT$.

A segunda estratégia de incorporação de *hashtags* é realizada a nível de *tweets* e de descrições de perfis de usuários, ou seja, as *hashtags* identificadas nas descrições dos perfis

dos usuários, armazenadas no *dataset* $DescPU$, são inseridas no final de cada mensagem postada pelo usuário ($TweetsPP$). A contribuição de cada categoria de *hashtag*, política e não-política, identificada na descrição do perfil do usuário, a partir do *dataset* $HashPDF$, é avaliada individualmente na última etapa do modelo. Por esse motivo, três cenários são propostos para configuração dos *tweets*: (D1) adiciona-se ao final dos *tweets* as *hashtags* não-políticas das descrições, (D2) adiciona-se ao final dos *tweets* as *hashtags* políticas das descrições e (D3) adiciona-se ao final dos *tweets* todas as *hashtags* (políticas/não-políticas) das descrições.

A última estratégia consiste em utilizar as *hashtags* contidas em *tweets* e em descrições de perfis de usuários e incorporá-las aos *tweets*, utilizando as abordagens supracitadas tanto para os *tweets* quanto para as descrições de perfis de usuários.

4.3.2 Representação de atributos

Após os *tweets* terem sido configurados na etapa anterior, são gerados os atributos preditivos para serem utilizados pelos algoritmos de classificação. Nesta etapa do modelo, são utilizados os formatos unigrama e bigrama, assim como em outros trabalhos [1, 10, 89, 92, 93, 94, 115, 131], para representação dos atributos aos classificadores.

Nesta etapa, dois atributos propostos neste trabalho, baseado em *hashtags* políticas, são utilizados para representação dos *tweets* aos algoritmos. O primeiro atributo baseia-se em *hashtags* políticas identificadas nos *tweets* da amostra, e o segundo em *hashtags* políticas identificadas nas descrições dos perfis dos usuários. O primeiro atributo recebe o nome de *Tweet Political Support Bit* ($TPSB$) e baseia-se na presença/ausência de *hashtags* políticas contidas em *tweets*, armazenados no *dataset* $TweetsPP$. Para representação desse atributo é utilizada uma tupla composta por quatro parâmetros binários (C1_THSN, C2_THSN, C1_THPR, C2_THPR), onde:

- **C1_THSN**: contém um dos seguintes valores: 1, se houver no *tweet* alguma *hashtag* contendo uma ou mais palavras utilizadas durante o período de campanha eleitoral para fazer referência ao candidato 1, e 0 caso contrário;
- **C2_THSN**: contém um dos seguintes valores: 1, se houver no *tweet* alguma *hashtag* contendo uma ou mais palavras utilizadas durante o período de campanha eleitoral para fazer referência ao candidato 2, e 0 caso contrário;
- **C1_THPR**: contém um dos seguintes valores: 1, se houver no *tweet* alguma *hashtag*

contendo uma ou mais expressões de repúdio sobre o candidato 1, e 0 caso contrário;e

- **C2_THPR**: contém um dos seguintes valores: 1, se houver no *tweet* alguma *hashtag* contendo uma ou mais expressões de repúdio sobre o candidato 2, e 0 caso contrário.

O segundo atributo proposto neste trabalho, o *Description Political Support Bit* (*DPSB*), baseia-se na presença/ausência de *hashtags* políticas contidas nas descrições dos perfis dos usuários, armazenadas em *DescPU*. Para representação desse atributo, é utilizada uma tupla composta por quatro parâmetros binários (C1_DHSN, C2_DHSN, C1_DHPR, C2_DHPR), onde:

- **C1_DHSN**: contém um dos seguintes valores: 1, se houver na descrição do perfil do usuário alguma *hashtag* contendo uma ou mais palavras utilizadas durante o período de campanha eleitoral para fazer referência ao candidato 1, e 0 caso contrário;
- **C2_DHSN**: contém um dos seguintes valores: 1, se houver na descrição do perfil do usuário alguma *hashtag* contendo uma ou mais palavras utilizadas durante o período de campanha eleitoral para fazer referência ao candidato 2, e 0 caso contrário;
- **C1_DHPR**: contém um dos seguintes valores: 1, se houver na descrição do perfil do usuário alguma *hashtag* contendo uma ou mais expressões de repúdio sobre o candidato 1 e 0 caso contrário; e
- **C2_DHPR**: contém um dos seguintes valores: 1, se houver na descrição do perfil do usuário alguma *hashtag* contendo uma ou mais expressões de repúdio sobre o candidato 2 e 0 caso contrário.

4.3.3 Geração/avaliação do modelo do classificador

Nesta última etapa, a abordagem de aprendizado de máquina supervisionado é utilizada para classificar o sentimento dos *tweets* políticos. Algoritmos que utilizam essa abordagem, são comumente encontrados na literatura sobre mineração de opiniões de *tweets* no cenário eleitoral [3, 10, 14, 31, 71, 77, 86, 95, 102] e, por isso, essa abordagem foi adotada para ser utilizada no modelo proposto.

Os *tweets* configurados nas seções 4.3.1 e 4.3.2 são utilizados nesta última etapa do modelo com o objetivo de analisar a contribuição de cada categoria de *hashtag*, contidas em *tweets* e em descrições de perfis de usuário, além dos atributos propostos, na melhoria do desempenho dos classificadores.

Para avaliar o desempenho dos classificadores utilizados nesta última etapa, foi escolhido o método de validação cruzada estratificada com 10 iterações [42] (*stratified k-fold cross validation, com k=10*) por ser um método comumente utilizado em vários trabalhos reportados na literatura [3, 11, 22, 42, 58, 57, 80, 102, 109]. Na abordagem de classificação utilizando esse tipo de validação, são gerados um total de 10 modelos por classificador, um em cada interação. Para avaliação dos classificadores, cada modelo gerado é utilizado para a classificação da base de teste. A soma dos acertos e erros do classificador de cada classe, considerando todas as iterações, é baseado na matriz de confusão apresentada na Tabela 4.1.

Tabela 4.1: Representação da matriz de confusão utilizada na classificação de *tweets* postados em cenários de eleições

		Classe Predita		
		Pró-Candidato1	Pró-Candidato2	Neutro
Classe Real	Pró-Candidato1	a	b	c
	Pró-Candidato2	d	e	f
	Neutro	g	h	i

A Acurácia, medida utilizada no modelo proposto para avaliação do desempenho dos classificadores, consiste na razão entre a quantidade de mensagens classificadas corretamente (a+e+i) e a quantidade total de mensagens da amostra, conforme definida na Equação 4.1.

$$Acuracia = \frac{a + e + i}{a + b + c + d + e + f + g + h + i} \tag{4.1}$$

Essa mesma medida de avaliação foi utilizada em outros trabalhos no domínio político para avaliação do desempenho de algoritmos de aprendizado de máquina supervisionado [3, 15, 71, 78, 102]. Nesta tese, é utilizada a mesma abordagem de [3], onde apenas essa medida foi adotada para avaliar o desempenho dos classificadores.

No capítulo seguinte, é apresentado o resultado de um estudo realizado para investigar a relevância de *hashtags* contidas em *tweets* e em descrições de perfis de usuários, a partir de duas amostras presidenciais. Nessa investigação, a primeira etapa do modelo proposto foi utilizada para realizar a coleta das mensagens. No capítulo subsequente, essa base de dados é utilizada nas duas últimas etapas do modelo para analisar a contribuição de *hashtags* e informações contidas em descrições de perfis de usuários, na melhoria do desempenho de classificadores utilizados para analisar o sentimento das amostras.

Capítulo 5

Estudo da relevância de hashtags em cenários de eleições

Neste capítulo, *tweets* no cenário eleitoral são analisados a partir de duas amostras presidenciais. As mensagens utilizadas nas análises correspondem as eleições presidenciais ocorridas no Brasil e nos EUA, nos anos de 2014 e 2016, respectivamente. O objetivo principal deste capítulo consiste em analisar a relevância de *hashtags* contidas em *tweets* e em descrições de perfis de usuários dessas duas amostras. O resultado dessa investigação, utilizando a base de dados americana, é reportado detalhadamente em [98]. A base de dados brasileira utilizada neste capítulo foi investigada em outro estudo conduzido por um dos autores deste trabalho, disponível em [95]. Com o objetivo de compreender também a relevância de *hashtags* na amostra presidencial brasileira, as análises propostas em [98] foram aplicadas a base de dados utilizada em [95]. As principais conclusões obtidas, utilizando essas duas bases de dados, são apresentadas neste capítulo.

Na Seção 5.1, são apresentadas as principais características das eleições presidenciais brasileira e americana. Na Seção 5.2, são apresentadas as bases de dados utilizadas nas análises. Na Seção 5.3, é investigada a relevância de *hashtags* em cenário de eleições.

5.1 *Background* das eleições brasileira e americana

5.1.1 Eleição brasileira

A eleição brasileira, realizada no ano de 2014, contou com a participação de 11 candidatos à Presidência da República no primeiro turno, realizada no dia cinco de outubro de 2014. Os dois candidatos mais votados nesse turno, Dilma Vana Rousseff e Aécio Neves da

Cunha, conhecidos mais popularmente durante as eleições apenas por Dilma Rousseff e Aécio Neves, foram os escolhidos pelos eleitores para disputarem no segundo turno, realizado no dia 26 de outubro de 2014, a vaga para presidente da República do Brasil. No mesmo dia da votação foi anunciada a vitória da candidata do Partido dos Trabalhadores (PT), Dilma Rousseff, com 51,4% dos votos em relação ao candidato do Partido da Social Democracia Brasileira (PSDB), Aécio Neves com 48,36%.

Durante o período de campanha eleitoral brasileiro, os *slogans* oficiais de campanha dos candidatos Aécio Neves e Dilma Rousseff foram "Muda Brasil" e "Mais mudanças, mais futuro", respectivamente. A corrida presidencial do ano de 2014 ficou marcada por diversos acontecimentos, como a morte de um dos candidatos, Eduardo Henrique Accioly Campos, vaias e xingamentos dirigidos à presidenciável Dilma Rousseff durante a realização da copa do mundo de futebol organizada no Brasil no mesmo ano, com palavras de ordem "Fora PT" e "Fora Dilma", a revelação da construção de um aeroporto dentro da fazenda de um tio do candidato Aécio Neves, acusado de utilizar dinheiro público para realização da obra, e diversos protestos realizados pelo Brasil.

5.1.2 Eleição americana

A primeira etapa da eleição americana para escolha do presidente dos EUA iniciou-se em 1 de fevereiro de 2016, quando os eleitores foram às urnas para escolher, de forma indireta, os pré-candidatos. Em julho do mesmo ano, na convenção nacional de cada partido, os candidatos de cada partido foram nomeados oficialmente. Os candidatos Donald John Trump e Hillary Diane Rodham Clinton foram escolhidos para representar os dois principais partidos americanos, Republicano e Democrata, respectivamente, na eleição presidencial. Ambos os candidatos tiveram três meses de campanha eleitoral para apresentar as suas propostas antes do dia das eleições, realizada no dia 8 de novembro de 2016. Durante esse período, os candidatos ficaram mais conhecidos apenas pelos nomes de Donald Trump e Hillary Clinton. Donald Trump, apesar de ter recebido uma quantidade inferior de votos nas urnas (47.01%), em relação a Hillary Clinton (48.03%), foi eleito presidente do EUA na 58ª eleição presidencial, pois obteve 306 dos 538 votos dos eleitores do Colégio Eleitoral, sendo 270 o número mínimo para vencer as eleições.

Durante o período de campanha eleitoral americano, os *slogans* oficiais de campanha dos candidatos Donald Trump e Hillary Clinton foram "*Make America Great Again*" (*#MAGA*) e "*I'm With Her*" (*#ImWithHer*). Durante esse período, os candidatos estiveram envolvidos em várias polêmicas. Em relação a Donald Trump, temas como falta de

pagamento de impostos, construção de um muro na fronteira entre os EUA e o México e denúncias de assédio, foram um dos assuntos mais polêmicos sobre o candidato. Em relação a Hillary, os resultados das investigações do *Federal Bureau of Investigation* (FBI)[1] sobre o uso de uma conta de *email* privada e vazamentos de *emails* pelo *Wikileaks*[2], foram um dos mais comentados durante as eleições.

5.2 Bases de dados

As bases de dados utilizadas nas análises realizadas neste capítulo, para verificar a relevância de *hashtags*, disponíveis em *tweets* e em descrições de perfis de usuários, em cenários de eleições, foram coletadas a partir da *Search API do Twitter*[3]. Para ambas as eleições, foram coletados *tweets* postados nas últimas semanas que antecederam as eleições. O período de coleta da base de dados americana compreendeu os dias seis de outubro à sete de novembro de 2016 e, a base brasileira, entre os dias seis à 25 de outubro de 2014. Foram coletadas mensagens até um dia antes do dia de votação, para ambas as bases de dados, assim como em outros trabalhos [14, 102, 108]. A Tabela 5.1 apresenta uma visão geral das bases de dados.

Tabela 5.1: Visão geral das bases de dados brasileira e americana

Parâmetro	Volume (Brasil)	Volume (EUA)
Tweets	704.260 (35,92%)	430.529 (21,81%)
Retweets	1.255.925 (64,08%)	1.543.872 (78,19%)
Usuários	337.102	432.289

Além do *tweet*, foi aquisionado e armazenado na base de dados a descrição do perfil do usuário. O critério utilizado para coletar os dados, para ambas as eleições, foi o mesmo adotado em outros trabalhos [35, 71, 95, 102, 125], ou seja, foram aquisicionados *tweets* a partir do primeiro nome e/ou o sobrenome dos candidatos.

O termo "mensagem" é utilizado neste e no capítulo seguinte para representar um *tweet* ou *retweet*, e o termo "mensagem individual" para representar uma mensagem contendo no mínimo o primeiro nome, sobrenome e/ou nome de partido de um candidato, sem fazer menção ao seu oponente na mesma mensagem.

[1]Unidade de polícia do Departamento de Justiça dos EUA

[2]Organização sem fins lucrativos que divulga postagens de fontes anônimas, documentos, fotos e informações confidenciais, vazadas de governos ou empresas

[3]Disponível em: https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html

Em relação a coleta das descrições dos perfis dos usuários, as abordagens utilizadas variaram de uma base para outra. Na base de dados americana, o texto contido na descrição do perfil de um usuário foi coletado juntamente as mensagens postada por ele, diferentemente da base brasileira, onde apenas a última atualização da descrição, antes do dia da eleição, foi coletada.

5.3 Análise do conteúdo de mensagens e descrições

Nesta seção, *hashtags* disponíveis nas mensagens e nas descrições dos perfis dos usuários, armazenadas nas bases de dados brasileira e americana, foram investigadas com o objetivo de analisar a relevância desse tipo de informação em cenários de eleições.

5.3.1 Análise de mensagens

Nesta seção, são realizadas duas investigações com o objetivo de verificar a relevância de *hashtags* contidas em *tweets*. Na primeira, o objetivo consiste em verificar se o número de mensagens que demonstra preferência política com *hashtags*, em cenário eleitoral, é relevante. Na segunda, é analisado se *hashtags* contendo o primeiro nome e/ou sobrenome de candidatos são utilizadas com frequência em *tweets* políticos.

Para investigar o primeiro objetivo, toda mensagem da base de dados contendo qualquer sequência de caracteres alfanuméricos, além dos símbolos de *underline* (_) e traço (-), precedida do símbolo "#", foi selecionada. Para a base de dados americana, foram encontradas 22,5% das mensagens satisfazendo esse critério, isto é, contendo pelo menos uma *hashtag*. Para a base brasileira, esse índice foi praticamente o mesmo, ou seja, foi equivalente a 22,4%.

Para verificar se usuários manifestam preferência política a partir do uso de *hashtags*, foram selecionadas mensagens incluindo pelo menos uma *hashtag* contendo o nome do partido e/ou o primeiro nome/sobrenome de algum presidenciável. Para a base americana, a relevância desse grupo de mensagens foi igual a 8,2% e, para a brasileira, foi equivalente a 11,6%. Segundo [98], é possível aumentar esse índice ao adicionar o *slogan* oficial de campanha de candidatos nas análises. Nesse trabalho, por exemplo, o aumento obtido foi de aproximadamente 1,5%. Ao utilizar essa mesma abordagem para a base de dados brasileira, a relevância passou para 11,8%.

Durante o período de campanha eleitoral é comum o surgimento de novos *slogans* de

campanha e de expressões de apoio/repúdio sobre presidenciáveis. Esse conjunto de novas palavras é criado por apoiadores/opositores dos candidatos e por empresas contratadas para fazer a criação das peças publicitárias dos políticos. Com o objetivo de identificá-los, para ambas as eleições, foi utilizada a mesma abordagem reportada em [102], onde foi realizada uma consulta manual ao *trending topics* do *Twitter* para seleção desses novos termos. O resultado dessa pesquisa é apresentado na Tabela 5.2.

Tabela 5.2: *Slogans* de campanha e expressões políticas utilizadas durante períodos de campanha eleitoral

Eleição	Candidato	*Slogans* de campanha	Expressões de apoio/repúdio
Brasileira	Aécio	#AecioPresidente, #SouAecio, #Aecio45Confirma, #Aecio45PeloBrasil, #Aecio45, #AecioPelaMudanca	#AecioNever, #AecioPorto, coxinha(s), tucanalha(s), reacionario(s), reaça
	Dilma	#QueroDilmaTreze, #Dilma13, #MelhorComDilma13	#ForaPT, #ForaDilma, lulista, ptralha, petralha, esquerdista, lula, petista, ptista, lulinha
Americana	Trump	*#DrainTheSwamp*, *#FollowTheMoney*, *#VoteTrump*	*#NeverTrump*
	Hillary	#Hillary2016	*#Wikileaks*, *#SpiritCooking*, *#PodestaEmails*, *#NeverHillary*

Em relação aos *slogans* de campanha, foram identificadas, para ambas as eleições, *hashtags* contendo o primeiro nome do candidato associado a outras palavras/números (Ex: #Aecio45Confirma, #Aecio45, #QueroDilmaTreze e #Dilma13, *#Hillary2016*, *#VoteTrump* etc). Na eleição americana, além dos *slogans* oficiais, foram outros tais como *#DrainTheSwamp* e *#FollowTheMoney*. A respeito do grupo de expressões de apoio e repúdio, foram encontradas basicamente *hashtags* e termos contendo palavras relacionadas a denúncias ou outros fatos envolvendo os candidatos durante a carreira política/vida pessoal deles. Nesse grupo, os termos faziam referência direta aos candidatos, a partir dos nomes deles ou de seus partidos (Ex: #AecioNever, #ForaPT, *#NeverTrump* e *#NeverHillary*) ou indiretamente, a partir de outras expressões (Ex: esquerdista, reacionario, ptralha, *#Wikileaks*, *#SpiritingCooking* etc).

Os novos *slogans* e expressões de repúdio selecionadas foram adicionados ao conjunto anterior, composto pelo nome do partido, primeiro nome e/ou sobrenome do candidato e *slogans* oficiais de campanha, com o objetivo de verificar a relevância deles no aumento do percentual de mensagens que expressavam preferência política com *hashtags*. Para a base americana, o incremento obtido foi de 3,1%, totalizando 12,8% das mensagens contendo

essa caracaterística. Para a base brasileira, o aumento foi de 0,2%, resultando num volume total de 12% das mensagens da base de dados.

Para investigar o segundo objetivo supracitado no início dessa seção, a mesma metodologia apresentada anteriormente, para selecionar mensagens contendo *hashtags*, foi utilizada. Com o objetivo de analisar se *hashtags* utilizadas com maior frequência são compostas pelo primeiro nome e/ou sobrenome de candidatos, inicialmente foram identificadas todas as *hashtags* contidas nas mensagens. Na base de dados americana, foram encontradas 22.277 *hashtags* diferentes nas mensagens e, na base brasileira, 15.715. Em seguida, as *hashtags* contidas nas mensagens individuais de cada candidado foram filtradas e o percentual de uso de cada uma delas foi computado, a partir da Equação 5.1 a seguir:

$$PU\_hash(x) = \frac{FU\_hash(x)}{\sum_{h=1}^{t} FU\_hash(h)} \tag{5.1}$$

O percentual de uso de uma *hashtag* "x" qualquer, representado por $PU\_hash(x)$, é calculado pela relação entre a frequência de uso dela, $FU\_hash(x)$, e o somatório das frequências de uso de todas as hashtags ($h = 1...t$), contidas nas mensagens individuais de determinado candidato, representado por $\sum_{h=1}^{t} FU\_hash(h)$. Por exemplo, na base de dados brasileira a *hashtag* utilizada com maior frequência nas mensagens individuais sobre o candidato Aécio Neves foi "#Aecio45", com frequência de uso igual a 33.971, ou seja, $FU\_hash('\#Aecio45') = 33.971$. Nas mensagens individuais sobre esse candidato, a frequência de uso de todas as *hashtags* foi equivalente a 208.677, ou seja, $\sum_{h=1}^{15.715} FU\_hash(h) = 208.677$. Portanto, o percentual de uso da *hashtag* "#Aecio45" nas mensagens individuais sobre o candidato Aécio Neves foi igual a 16,3%, isto é, $PU\_hash('\#Aecio45') = 16,3\%$. A Tabela 5.3 apresenta o conjunto das 15 *hashtags* utilizadas com maior frequência nas mensagens individuais sobre cada um dos candidatos presidenciais, tanto para a base de dados brasileira quanto para a americana, e o percentual de uso de cada uma delas. No Apêndice A, são apresentados alguns exemplos de *tweets* contendo pelo menos uma das *hashtags* apresentadas na Tabela 5.3.

Na base de dados brasileira, a *hashtag* utilizada com maior frequência nas mensagens individuais sobre o candidato Aécio Neves foi #Aecio45 e nas mensagens individuais sobre a candidata Dilma foi #Dilma, com 16,3% e 8,4% de percentual de uso, respectivamente. Na base americana, a *hashtag* mais frequente sobre os candidatos Trump e Hillary, foi *#DrainTheSwamp* (27,2%) e *#Wikileaks* (5,4%), respectivamente. Ao analisar o conjunto das 15 *hashtags* mais frequentes da base brasileira, conclui-se que usuários brasileiros tem

Tabela 5.3: *Hashtags* utilizadas com maior frequência em mensagens individuais sobre candidatos

Base Brasileira			
Hashtag sobre Aécio	*PU_hash*	*Hashtag* sobre Dilma	*PU_hash*
#Aecio45	16,3%	#DILMA	8,4%
#Aecio	5,0%	#QueroDilmaTreze	5,9%
#Aecio45PeloBrasil	4,0%	#13rasilTodoComDilma	3,1%
#AecioPeloBR45IL	3,4%	#desesperodaveja	2,5%
#VotoAecioPeloBR45IL	2,2%	#Dilma13	2,5%
#MudaBrasil	2,1%	#MelhorcomDilma13	2,3%
#AecioNever	2,0%	#SomosTodosDilma	2,0%
#Eleicoes2014	2,0%	#eleicoes2014	1,9%
#AecioPelaMudanca	1,8%	#MenosOdioMaisNordeste	1,6%
#Aecioporto	1,6%	#ForaDilma	1,6%
#Aecio45Confirma	1,3%	#PretonoBranco	1,3%
#EmTodoBrasilAecio45	1,3%	#Dilma13PraVencer	1,3%
#45AecioConfirma	1,2%	#MenosOdioMaisDilma	1,3%
#debatenosbt	1,2%	#ForaPT	1,3%
#MenosOdioMaisNordeste	1,1%	#Dilma13MaisNordeste	1,2%
Base Americana			
Hashtag sobre Trump	*PU_hash*	*Hashtag* sobre Hillary	*PU_hash*
#*DrainTheSwamp*	27,2%	#*Wikileaks*	5,4%
#Trump	5,3%	#Hillary	5,1%
#*NeverTrump*	5,3%	#*ImWitHer*	3,2%
#*MAGA*	4,4%	#Clinton	2,5%
#DonaldTrump	3,0%	#*MAGA*	2,5%
#*ImWithHer*	1,8%	#*HillaryClinton*	2,1%
#*debate*	1,7%	#*debate*	1,9%
#*FollowTheMoney*	1,7%	#*PodestaEmails*	1,7%
#*Breaking*	1,6%	#*tcot*	1,6%
#*Election2016*	1,4%	#*SpiritCooking*	1,3%
#*VOTETRUMP*	1,3%	#*debatenight*	1,3%
#*AmericaFirst*	1,1%	#AnthonyWeiner	1,2%
#*USA*	1,0%	#*NeverHillary*	1,1%
#*TrumpTrain*	1,0%	#*CrookedHillary*	1,0%
#*BananaRepublic*	0,8%	#*TruePundit*	0,8%

o hábito de utilizar o primeiro nome de candidatos em *hashtags* com maior frequência do que usuários americanos. No conjunto apresentado na Tabela 5.3, 73,4% das 15 *hashtags* mais frequentes fizeram referência ao candidato Aécio Neves nas *hashtags*, a partir do primeiro nome dele e, em relação a candidata Dilma Rousseff, esse índice foi equivalante a 66,7%. Já na base americana esse índice foi igual a 33,4% e 26,7% para os candidatos Donald Trump e Hillary Clinton, respectivamente.

Além do primeiro nome e/ou sobrenome de candidato, as *hashtags* apresentadas na

Tabela 5.3 são compostas basicamente por palavras contendo *slogans* de campanha de candidatos (Ex: #Aecio45, #MudaBrasil, #*MAGA*, #*ImWithHer*, #*DrainTheSwamp* etc), informações sobre acontecimentos relacionados ao período eleitoral (Ex: #DebateNaGlobo, #eleicoes2014, #*Breaking*, #*news*, etc) e expressões de repúdio relacionadas aos presidenciáveis (Ex: #AecioPorto, #ForaPT, #*NeverTrump*, #*Wikileaks*, #*SpiritCooking*, #*NeverHillary*, etc). O percentual de uso de cada uma das categorias de *hashtags* identificadas nas análises é apresentado na Tabela 5.4.

Tabela 5.4: Percentuais de uso dos grupos de *hashtags* identificados em mensagens individuais sobre candidatos

	Base brasileira		Base Americana	
Categoria da *hashtag*	Aécio	Dilma	Trump	Hillary
# Primeiro nome e/ou sobrenome	60,0%	60,0%	20,0%	26,7%
# Expressões de repúdio	13,3%	13,3%	6,7%	26,7%
# *Slogans* de campanha	6,7%	0,0%	40,0%	13,3%
# Outros assuntos	20,0%	26,7,0%	33,3%	33,3%

Nesta seção, foi investigada a relevância de *hashtags* disponíveis em *tweets* postados durante períodos de campanha eleitoral. Primeiro, chegou-se a conclusão que o número de mensagens contendo *hashtags* é frequente nesse cenário. Para ambas as bases analisadas, o percentual de mensagens contendo essa característica foi de aproximadamente 23%. Em seguida, foi proposta uma abordagem para verificar se usuários expressam opinião política com *hashtags*. A presença de palavras contendo o primeiro nome/sobrenome do candidato, nome do partido político, *slogans* oficiais e não-oficiais de campanha ou expressões de repúdio em *hashtags*, foi um dos critérios utilizados para realizar essa classificação. As *hashtags* contendo pelo menos um (nenhum) desses termos comporão o conjunto das "*hashtags* políticas" ("*hashtags* não políticas"). O Apêndice B apresenta exemplos de *tweets* contendo *hashtags* políticas obtidas a partir das duas bases de dados. Os gráficos apresentados na Figura 5.1 mostram a distribuição de mensagens contendo cada um desses dois tipos de *hashtags*, identificadas em ambas as bases de dados.

Em relação ao grupo das 15 *hashtags* utilizadas com maior frequência, localizadas nas mensagens individuais sobre os candidatos, foram identificadas *hashtags* contendo *slogans* de campanha, primeiro nome e/ou sobrenome de candidato associado a outras palavras/números e expressões de repúdio. Os percentuais de uso de cada uma dessas categorias, para ambas as eleições, são apresentados nos gráficos da Figura 5.2.

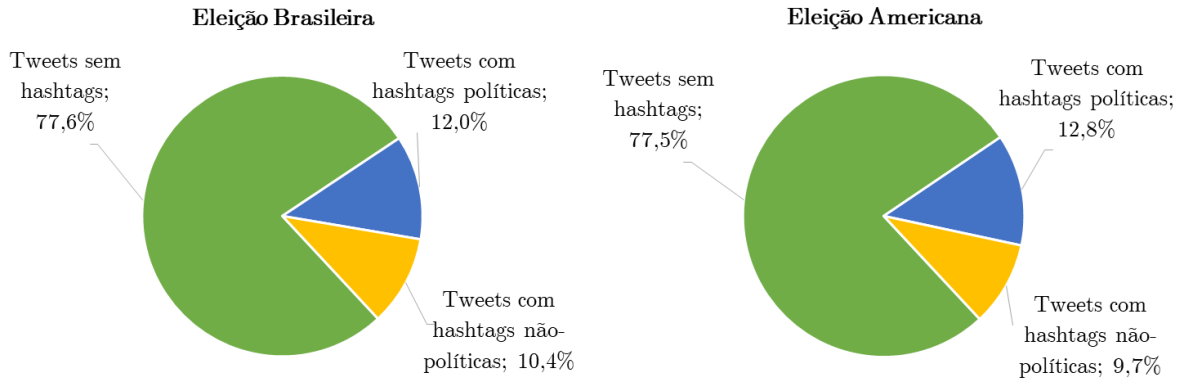

Figura 5.1: Representatividade de *tweets* contendo *hashtags* políticas/não-políticas nas eleições brasileira e americana

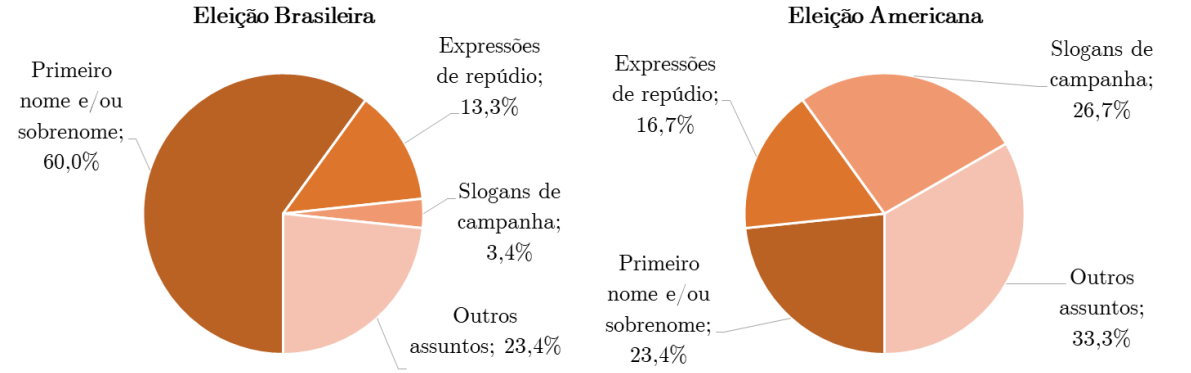

Figura 5.2: Representatividade das categorias de *hashtags* identificadas em mensagens individuais sobre candidatos

5.3.2 Análise de descrições de perfis de usuários

Nesta seção, são realizadas duas investigações com o objetivo de verificar a relevância de *hashtags* disponíveis em descrições de perfis de usuários em cenário de eleições, a partir da análise das bases de dados brasileira e americana. Na primeira investigação, o objetivo consiste em verificar se a posição política expressa por usuários em descrições de perfis é relevante. Na segunda, é analisado se o número de descrições de perfis de usuários que expressam preferência política com *hashtags* é representativa.

Para investigar o primeiro objetivo, foram selecionadas todas as descrições de perfis de usuários contendo as expressões de repúdio e os *slogans* de campanha identificados na seção anterior (Tabela 5.2), juntamente com o primeiro nome e/ou sobrenome, nome de partido e *slogans* oficiais de campanha dos candidatos. A última atualização da descrição do perfil de cada usuário, contendo alguma dessas expressões políticas, foi utilizada nas análises. Para as bases de dados americana e brasileira, foram identificadas 7,1% e

0,7% das descrições satisfazendo esse critério, respectivamente, conforme apresentado nos gráficos das Figuras 5.3 e 5.4.

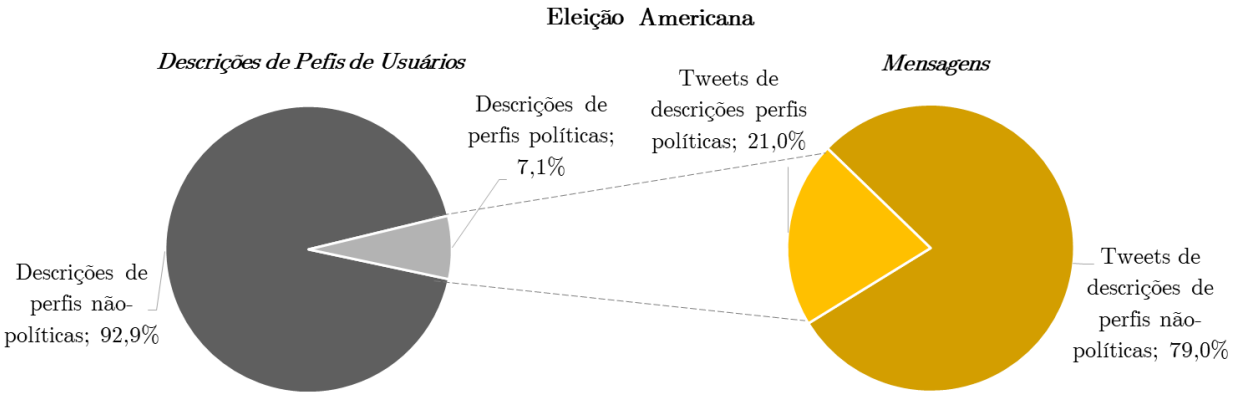

Figura 5.3: Representatividade do volume de *tweets* postados por usuários contendo descrição de perfil política, na eleição americana

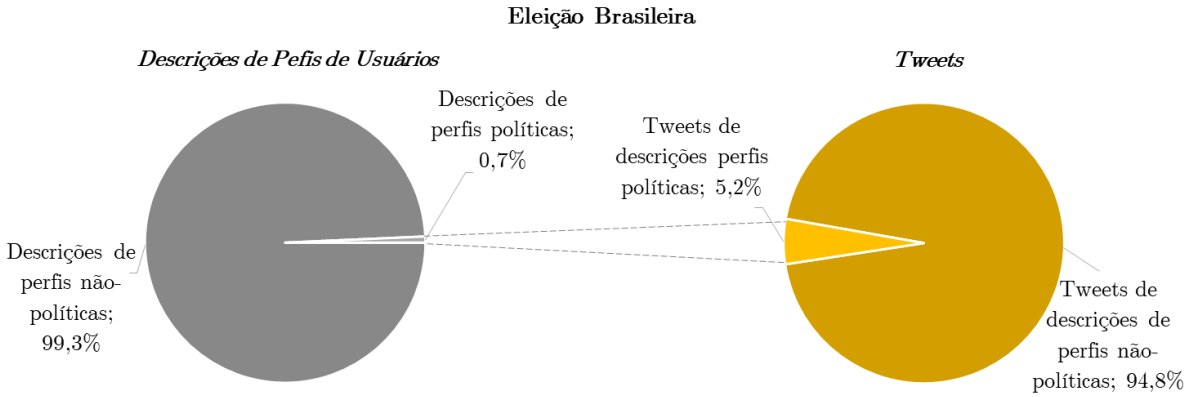

Figura 5.4: Representatividade do volume de *tweets* postados por usuários contendo descrição de perfil política, na eleição brasileira

As descrições de perfis de usuários contendo pelo menos uma (nenhuma) das expressões políticas supracitadas no parágrafo anterior, constituem as "descrições de perfis políticas" ("descrições de perfis não-políticas").

Com o objetivo de investigar a relevância de descrições de perfis políticas em cenários de eleições, foi analisada a representatividade do volume de mensagens postadas por esse grupo de usuários, utilizando as duas bases de dados. Chegou-se a conclusão que aproximadamente 21% de todas as mensagens da base de dados americana haviam sido publicadas por esse tipo de usuário, enquanto que esse mesmo índice para a base brasileira foi de 5,2%, conforme apresentado nos gráficos das Figuras 5.3 e 5.4. Acredita-se que essa diferença possa estar relacionada com o método utilizado para realizar a coleta das descrições dos perfis dos usuários para ambas as bases de dados, ou seja, enquanto que

a descrição de determinado usuário da base brasileira havia sido coletada apenas uma única vez, na americana a descrição de um usuário era coletada sempre que uma nova mensagem era postada por ele.

Para investigar o segundo objetivo apresentado no início desta seção, que consiste em analisar se o número de descrições de perfis de usuários que expressam preferência política com *hashtags* é relevante, foram selecionadas, inicialmente, todas as descrições de perfis de usuários contendo pelo menos uma *hashtag*. Para a base americana, 12,5% das descrições filtradas atenderam a esse critério. Utilizando a base brasileira, esse mesmo índice foi correspondente a 1,9%. Em seguida, o conteúdo das *hashtags* localizadas nas descrições dos perfis dos usuários foram analisadas, com o ojetivo de identificar a presença de alguma expressão política nelas. Utilizando a base de dados americana, 30% das descrições contendo pelo menos uma *hashtag* continham nessa o primeiro nome e/ou sobrenome, *slogan* de campanha, nome de partido e alguma expressão de repúdio sobre candidatos. Para a base brasileira, esse percentual foi de aproximadamente 3,2%. Os gráficos apresentados nas Figuras 5.5 e 5.6 apresentam as distribuições desses percentuais para ambas as bases de dados.

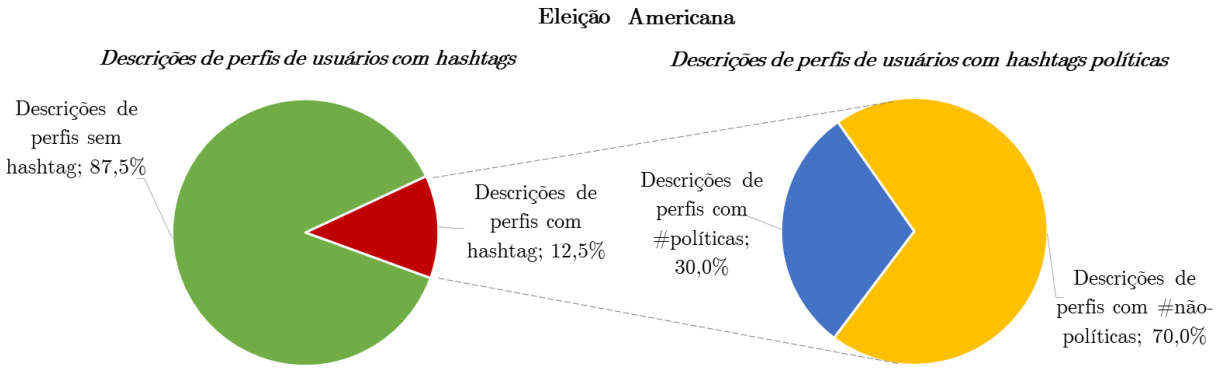

Figura 5.5: Distribuição do percentual de descrições de perfis contendo *hashtags* políticas/não-políticas para a eleição americana

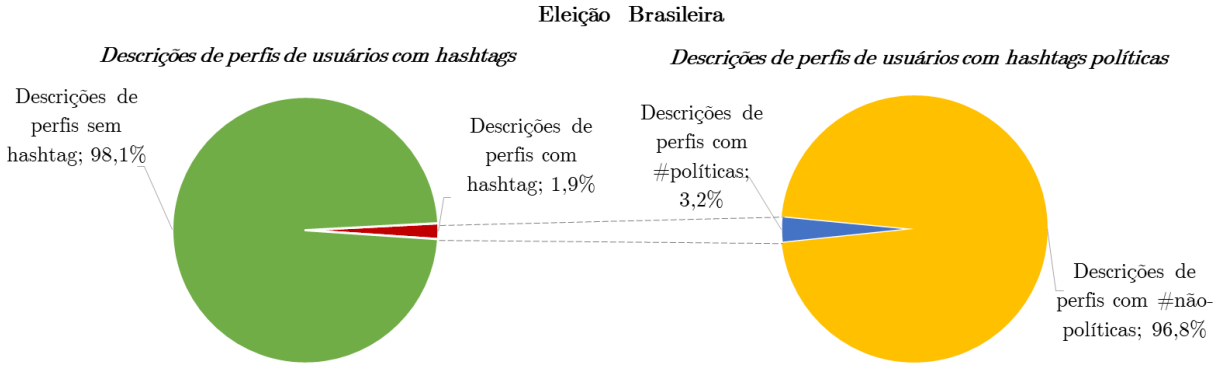

Figura 5.6: Distribuição do percentual de descrições de perfis contendo *hashtags* políticas/não-políticas para a eleição brasileira

O volume de mensagens publicadas por usuários que expressavam preferência política em seus perfis, a partir do uso de *hashtags*, também foi investigado. Primeiro, analisou-se o volume de mensagens postadas por usuários cujas descrições de perfis haviam pelo menos uma *hashatg*. Chegou-se a conclusão que o percentual de mensagens satisfazendo esse critério foi igual a 19,6% e 3,1%, para as bases americana e brasileira, respectivamente. Ao considerar apenas as *hashtags* políticas nessa análise, o volume encontrado para as bases americana e brasileira foram iguais a 11,2% e 0,7%, respectivamente. Os gráficos das Figuras 5.7 e 5.8 apresentam as distribuições dos percentuais de *tweets* postados para cada um dos tipos de descrições de perfis de usuários, sem *hashtag* e com *hashtag* política/não-política, para ambas as eleições.

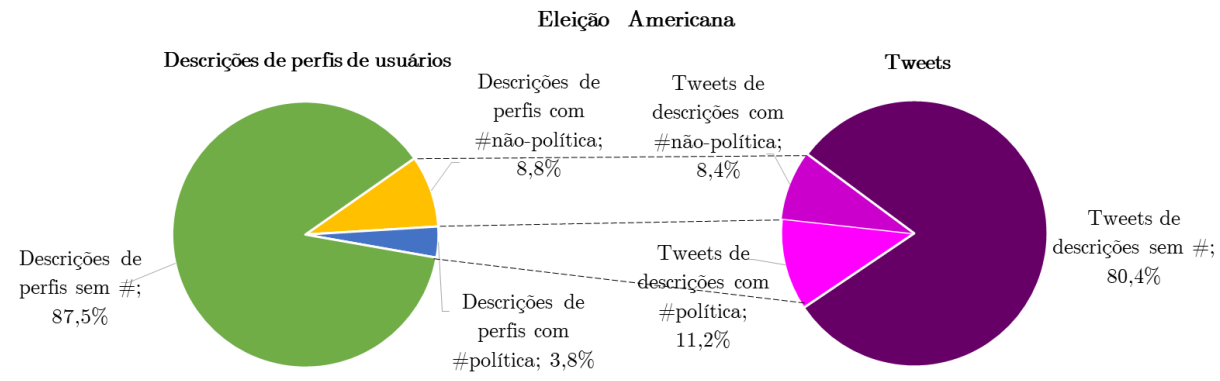

Figura 5.7: Distribuição do percentual de descrições de perfis contendo *hashtags* políticas/não-políticas para a eleição brasileira

Com o objetivo de identificar as categorias de *hashtags* utilizadas com maior frequência nas descrições dos perfis dos usuários, foi utilizada a mesma metodologia apresentada na seção anterior. Nas bases de dados americana e brasileira, foram encontradas 9.132 e 6.973

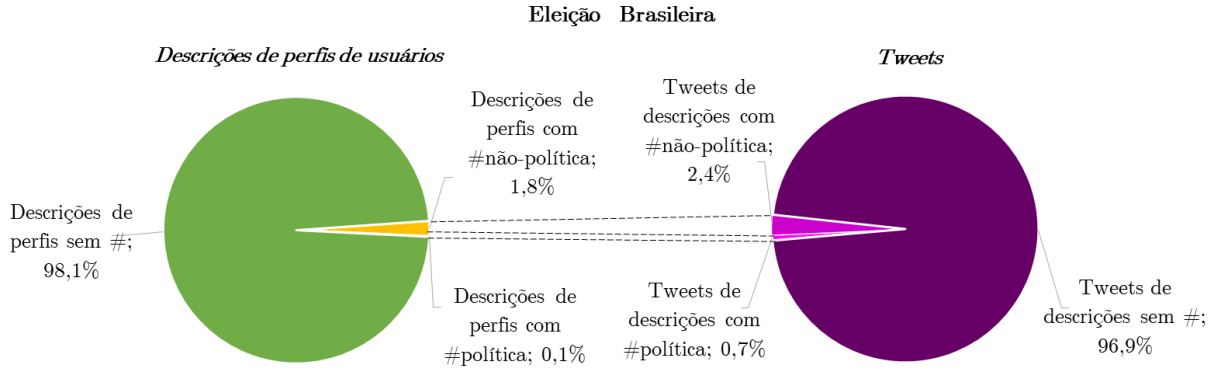

Figura 5.8: Distribuição do percentual de descrições de perfis contendo *hashtags* políticas/não-políticas para a eleição brasileira

hashtags diferentes, respectivamente. A Tabela 5.5 apresenta o conjunto das 15 *hashtags* de maior percentual de uso identificadas nas descrições dos perfis dos usuários, para as bases brasileira e americana, respectivamente.

Na base de dados brasileira, as *hashtags* com maior frequência de uso utilizadas nas descrições dos perfis dos usuários, sobre os candidatos Aécio Neves e Dilma Rousseff, eram compostas pelo *slogan* oficial de campanha (#MudaBrasil: 19,5%) e por palavras de ordem (#ForaPT: 15,1%), respectivamente. Na base americana, as *hashtags* mais frequentes sobre os candidatos Trump e Hillary, foram *#MAGA* (11,3%) e *#ImWithHer* (21,5%), respectivamente. Conforme pode ser visto na Tabela 5.5, a categoria de *hashtag* utilizada com mais frequência nas descrições dos perfis dos usuários, para ambas as eleições, é composta pelo *slogan* oficial de campanha do candidato, com exceção da análise realizada para a candidata Dilma, cuja *hashtag* mais frequente é composta por palavras de ordem contra a candidata (expressão de repúdio).

Em relação ao grupo das 15 *hashtags* utilizadas com maior frequência nas descrições dos perfis dos usuários (Tabela 5.5), sobre cada candidato individualmente, foram obtidas as mesmas categorias de *hashtags* identificadas na seção anterior (quando *tweets* foram utilizados nas análises): *slogans* de campanha, primeiro nome e/ou sobrenome do candidato associado a palavras/números e expressões de repúdio. No caso da base de dados brasileira, foi identificada ainda uma nova categoria, composta pelo nome do partido político do candidato. A distribuição do percentual de uso de cada uma das categorias identificadas nas *hashtags* com o maior percentual de uso é apresentada na Tabela 5.6.

Nesta seção, a representatividade de informações contidas em descrições de perfis de usuários do *Twitter*, foi investigada com o objetivo principal de verificar se a quantidade de informações contidas nesse espaço da conta do usuário é relevante em cenário de eleições.

Tabela 5.5: *Hashtags* com maior frequência de uso identificadas nas descrições dos perfis dos usuários brasileiros e americanos

Base Brasileira			
Hashtag sobre Aécio	*PU_hash*	*Hashtag* sobre Dilma	*PU_hash*
#MudaBrasil	19,5%	#ForaPT	15,1%
#PSDB	2,7%	#ForaDilma	5,7%
#EuVouDeAecio	1,8%	#Dilma13	2,1%
#antiPT	1,8%	#DilmadeNOVO	1,8%
#FaltouAgua	1,8%	#DilmaPresidente	1,3%
#aecio45	1,8%	#PT	1,3%
#45	0,9%	#Petralha	1,0%
#AecioEmTodoBrasil	0,9%	#CRISEnaPF	1,0%
#AgoraEAecioBrasil	0,9%	#Petista	1,0%
#AutoModelismo	0,9%	#Dilma	0,8%
#BRASIL	0,9%	#DilmaReeleita	0,8%
#Corinthiana	0,9%	#DeclarandoVOTOemDilma	0,5%
#Crista	0,9%	#gentediferenciada	0,5%
#EternoAmor	0,9%	#MaisMudancasMaisFuturo	0,5%
#forcadonovo	0,9%	#ReformaPolitica	0,5%
Base Americana			
Hashtag sobre Trump	*PU_hash*	*Hashtag* sobre Hillary	
#*MAGA*	11,3%	#*ImWithHer*	21,5%
#Trump2016	6,9%	#*NeverHillary*	6,2%
#Trump	3,7%	#Hillary2016	2,8%
#*MakeAmericaGreatAgain*	3,1%	#*UniteBlue*	2,1%
#*nevertrump*	3,0%	#*StrongerTogether*	1,7%
#*TrumpTrain*	2,9%	#*blacklivesmatter*	1,2%
#*TrumpPence16*	2,4%	#ClintonKaine2016	0,9%
#2A	2,4%	#2A	0,8%
#*TrumpPence2016*	1,9%	#*VoteBlue*	0,7%
#*AmericaFirst*	1,2%	#ClintonKaine	0,6%
#NRA	1,2%	#*HillaryForPrison*	0,6%
#TCOT	1,0%	#HillaryClinton	0,6%
#*DrainTheSwamp*	1,0%	#*Hillary*	0,6%
#*conservative*	0,9%	#LGBT	0,5%
#*Deplorable*	0,7%	#*GunSense*	0,5%

A partir das bases de dados utilizadas nas análises, chegou-se a conclusão que usuários têm o hábito de expressar opinião política não somente a partir de *tweets*, mas também a partir de informações localizadas nas descrições de seus perfis. Verificou-se também que, em média, 13,1% das mensagens foram publicadas por usuários manifestando algum tipo de expressão política, a partir de seus perfis.

Neste capítulo, foi apresentado um estudo para analisar a relevância de *hashtags* presentes em *tweets* e em descrições de perfis de usuários, no cenários de eleições, a partir de

Tabela 5.6: Percentual de uso dos grupos de *hashtags* mais frequentes obtidas a partir de descrições de perfis de usuários

	Base brasileira		Base Americana	
Categoria da *hashtag*	Aécio	Dilma	Trump	Hillary
# Primeiro nome e/ou sobrenome	26,6%	40,0%	26,7%	33,4%
# Expressão de repúdio	6,7%	20,0%	6,7%	13,3%
# *Slogan* de campanha	6,7%	6,7%	33,3%	13,3%
# Partido Político	13,3%	6,7%	0,0%	0,0%
# Outros assuntos	46,7%	26,6%	33,3%	40,0%

duas bases de dados. No próximo capítulo, as duas últimas etapas do modelo proposto nesta tese serão utilizadas para investigar a contribuição das informações analisadas neste capítulo na melhoria do desempenho de algoritmos de aprendizado de máquina.

Capítulo 6

Análise experimental

Neste capítulo, um conjunto de experimentos computacionais, utilizando as bases de dados brasileira e americana, investigadas no capítulo anterior, foi realizado com o objetivo de avaliar o modelo proposto. Na Seção 6.1, são apresentados os *datasets* obtidos na primeira fase do modelo, as técnicas de pré-processamento definidas para limpeza dos dados e os algoritmos de aprendizado de máquina supervisionados utilizados na etapa de análise de dados. Na Seção 6.2, são apresentados os experimentos propostos e os resultados computacionais obtidos.

6.1 Configurações do modelo

6.1.1 *Datasets*

Nesta seção, são apresentadas as abordagens utilizadas para obtenção dos *datasets* $TweetsAL$, $HashPT$, $HashPDF$ e $DescPU$ do modelo proposto, detalhados no Capítulo 4. Para obtenção destes *datasets,* foram utilizadas as bases de dados brasileira e americana e os resultados das análises apresentados no capítulo anterior.

A abordagem utilizada para obtenção do *dataset* $TweetsAL$ de cada eleição, consistiu numa escolha aleatória de *tweets* respeitando a distribuição de mensagens contendo *hashtags* políticas e não-políticas, conforme percentuais apresentados no capítulo anterior. Para as bases americana e brasileira, foram selecionadas 3.996 e 3.720 mensagens aleatórias, respectivamente. Em seguida, as mensagens contidas no *dataset* $TweetsAL$, de cada base de dados, foram rotuladas manualmente por um conjunto de indivíduos. Em relação a base brasileira, as amostras foram rotuladas nas classes Pró-Dilma (PD), Pró-Aécio (PA) e em N (Neutro), por 18 voluntários. Na base americana, as amostras

foram rotuladas em Pró-Trump (PT), Pró-Hillary (PH) e N, por um conjunto de 14 pessoas. A Tabela 6.1 apresenta a distribuição de classes dos *datasets* $TweetsAL$ de ambas as eleições.

Tabela 6.1: Distribuição de classes das amostras americana e brasileira

Eleição Americana		Eleição Brasileira	
Classe	Volume	Classe	Volume
PT	2.355 (58,9%)	PA	1.374 (36,9%)
PH	873 (21,8%)	PD	1.374 (36,9%)
N	768 (19,3%)	N	972 (26,2%)

Ao comparar a distribuição de classes das amostras americana e brasileira (Tabela 6.1), com o resultado final das eleições presidenciais (Seção 5.1), conclui-se que o comportamento de usuários brasileiros no *Twitter* se aproximou de maneira mais homogênea com o comportamento dos eleitores brasileiros nas urnas. Já, no caso da eleição americana, a diferença entre a distribuição de classes de ambos os presidenciáveis, com o resultado final das eleições, foi mais discrepante. Por esse fato, conclui-se que a ação de usuários americanos no *Twitter* a favor do candidato Donald Trump foi maior, quando comparado com o percentual de classes a favor da candidata Hillary Clinton.

Para obtenção do *dataset* $HashPT$, isto é, do conjunto de *hashtags* políticas contidas em *tweets*, a abordagem utilizada para selecioná-las consistiu em escolher *hashtags* contendo o primeiro nome e/ou sobrenome, *slogans* oficiais e não-oficiais de campanha e expressões de repúdio sobre os candidatos. Esses termos foram escolhidos, pois segundo as análises realizadas no capítulo anterior, eles representam as categorias de *hashtags* utilizadas com maior frequência nas mensagens postadas em cenários de eleições. Basicamente, as *hashtags* políticas apresentadas na Tabela 5.3 foram utilizadas para composição desse *dataset*.

Neste trabalho, assume-se a hipótese de que *hashtags* compostas apenas pelo nome do candidato, disponíveis em *tweets* e em descrições de perfis de usuários, postadas durante períodos de campanha eleitoral, não são utilizadas para expressar opinião sobre os políticos. Por este motivo, as *hashtags* utilizadas com maior frequência de uso, apresentadas no capítulo anterior, #Aecio, #Dilma, #Trump, #DonaldTrump, #Hillary e #Clinton, não foram selecionadas para composição do *dataset* $HashPT$. Para ilustrar essa hipótese, foi selecionada do *dataset* $TweetsAL$, de cada eleição, uma mensagem aleatória contendo pelo menos uma *hashtag* composta pelo primeiro nome e/ou sobrenome de algum candidato. O *tweet*[1] "*#Trump and #Wikileaks will expose FBI and DOJ #Hillary cover up*

[1] O endereço eletrônico contido no *tweet* do exemplo foi substituído pelo termo URL

URL please SHARE" (#Trump e *#Wikileaks* irão expor o FBI e DOJ #Hillary cobertura em URL por favor, compartilhe), obtido da amostra americana, faz referência aos candidatos Donald Trump e Hillary Clinton, a partir das *hashtags* #Trump e #Hillary, respectivamente. Em relação a amostra brasileira, o autor do *tweet* "#Datafolha #Ibope: Aprovação de #Dilma e reprovação de #Aécio sobem quatro pontos URL #MenosOdioMaisDilma .....", faz referência aos candidatos Dilma Rousseff e Aécio Neves, a partir das *hashtags* #Dilma e #Aécio, respectivamente. Conforme apresentado nesses dois exemplos, as *hashtags* contendo apenas os nomes dos candidatos (primeiro nome e/ou sobrenome) foram utilizadas para fazer referência aos políticos e não para expressar qualquer opinião sobre eles nos *tweets*.

O *dataset* $HashPDF$, que consiste no conjunto de *hashtags* políticas localizadas nas descrições dos perfis dos usuários, foi obtido a partir da mesma estratégia utilizada para obtenção do *dataset* $HashPT$.

Para obtenção do último *dataset*, $DescPU$, para cada base de dados, foi selecionada a última atualização da descrição do perfil do usuário contendo pelo menos uma das seguintes expressões políticas: primeiro nome/sobrenome do candidato, *slogan* de campanha ou alguma expressão de repúdio sobre os políticos. Esses termos foram obtidos a partir das Tabelas 5.2, 5.3 e 5.5.

6.1.2 Conjuntos de pré-processamento

Na etapa de limpeza de dados do modelo proposto, o *dataset* $TweetsPP$ de cada base de dados (brasileira e americana) é obtido a partir do seu *dataset* $TweetsAL$ correspondente. Conforme apresentado no Capítulo 4, a contribuição do uso das técnicas de remoção de *stopwords* e *stemming*, na fase de pré-processamento do modelo, é investigada na melhoria do desempenho de algoritmos de aprendizado de máquina supervisionado, utilizados para classificar o sentimento de mensagens do cenário eleitoral.

Em [30], os autores propuseram o conceito de camadas de pré-processamento incremental para avaliar o melhor conjunto de técnicas de pré-processamento utilizando diversas bases de dados. Segundo os autores, foram utilizadas quatro camadas de pré-processamento. A cada camada, novas técnicas de pré-processamento eram inseridas e acumuladas com as técnicas da camada anterior. A desvantagem dessa abordagem consiste em não saber a contribuição individual das técnicas contidas em cada camada. Neste trabalho, é utilizado o conceito de Conjunto de Pré-processamento (CPP), onde é possível identificar o ganho individual de cada grupo de técnicas de pré-processamento. Os

três conjuntos, para avaliar a contribuição das técnicas de pré-processamento do modelo proposto nesta tese, são apresentados a seguir.

- **Conjunto de Pré-processamento 1 (CPP1)**: neste conjunto, são utilizadas as seguintes técnicas: conversão de letras, remoção de *URL's*, remoção de estruturas do *Twitter*, remoção de pontuações e tokenização;

- **Conjunto de Pré-processamento 2 (CPP2)**: este conjunto é composto apenas pela técnica de remoção de *stopwords*. A lista de *stopwords*[2] utilizada neste trabalho, para limpar informações indesejadas dos *tweets* da base americana, é a mesma utilizada em outros estudos [1, 30, 32, 47]. O mesmo repositório de onde foi obtida a lista de *stopwords* do idioma inglês foi acessado para obtenção da lista de *stopwords*[3] utilizada para limpar as mensagens da amostra brasileira; e

- **Conjunto de Pré-processamento 3 (CPP3)**: este conjunto é composto unicamente pela técnica de *stemming*. Implementações dos algoritmos *Lovins* e *Snowball*, para analisar palavras no idioma inglês, são facilmente encontradas na literatura [84, 85]. Já a versão do algoritmo *Lovins* para analisar palavras no idioma português não é encontrada com facilidade. Por esse motivo, nos experimentos apresentados na próxima seção, é investigada a contribuição da técnica de *stemming* no *dataset* brasileiro (BR), a partir apenas do uso do algoritmo *Snowball*. Neste trabalho, foi utilizado o *Natural Language Tookit* (NTLK)[4] para avaliar o desempenho dessa técnica.

6.1.3 Algoritmos de aprendizado de máquina

Os algoritmos de aprendizado de máquina supervisionado, apresentados na Seção 2.4, NB, SVM e MNB, são utilizados nos experimentos propostos neste capítulo, para avaliação do modelo proposto, assim como em outros trabalhos encontrados na literatura sobre mineração de opiniões de *tweets* políticos [3, 10, 14, 31, 71, 77, 86, 95, 102]. Neste trabalho, foram utilizadas as implementações desses algoritmos disponíveis na ferramenta *Weka*[5] .

[2]Lista de *stopwords* do idioma inglês: http://snowball.tartarus.org/algorithms/english/stop.txt
[3]Lista de *stopwords* do idioma português: http://snowball.tartarus.org/algorithms/portuguese/stop.txt
[4]https://www.nltk.org/
[5]Disponível em http://www.cs.waikato.ac.nz/ml/weka/

6.2 Avaliação do modelo proposto

Nesta seção, são apresentados quatro experimentos com o objetivo de avaliar o modelo proposto. No Experimento I, são analisadas as contribuições das técnicas de remoção de *stopwords* e *stemming* na fase de pré-processamento. As melhores configurações obtidas nesse experimento são utilizadas como *baseline* nos experimentos seguintes. No Experimento II, são avaliadas as contribuições das duas categorias de *hashtags* propostas nesta tese, as políticas e as não-políticas, contidas nas mensagens das amostras de cada eleição, e do atributo $TPSB$. No Experimento III, são discutidas as contribuições das mesmas categorias de *hashtags*, porém contidas nas descrições dos perfis dos usuários, e do atributo $DPSB$. No Experimento IV, é investigada a contribuição em conjunto das *hashtags* contidas nos *tweets* e nas descrições dos perfis dos usuários.

6.2.1 Experimento I: Avaliação dos conjuntos de pré-processamento

Inicialmente, a análise do sentimento das mensagens contidas no *dataset* $TweetsAL$, de cada eleição, foi realizada utilizando apenas as técnicas do CPP1. Posteriormente, foi avaliada a contribuição da técnica de remoção de *stopwords* (CPP2) em conjunto com as técnicas do CPP1 (CPP1 + CPP2). Logo após, foi avaliada a contribuição dos algoritmos de *stemming Lovins* e *Snowball* (CPP3), em conjunto com as técnicas do CPP1 (CPP1 + CPP3). Por fim, foi analisada a contribuição das técnicas contidas em CPP1, CPP2 e CPP3, em conjunto (CPP1 + CPP2 + CPP3). As acurácias obtidas pelos classificadores em cada um destes cenários são apresentadas na Tabela 6.2. Os algoritmos *Lovins* e *Snowball* são representados nesta tabela a partir das nomeclaturas CPP3(L) e CPP3(S), respectivamente.

Conforme apresentado na Tabela 6.2, a técnica de remoção de *stopwords* (CPP1+CPP2) não foi capaz de melhorar as acurácias dos algoritmos em todas as configurações analisadas. Ao considerar apenas os *tweets* da amostra americana, conclui-se que o uso desta técnica, na fase de limpeza de dados, foi capaz de melhorar o desempenho de todos os classificadores, porém em apenas 50% das configurações analisadas (unigrama/bigrama) as acurácias dos algoritmos *NB*, *MNB* e *SVM* foram incrementadas. Ao considerar o *dataset* brasileiro nas análises, verificou-se que esta técnica foi capaz de aumentar o desempenho de apenas 33% das configurações, isto é, apenas dos algoritmos *NB* (bigrama) e *SVM* (unigrama). Ao analisar a melhoria do desempenho de todos os algoritmos, independentemente do *dataset* utilizado, verificou-se que esta técnica foi capaz de melhorar

Tabela 6.2: Contribuições das técnicas de remoção de *stopwords* e *stemming*

		NB		*MNB*		*SVM*	
	Conjunto de PP	*unigram*	*bigram*	*unigram*	*bigram*	*unigram*	*bigram*
Dataset EUA	CPP1	62,9%	59,8%	68,9%	67,8%	63,8%	65,8%
	CPP1+CPP2	61,3%	59,9%	68,7%	68,7%	64,0%	64,3%
	CPP1+CPP3(L)	63,1%	60,0%	68,3%	68,6%	63,4%	65,7%
	CPP1+CPP3(S)	63,5%	59,7%	68,7%	68,4%	64,2%	65,5%
	CPP1+CPP2+CPP3(L)	61,8%	60,6%	68,3%	69,9%	64,1%	64,7%
	CPP1+CPP2+CPP3(S)	62,1%	60,5%	68,3%	70,0%	64,6%	64,7%
Dataset BR	CPP1	49,9%	49,9%	58,2%	63,3%	55,9%	60,2%
	CPP1+CPP2	49,9%	53,0%	57,6%	56,5%	56,7%	58,5%
	CPP1+CPP3(L)	-	-	-	-	-	-
	CPP1+CPP3(S)	49,5%	49,9%	56,5%	61,8%	55,1%	60,1%
	CPP1+CPP2+CPP3(L)	-	-	-	-	-	-
	CPP1+CPP2+CPP3(S)	48,6%	53,8%	56,2%	56,9%	55%	58,3%

a eficiência dos algoritmos *NB* e *SVM*, a partir dos formatos bigrama e unigrama, respectivamente. Os incrementos médios obtidos nas acuráricas destes dois classificadores foram de 1,6% e 0,5%, respectivamente. Baseado nesta análise, conclui-se ainda que o uso indiscriminado desta técnica no cenário de eleições, antes de avaliar a sua contribuição, poderá influenciar no resultado final da classificação do sentimento deste tipo de mensagem. Este fato pode ser comprovado ao analisar o desempenho do algoritmo *MNB*, a partir da utilização do formato bigrama. No *dataset* brasileiro, por exemplo, verificou-se um decremento de 6,8% na acurácia desse classificador, ao utilizar esta técnica nas análises.

Neste primeiro experimento a contribuição da técnica de *stemming*, na melhoria do desempenho dos algoritmos, também foi analisada. Na Tabela 6.2, as acurácias obtidas pelos classificadores, ao utilizar esta técnica, é representada pela nomenclatura CPP1+CPP3. A partir desta tabela, conclui-se que a contribuição desta técnica, no cenário eleitoral, não é unânime. Ao considerar o *dataset* americano, conclui-se que ela contribuiu para incrementar as acurácias dos algoritmos *NB*, independente do formato (unigrama/bigrama), e dos algoritmos *MNB* e *SVM*, a partir dos formatos bigrama e unigrama, respectivamente. Em relação a amostra brasileira, verificou-se que não houve melhoria nas acurácias de nenhum algoritmo. Outro ponto que merece destaque, refere-se ao tipo de algoritmo de *stemming* utilizado nas análises. Para a amostra americana, a melhoria na acurácia do algoritmo *NB*, utilizando os formatos unigrama e bigrama, foi obtida a partir dos algoritmos de *stemming Snowball* e *Lovins*, respectivamente. Já as acurácias dos algoritmos *MNB* (bigrama) e *SVM* (unigrama) foram incrementadas, a partir dos algoritmos *Lovins* e *Snowball*, respectivamente.

A terceira e a última análise realizada no primeiro experimento consistiu em verificar se as duas técnicas, remoção de *stopwords* e *stemming*, são capazes de melhorar o desempenho de algoritmos de aprendizado de máquina, em conjunto. Conforme apresentado na Tabela 6.2, 50% das melhores acurácias obtidas neste experimento, utilizando a amostra americana, foram obtidas ao considerar estas duas técnicas simultaneamente na fase de pré-processamento. Em relação a base brasileira, houve incremento em apenas 16% dos casos analisados.

Discussão

Com base nos resultados obtidos neste experimento, pode-se concluir que a contribuição da técnica de remoção de *stopwords*, na fase de pré-processamento, deve ser investigada antes do seu uso, ao invés de ser utilizada fortuitamente, conforme encontrado em alguns trabalhos [86, 108] ou ignorada em outros [23, 36, 88, 95, 109, 114, 125]. De forma análoga, a técnica de *stemming* é pouco explorada no cenário político [77, 88] e, quando utilizada, a sua contribuição não é investigada [20]. Sugere-se também que, no cenário político seja realizada uma investigação individual da contribuição de cada uma destas técnicas em relação a cada algoritmo de aprendizado de máquina supervisionado utilizado nas análises, independentemente da forma de representação dos atributos utilizados aos classificadores.

Conforme apresentado no Capítulo 4, a abordagem utilizada pelo modelo proposto consiste em usar as configurações obtidas na fase de pré-processamento responsáveis pela obtenção das melhores acurácias dos algoritmos de aprendizado de máquina utilizados nas análises. Desta forma, a Tabela 6.3 apresenta as melhores acurácias obtidas no primeiro experimento e as configurações utilizadas para se chegar a esta conclusão. Os valores apresentados nesta tabela serão utilizados como *baseline* nos próximos experimentos.

Tabela 6.3: Melhores acurácias obtidas pelos classificadores ao utilizar os conjuntos CPP2 e CPP3

	Dataset EUA					
	NB		*MNB*		*SVM*	
	unigrama	*bigrama*	*unigrama*	*bigrama*	*unigrama*	*bigrama*
Baseline	63,5%	60,6%	68,9%	70,0%	64,6%	65,8%
CPP	1+3(S)	1+2+3(L)	1	1+2+3(S)	1+2+3(S)	1
	Dataset BR					
	NB		*MNB*		*SVM*	
Baseline	49,9%	53,8%	58,2%	63,3%	56,7%	60,2%
CPP	1	1+2+3(S)	1	1	1+2	1

6.2.2 Experimento II: Avaliação de *hashtags* em mensagens

Neste experimento, a contribuição de *hashtags* é investigada sob duas perspectivas diferentes. Na primeira, é analisada a contribuição das duas categorias de *hashtags*, as políticas e as não-políticas, na classificação do sentimento dos *tweets* das amostras brasileira e americana. Na segunda perspectiva, é avaliada a contribuição do atributo $TPSB$, proposto nesta tese e utilizadas nas análises, na melhoria do desempenho dos algoritmos de aprendizado de máquina supervisionado.

Conforme apresentado na Seção 4.3.1, o modelo proposto avalia a contribuição de *hashtags* contidas em *tweets*, a partir de três cenários distintos: (T1) avaliação da contribuição de *hashtags* não-políticas, (T2) avaliação da contribuição de *hashtags* políticas e (T3) avaliação, em conjunto, da contribuição de *hashtags* políticas e não-políticas. O resultado da classificação do sentimento dos *tweets* das amostras americana e brasileira, utilizando cada um desses cenários, é mostrado na Tabela 6.4.

Tabela 6.4: Contribuição de *hashtags* políticas e não-políticas contidas em *tweets*

Dataset EUA						
	NB		*MNB*		*SVM*	
Cenário	*unigram*	*bigram*	*unigram*	*bigram*	*unigram*	*bigram*
Baseline	63,5%	60,6%	68,9%	70,0%	64,6%	65,8%
T1	63,5%	**60,7%**	68,7%	69,8%	64,7%	66,2%
T2	64,3%	60,6%	**69,6%**	**70,2%**	65,2%	66,1%
T3	**64,4%**	**60,7%**	**69,6%**	**70,2%**	**65,6%**	**66,4%**
Dataset BR						
	NB		*MNB*		*SVM*	
Baseline	49,9%	53,8%	58,2%	63,3%	56,7%	60,2%
T1	50,1%	54,1%	58,6%	63,5%	55,9%	60,1%
T2	**52,3%**	54,5%	61,0%	**64,1%**	**59,0%**	**60,6%**
T3	**52,3%**	**54,6%**	**61,1%**	**64,1%**	58,6%	60,4%

Conforme pode ser visto na Tabela 6.4, a contribuição de *hashtags* não-políticas (T1), na melhoria do desempenho dos algoritmos *NB*, *MNB* e *SVM*, não foi unânime entre os dois *datasets* utilizados nas análises. Em relação ao *dataset* americano, as *hashtags* não-políticas melhoraram as acurácias dos algoritmos *NB* (unigrama) e *SVM* (unigrama e bigrama). Na amostra brasileira, elas não foram capazes de melhorar apenas a acurácia do algoritmo *SVM* (unigrama e bigrama). Ao analisar a contribuição das *hashtags* políticas (T2), conclui-se que elas tiveram um desempenho superior às *hashtags* não-políticas, pois elas conseguiram incrementar as acurácias de todos os classificadores, independentemente do *dataset* utilizado, com exceção da configuração *NB* (bigrama), cuja acurácia se manteve

inalterada ao utilizar a amostra americana.

Neste experimento, foi analisada também a contribuição das *hashtags* políticas e não-políticas (T3), em conjunto. Conforme apresentado na Tabela 6.4, verificou-se uma melhoria nas acurácias de todos os algoritmos ao manter esses dois tipos de *hashtags*, independente do formato de atributo utilizado (unigrama/bigrama). Porém, conforme apresentado nessa tabela, as melhores acurácias obtidas pelo algoritmos não ocorreram nesse cenário. No *dataset* americano, 50%, 33% e 16% dos melhores resultados foram obtidos nos cenários T3, T2 e T1, respectivamente. No *dataset* brasileiro, 33% e 67% das melhores acurácias foram obtidas nos cenários T3 e T2, respectivamente. Na Tabela 6.4, os melhores resultados obtidos neste experimento estão representados em negrito.

As melhores acurácias obtidas pelos algoritmos (Tabela 6.4), representados a seguir pela nomenclatura *TH*, foram comparadas com o *baseline* obtido no primeiro experimento com o objetivo de verificar o ganho obtido na acurácia de cada algoritmo. Um teste de hipótese foi realizado com a finalidade de verificar se o aumento obtido no desempenho dos algoritmos foi estatisticamente significante. Para o teste estatístico, estabeleceu-se, como hipótese nula (H_0), que não há diferença estatística entre amostra de *tweets* com e sem *hashtags*. Os testes *Test-t* [120] e *Wilcoxon* [133], ambos pareados, foram escolhidos para serem utilizados nas análises. Para os testes, adotou-se o nível de significância de 5%. Na Tabela 6.5, é apresentado o resultado da aplicação destes testes.

Tabela 6.5: Análise da significância estatística da contribuição de *hashtags* contidas em *tweets* no cenário eleitoral

	Dataset EUA					
	NB		*MNB*		*SVM*	
	unigrama	*bigrama*	*unigrama*	*bigrama*	*unigrama*	*bigrama*
Baseline	63,5%	60,6%	68,9%	70,0%	64,6%	65,8%
TH	64,4%	60,7%	69,6%	70,2%	65,6%	66,4%
Test-t	-	0,46	0,13e-3	-	-	-
Wilcoxon	**0,57e-2**	0,67	**0,55e-2**	0,27	0,12	**0,9e-2**
	Dataset BR					
	NB		*MNB*		*SVM*	
Baseline	49,9%	53,8%	58,2%	63,3%	56,7%	60,2%
TH	52,3%	54,6%	61,1%	64,1%	59,0%	60,6%
Test-t	-	-	0,27e-7	0,31e-2	-	-
Wilcoxon	**0,19e-2**	**0,16e-1**	**0,58e-2**	**0,13e-1**	**0,91e-2**	0,52e-1

Conforme discutido anteriormente, as *hashtags* contidas nos *tweets* nos cenários eleitorais brasileiro e americano foram capazes de aumentar as acurácias dos algoritmos em 100% dos casos analisados. Em relação aos incrementos obtidos no *dataset* americano,

Tabela 6.5, 50% deles foram estatisticamente significantes (representados nesta tabela em negrito). Em relação a amostra brasileira, o uso de *hashtags* foi mais eficaz na melhoria do desempenho dos algoritmos, pois 83,4% dos aumentos obtidos nas acurácias foram estatisticamente significantes, ou seja, apenas a melhoria da acurácia do algoritmo *SVM* (bigrama), de 60,2% para 60,6%, não obteve o mesmo resultado.

A significância estatística da contribuição das *hashtags* políticas contidas nos *tweets* das duas amostras, também foi analisada neste experimento, a partir da comparação dos resultados do cenário T2 com o *baseline*. O resultado desta investigação é apresentado na Tabela 6.6.

Tabela 6.6: Contribuição de *hashtags* políticas no desempenho de classificadores

Dataset EUA						
	NB		*MNB*		*SVM*	
	unigrama	*bigrama*	*unigrama*	*bigrama*	*unigrama*	*bigrama*
Baseline	63,5%	60,6%	68,9%	70,0%	64,6%	65,8%
T2	64,3%	60,6%	69,6%	70,2%	65,2%	66,1%
Test-t	-	-	**0,13e-3**	-	-	-
Wilcoxon	**0,14e-1**	1	**0,55e-2**	0,27	0,25	0,23
Dataset BR						
	NB		*MNB*		*SVM*	
Baseline	49,9%	53,8%	58,2%	63,3%	56,7%	60,2%
T2	52,3%	54,5%	61,0%	64,1%	59,0%	60,6%
Test-t	-	-	**0,71e-7**	**0,31e-2**	-	-
Wilcoxon	**0,19e-2**	**0,77e-2**	**0,59e-2**	**0,13e-1**	**0,91e-2**	0,52e-1

Com base nos resultados apresentados na Tabela 6.6, há evidências para se concluir que *hashtags* políticas são informações que não podem ser descartadas na fase de pré-processamento, em cenários de eleições. Em relação a amostra brasileira, de todos os aumentos obtidos (100% dos casos), 83,4% deles foram estatisticamente significantes. Já na base americana 40% dos 83,4% dos incrementos obtidos, nas acurácias dos algoritmos, foram estatisticamente significantes.

Neste segundo experimento foi avaliada também a contribuição do atributo $TPSB$, apresentado na Seção 4.3.2, na melhoria do desempenho dos algoritmos *NB*, *MNB* e *SVM*. Para ilustrar como este atributo é representado, considere a mensagem obtida a partir do *dataset* brasileiro[6]: "Aécio ensina como deve ser tratado o idioma que Lula e Dilma torturam há 12 anos URL #Aécio45 #MudaBrasil #ForaDilma". Nesse exemplo, a entidade Dilma Rousseff está sendo utilizada para representar o Candidato 1 e, Aécio Neves, o Candidato 2. Para essa mensagem, o atributo $TPSB$ é representado pela tupla (0,1,1,0).

[6]O endereço eletrônico contido no exemplo do *tweet* foi substituído pelo termo *URL*

O valor do primeiro parâmetro da tupla, C1_THSN, é zero, pois não há nenhuma *hashtag* contendo algum *slogan* de campanha ou o primeiro nome e/ou sobrenome do Candidato 1. O segundo parâmetro, C2_THSN, é igual a 1, pois há pelo menos uma *hashtag* contendo o *slogan* de campanha do presidenciável Aécio Neves, representada por #MudaBrasil. O terceiro parâmetro, C1_THPR, é igual a 1, devido a presença da *hashtag* de repúdio #ForaDilma, sobre o Candidato 1. Por fim, o último parâmetro, C2_THPR, é igual a zero, pois não há nenhuma *hashtag* de repúdio sobre o Candidato 2.

O resultado da análise da contribuição do atributo *TPSB* na melhoria do desempenho dos classificadores *NB*, *MNB* e *SVM*, é apresentado na Tabela 6.7. Nesta tabela, é apresentado também o resultado da aplicação dos mesmos testes estatísticos, utilizados anteriormente, para verificar se os aumentos obtidos nas acuráricas dos algoritmos foram estatisticamente significantes, ao adicionar este novo atributo nas análises. O desempenho dos algoritmos foi comparado com os melhores resultados obtidos na análise anterior (*TH*).

Tabela 6.7: Contribuição do atributo *PSBT* na melhoria das acurácias dos classificadores

Dataset EUA						
	NB		*MNB*		*SVM*	
	unigrama	*bigrama*	*unigrama*	*bigrama*	*unigrama*	*bigrama*
TH	64,4%	60,7%	69,6%	70,2%	65,6%	66,4%
TPSB	65,0%	60,7%	69,9%	71,1%	65,6%	67,1%
Test-t	-	-	0,16	**0,19e-3**	-	**0,21e-2**
Wilcoxon	**0,78e-2**	1	0,14	**0,9e-2**	1	**0,13e-1**
Dataset BR						
	NB		*MNB*		*SVM*	
TH	52,3%	54,6%	61,1%	64,1%	59,0%	60,6%
TPSB	53,1%	56,6%	61,2%	65,7%	59,1%	62,1%
Test-t	-	-	0,79	-	-	**0,3e-2**
Wilcoxon	0,18	**0,58e-2**	0,72	**0,58e-2**	0,71	**0,1e-1**

Conforme apresentado na Tabela 6.7, o atributo *TPSB* contribuiu para melhorar as acurácias de todos os classificadores, independentemente do *dataset* e formato (unigrama/bigrama) utilizados, com exceção nas configurações *NB* (bigrama) e *SVM* (unigrama), utilizando a amostra americana, cujas acurácias se mantiveram constantes. Pelos resultados apresentados nesta tabela, conclui-se também que o desempenho dos algoritmos foram melhores quando os atributos foram representados no formato bigrama, ou seja, em 100% destes casos os aumentos nas acurácias foram estatisticamente significantes.

Discussão

Com os resultados obtidos neste segundo experimento, conclui-se que *hashtags* desempenham um papel importante na classificação do sentimento de *tweets* postados em cenários de eleições. No segundo experimento, chegou-se a conclusão que neste domínio, isto é, *hashtags* contendo *slogans* de campanha, o primeiro nome e/ou sobrenome de candidato associado a outras palavras/números e palavras de repúdio sobre presenciáveis, podem ser utilizadas para melhorar o desempenho de algoritmos de aprendizado de máquina, utilizados para classificar o sentimento de *tweets* postados em períodos de campanha eleitoral. Em [94], os autores concluíram que a variação de atributos podem melhorar o desempenho de classificadores. Neste xperimento, esta conclusão foi corroboraba ao propor um novo atributo baseado no conceito de *hashtags* políticas, o *TPSB*. Nas análises realizadas, foi mostrado também que este tipo de atributo apresenta um melhor desempenho quando os *tweets* são representados aos classificadores no formato bigrama. No experimento a seguir é investigada a contribuição de *hashtags* contidas nas descrições dos perfis dos usuários.

6.2.3 Experimento III: Avaliação de *hashtags* em descrições de perfis

Neste terceiro experimento, a contribuição de *hashtags* contidas em descrições de perfis de usuários, no cenário eleitoral, são investigadas sob duas perspectivas diferentes. Na primeira, é analisada a contribuição das duas categorias de *hashtags*, as políticas e as não-políticas, contidas nas descrições dos perfis dos usuários, na classificação do sentimento de *tweets* no cenário eleitoral, a partir das amostras brasileira e americana. Na segunda perspectiva, é avaliada a contribuição do atributo *DPSB*, proposto neste trabalho e utilizado nas análises, na melhoria do desempenho dos algoritmos de aprendizado de máquina supervisionado.

Conforme apresentado na Seção 4.3.1, o modelo proposto avalia a contribuição de *hashtags* contidas em descrições de perfis de usuários, a partir de três cenários distintos: (D1) avaliação da contribuição de *hashtags* não-políticas, (D2) avaliação da contribuição de *hashtags* políticas e (D3) avaliação da contribuição de *hashtags* políticas e não-políticas, em conjunto. Conforme apresentado no Capítulo 4, as *hashtags* de cada um destes três cenários são incorporadas ao final dos *tweets*. O resultado da classificação do sentimento dos *tweets* das amostras americana e brasileira, utilizando cada um destes cenários, é mostrado na Tabela 6.8.

Tabela 6.8: Contribuição de *hashtags* políticas e não-políticas contidas em descrições de perfis de usuários

	Dataset EUA					
	NB		*MNB*		*SVM*	
	unigrama	*bigrama*	*unigrama*	*bigrama*	*unigrama*	*bigrama*
Baseline	63,5%	60,6%	68,9%	70,0%	64,6%	65,8%
D1	63,6%	60,5%	68,4%	67,0%	**65,0**%	65,9%
D2	63,8%	60,6%	**69,4**%	**70,1**%	64,5%	**66,1**%
D3	**63,9**%	60,5%	68,8%	66,9%	65,0%	65,7%
	Dataset BR					
	NB		*MNB*		*SVM*	
Baseline	49,9%	53,8%	58,2%	63,3%	56,7%	60,2%
D1	49,9%	53,8%	58,1%	**63,4**%	56,5%	60,2%
D2	49,9%	53,8%	58,1%	63,3%	56,3%	60,2%
D3	49,9%	53,8%	58,1%	63,4%	56,3%	60,2%

Conforme pode ser observado na Tabela 6.8, *hashtags* contidas em descrições de perfis de usuários, incorporadas a *tweets*, podem contribuir para a melhoria do desempenho de algoritmos de aprendizado de máquina supervisionado, utilizados para classificar o sentimento de *tweets* no cenário eleitoral. Ao utilizar o *dataset* americano nas análises, observou-se que as *hashtags* políticas contidas nas descrições dos perfis dos usuários (D2), foram capazes de melhorar o desempenho de todos os classificadores, com exceção das configurações *NB* (bigrama) e *SVM* (unigrama), cujas acurácias ficaram inalterada e reduzida, respectivamente. Já as *hashtags* não-políticas, foram capazes de melhorar as acurácias dos algoritmos *NB* (bigrama) e *SVM* (unigrama e bigrama). Ao considerar o uso em conjunto desses dois tipos de *hashtags*, verificou-se um melhoria na acurácia apenas do algoritmo *NB* (unigrama). Estes mesmos resultados já não foram obtidos ao utilizar o *dataset* brasileiro, onde o único incremento obtido ocorreu somente ao considerar as *hashtags* não-políticas (D1) nas análises. Neste caso, apenas a acurácia do algoritmo *NB* (bigrama) foi incrementado.

As melhores acurácias obtidas pelos algoritmos, Tabela 6.8, representadas a seguir pela nomenclatura *DH*, foram comparadas com o *baseline* obtido no primeiro experimento com o objetivo de verificar se o ganho obtido na acurácia de cada algoritmo foi estatisticamente significante. Para isto, foi realizado um teste de hipótese semelhante ao do experimento da Seção 6.2.2. Para estes testes, foi estabelecido, como hipótese nula (H_0), que não há diferença estatística entre amostra de *tweets* com e sem *hashtags* contidas em descrições de perfis de usuários. Na Tabela 6.9, são apresentados os resultados destes testes.

Com base nos resultados apresentados na Tabela 6.9, conclui-se que há indícios para

Tabela 6.9: Análise da significância estatística da contribuição de *hashtags* contidas em descrições de perfis no desempenho de classificadores

	Dataset EUA					
	NB		*MNB*		*SVM*	
	unigram	*bigram*	*unigram*	*bigram*	*unigram*	*bigram*
Baseline	63,5%	60,6%	68,9%	70,0%	64,6%	65,8%
DH	63,9%	60,6%	69,4%	70,1%	65,0%	66,1%
Test-t	-	-	**0,33e-2**	-	-	-
Wilcoxon	**0,21e-1**	1	**0,14e-1**	0,20	0,61	0,63
	Dataset BR					
	NB		*MNB*		*SVM*	
Baseline	49,9%	53,8%	58,2%	63,3%	56,7%	60,2%
DH	49,9%	53,8%	58,1%	63,4%	56,5%	60,2%
Test-t	-	-	-	0,86e-1	-	-
Wilcoxon	1	1	-	0,17	-	1

afirmar que usuários, durante períodos de campanha eleitoral, utilizam outros espaços em suas contas no *Twitter*, além de *tweets*, para expressar opinião política em cenários de eleições. Em relação ao *dataset* americano, por exemplo, verificou-se que *hashtags* contidas nas descrições dos perfis dos usuários foram responsáveis pelo aumento das acurácias de todos os classificadores, com exceção da configuração *NB* (bigrama), que se manteve constante. Nesta amostra, 40% dos 83,4% dos incrementos obtidos nas acurácias dos algoritmos foram estatisticamente significantes. Em relação ao *dataset* brasileiro, o único aumento registrado na acurácia, a do algoritmo MNB (bigrama), não foi estatisticamente significante.

Neste terceiro experimento, foi avaliada também a contribuição do atributo *DPSB*, apresentado na Seção 4.3.2, na melhoria do desempenho dos algoritmos *NB*, *MNB* e *SVM*. Para ilustrar a representação deste atributo, foi obtida do *dataset* brasileiro a seguinte mensagem: "Com uma campanha honrada Aécio mudará o Brasil para melhor! #AecioEmTodoBrasil #DebateNaRecord". Em seguida, foi identificada no *dataset DescPU* a descrição do perfil do autor dessa mensagem[7]: "SIGA-NOS NO INSTAGRAM: @foraptforadilma Quer mudar o Brasil? #foraPT #foraDilma - perfil apartidário e pró-Brasil! Chega de roubalheira! -- @USER". As *hashtags* políticas #foraPT e #foraDilma foram identificadas e obtidas das descrições. Neste exemplo, utilizou-se a seguinte convenção: Candidato 1: Dilma Rousseff e Candidato 2: Aécio Neves. Para este exemplo, a representação do atributo *DPSB* será dada pela tupla (0,0,1,0). O primeiro valor desta tupla é zero, pois não há na descrição do perfil nenhuma *hashtag* contendo *slogan* de campanha

[7]A identificação do usuário no final do *tweet* do exemplo foi substituído por @USER

e/ou o primeiro nome e/ou sobrenome do Candidato 1. O segundo valor é zero, pois não há nenhuma *hashtag* contendo *slogan* de campanha e/ou o primeiro nome e/ou sobrenome do Candidato 2. O terceiro valor é 1, pois as duas *hashtags* encontradas possuem palavras de repúdio sobre o Candidato 1. Por fim, o último valor é 0, pois não foi encontrada nenhuma *hashtag* contendo palavras de repúdio sobre o Candidato 2.

A contribuição do atributo *DPSB* na melhoria do desempenho dos classificadores é apresentada na Tabela 6.10. Nesta tabela, também é apresentado o resultado da aplicação do mesmo teste estatístico utilizado nas análises da Seção 6.2.2, com o objetivo de verificar se os aumentos obtidos pelos algoritmos são estatisticamente significantes, ao considerar esse atributo nas análises.

Tabela 6.10: Constribuição do atributo *DPSB* na melhoria da acurácia dos classificadores

Dataset EUA						
	NB		*MNB*		*SVM*	
	unigram	*bigram*	*unigram*	*bigram*	*unigram*	*bigram*
DH	63,9%	60,6%	69,4%	70,1%	65,0%	66,1%
DPSB	64,4%	60,6%	69,4%	70,2%	65,0%	66,1%
Test-t	-	-	-	-	-	-
Wilcoxon	**0,28e-1**	1	1	0,40	1	1
Dataset BR						
]	*NB*		*MNB*		*SVM*	
DH	49,9%	53,8%	58,1%	63,4%	56,5%	60,2%
DPSB	49,9%	53,8%	58,1%	63,3%	56,3%	60,2%
Test-t	-	-	-	-	-	-
Wilcoxon	1	1	1	-	-	1

O atributo *DPSB* foi adicionado a configuração responsável pela obtenção dos melhores resultados apresentados na Tabela 6.9. Para o *dataset* americano, dos aumentos obtidos nas acurácias dos classificadores, após a adição deste atributo nas análises, um deles foi estatisticamente significante, com valor de *p-value* igual a 0,028, conforme destacado em negrito na Tabela 6.10. No *dataset* brasileiro, este atributo influenciou negativamente a acurácia dos algoritmos MNB (bigrama) e SVM (unigrama). Com base nestes resultados, conclui-se que este atributo, apesar de ter incrementado (decrementado) as acurácias de dois algoritmos, utilizado o *dataset* americano (brasileiro), ainda carece de outras análises, como o uso de outros *datasets* no cenário eleitoral, para se avaliar a contribuição dele neste tipo de cenário.

Discussão

Com base nos resultados obtidos neste terceiro experimento, conclui-se que *hashtags*

políticas, presentes em descrições de perfis de usuários, consistem num conjunto de informações que podem indicar uma possibilidade de melhoria no desempenho de algoritmos de aprendizado de máquina supervisionado, utilizados para classificar o sentimento de *tweets* no cenário político. Em um dos *datasets* utilizados nas análises, verificou-se que as acurácias de 83,4% dos casos analisados foram incrementados. Conforme apresentado no Capítulo 5, a diferença de periodicidade na coleta das descrições dos perfis dos usuários brasileiros e americanos foram diferentes. Enquanto para o *dataset* americano foram coletadas todas as atualizações das descrições dos perfis dos usuários até o dia da eleição, para o *dataset* brasileiro foi coletado apenas a última atualização da descrição, antes do dia da votação. Acredita-se que essa diferença de abordagem na coleta deste tipo de informação, em cenário de eleições, tenha influenciado os resultados obtidos neste experimento. Portanto, conclui-se que o uso deste tipo de informação na classificação do sentimento de *tweets*, postados em períodos de eleições, ainda carece de investigações. No próximo experimento, é investigada a contribuição em conjunto das *hashtags* contidas nas mensagens e nas descrições dos perfis dos usuários.

6.2.4 Experimento IV: Avaliação de *hashtags* contidas em mensagens e em descrições de perfis

No último experimento, é investigada a contribuição de *hashtags* contidas em *tweets* e em descrições de perfis de usuários, em conjunto, na classificação do sentimento das amostras brasileira e americana. Para realizar esta investigação foram propostas duas análises. Na primeira análise (Análise I), são combinadas as melhores configurações obtidas nos Experimentos II e III e, na segunda análise (Análise II), os dois atributos propostos neste trabalho, $TPSB$ e $DPSB$, são adicionados às configurações da Análise I. Os resultados dessas análises são apresentados na Tabela 6.11.

Na Análise I, ao considerar o *dataset* americano nas análises, verificou-se que os classificadores *NB* (unigrama e bigrama) e *MNB* (unigrama) obtiveram suas acurárias incrementadas, em relação aos resultados obtidos nos experimentos II e III. Em relação ao *dataset* brasileiro, não foi identificada nenhuma melhoria nas acurácias dos classificadores, quando comparadas com os experimentos II e III.

Na Análise II, ao utilizar a amostra americana, verificou-se que combinar *hashtags* e atributos baseados em *hashtags* políticas foi uma boa alternativa para melhorar as acurácias dos algoritmos *NB* (unigrama), *MNB* (unigrama) e *SVM* (bigrama). Neste cenário, por exemplo, foram obtidas as melhores acurácias por estes algoritmos até presente mo-

Tabela 6.11: Constribuição de *hashtags* contidas em mensagens e em descrições de perfis de usuários

	Dataset EUA					
	NB		*MNB*		*SVM*	
	unigram	*bigram*	*unigram*	*bigram*	*unigram*	*bigram*
Baseline	63,5%	60,6%	68,9%	70,0%	64,6%	65,8%
Experimento II	65,0%	**60,7%**	69,9%	**71,1%**	**65,6%**	67,1%
Experimento III	64,4%	60,6%	69,4%	70,2%	65,0%	66,1%
Experimento IV (Análise I)	65,2%	**60,7%**	70,1%	67,8%	65,4%	66,6%
Experimento IV (Análise II)	**65,7%**	**60,7%**	**70,4%**	69,5%	65,3%	**67,5%**
	Dataset BR					
	NB		*MNB*		*SVM*	
Baseline	49,9%	53,8%	58,2%	63,3%	56,7%	60,2%
Experimento II	**53,1%**	**56,6%**	**61,2%**	65,7%	**59,1%**	62,1%
Experimento III	49,9%	53,8%	58,1%	63,4%	56,5%	60,3%
Experimento IV (Análise I)	52,3%	54,6%	61,1%	64,2%	59,0%	60,6%
Experimento IV (Análise II)	**53,1%**	**56,6%**	**61,2%**	**65,9%**	58,9%	**62,3%**

mento. O mesmo ocorreu para os algoritmos MNB (bigrama) e SVM (bigrama) ao utilizar o *dataset* brasileiro.

Com base nos resultados apresentados na Tabela 6.11, conclui-se que é possível: (1) melhorar o desempenho de algoritmos de aprendizado de máquina supervisionados, utilizados para classificar o sentimento de *tweets* em cenários de eleições, por meio da combinação de *hashtags* contidas em *tweets* e em descrições de perfis de usuários; (2) adicionar nas análises a representação de atributos baseados em *hashtags* políticas. Conforme apresentado na Tabela 6.11, *hashtags* e atributos baseados em *hashtags* políticas foram capazes de aumentar as acurácias de todos os classificadores, independentemente da amostra utilizada. Por exemplo, o algoritmo *NB* (unigrama) teve a sua acurácia incrementada em 2,7%, em média, em relação ao *baseline*. Já o algoritmo *SVM*, utilizando o formato bigrama, incrementou o seu desempenho em 1,9%, em média, quando comparado com o *baseline*.

Capítulo 7

Conclusões

Nesta tese, foi realizado um estudo para investigar a relevância de *hashtags* contidas em *tweets* e em descrições de perfis de usuários, no cenário eleitoral. Foi proposto um modelo para analisar a contribuição destas informações na melhoria do desempenho de algoritmos de aprendizado de máquina supervisionado, utilizados para classificar o sentimento de *tweets* em cenários de eleições. Ainda foi investigada a contribuição de *hashtags* políticas, contidas em *tweets* e em descrições de perfis de usuários, na classificação do sentimento de *tweets* políticos. Por último, foi examinado como atributos baseados em *hashtags* políticas podem ser levados em consideração para melhorar a acurácia de classificadores.

Hashtags, contidas em *tweets* no cenário eleitoral, já foram utilizadas em pesquisas científicas sob diferentes perspectivas, por exemplo para coletar e selecionar *tweets* [6, 12, 13, 16, 22, 31, 102], para rotular um *corpus* de *tweets* [3, 60], etc. Nesses estudos, o *tweet* foi a principal fonte de dados utilizada nas análises. No levantamento bibliográfico realizado neste trabalho não foi encontrado nenhum estudo que tenha examinado a contribuição de *hashtags*, contidas em *tweets* e em descrições de perfis de usuários, na melhoria do desempenho de classificadores, utilizados para analisar o sentimento de *tweets* nesse cenário.

Para investigar a questão de pesquisa proposta neste trabalho, foram realizados quatro experimentos computacionais. As bases de dados utilizadas nas análises, foram as mesmas reportadas em [95] e em [98]. O primeiro experimento realizado teve como objetivo avaliar o desempenho das técnicas de pré-processamento, remoção de *stopwords* e *stemming*, no cenário eleitoral. Neste experimento, foram aplicadas também as técnicas de remoção de *URL's*, conversão de letras maiúsculas em minúsculas, remoção de estruturas específicas do *Twitter* e símbolos de pontuação. As configurações responsáveis pela obtenção das melhores acurácias obtidas pelos classificadores *NB*, *SVM* e *MNB*, neste experimento, foram

utilizadas como *baseline* para os outros experimentos. No segundo experimento, foram avaliadas a contribuição das *hashtags* políticas e não-políticas, propostas neste trabalho, e de um atributo baseado em *hashtags* políticas, o *TPSB*. No terceiro experimento, foram avaliadas a contribuição das *hashtags* políticas e não-políticas, e de um atributo baseado em *hashtags* políticas, o *DPSB*, obtidos a partir de informações localizadas nas descrições dos perfis dos usuários. No último experimento, foi avaliada a contribuição de *hashtags* contidas tanto em *tweets* quanto em descrições de perfis de usuários, em conjunto com os atributos *TPSB* e *DPSB*.

As investigações realizadas neste trabalho, mostraram que *hashtags* contidas em *tweets* se mostraram mais relevantes do que *hashtags* contidas em descrições de perfis de usuários, no cenário eleitoral. Os resultados obtidos, a partir dos experimentos realizados, mostraram que estas informações são capazes de melhorar o desempenho de algoritmos de aprendizado supervisionado, utilizados para classificar o sentimento de *tweets* no cenário político. Nos testes realizados com a amostra brasileira, *hashtags* políticas foram responsáveis por aumentar com significância estatística as acurácias de todos os classificadores, independentemente do formato utilizado (unigrama/bigrama). Nos testes realizados com a base americana, as acurácias de todos os classificadores também foram incrementadas e, em 66% dos casos, os aumentos obtidos foram estatisticamente significantes.

Os resultados encontrados sugerem que *hashtags* contendo palavras fazendo referência a candidatos, por exemplo a partir do nome deles e de seus *slogans* de campanha, podem ser úteis na classificação do sentimento de *tweets* políticos, e que usuários do *Twitter*, durante períodos de campanha eleitoral, expressam opinião política não somente a partir de *tweets*, mas também a partir de informações localizadas em suas descrições de perfis. Nos experimentos propostos, também foi analisada a contribuição de atributos baseados em *hashtags* políticas contidas em *tweets* e em descrições de perfis de usuários. Com base nos resultados obtidos, chegou-se a conclusão que o atributo *TPSB*, baseado em *hashtags* políticas contidas em *tweets*, contribuiu para melhorar as acurácias de todos os classificadores utilizados nas análises, com exceção de dois cenários, onde a acurácia se manteve constante. Já os resultados obtidos ao utilizar atributo *DPSB*, baseado em *hashtags* políticas contidas em descrições de perfis de usuários, foi capaz de melhorar as acurácias dos classificadores *NB* (unigrama) e *MNB* (bigrama), utilizando uma das duas bases de dados utilizadas nos experimentos. Porém, para afirmar que esse segundo tipo de informação contribui, de fato, para a melhoria do desempenho de algoritmos de aprendizado supervisionado, no cenário político, sugere-se a realização de novos experimentos, a partir da utilização de novas bases de dados.

O modelo proposto neste trabalho, utilizado para avaliar a contribuição de *hashtags* contidas em *tweets* e em descrições de perfis de usuários, na melhoria do desempenho de algoritmos de aprendizado de máquina supervisionado, mostrou-se eficiente ao avaliar duas amostras presidenciais. O modelo também foi capaz de avaliar a contribuição de dois atributos baseados em *hashtags* políticas. O modelo foi proposto para o cenário eleitoral, porém pode ser facilmente adaptado para avaliar *hashtags* de outros domínios.

Durante o processo de investigação do problema abordado nesta tese, alguns trabalhos foram publicados. Por exemplo, em [95] foi apresentado um estudo para predizer o resultado da eleição presidencial brasileira no ano de 2014, a partir da análise de informações contidas em *tweets* postados durante o período de campanha eleitoral. Nesse trabalho, foram utilizadas as abordagens de contagem e análise de sentimento das mensagens, para predizer o resultado final daquela eleição. Em [96], foi apresentada uma proposta de método semiautomático para rotulação de *tweets* políticos, a partir da análise de informações contidas nas descrições dos perfis dos usuários. Em [97], o método apresentado em [96], para rotulação de *tweets* de cunho político, foi ampliado. Por último, em [98] foi realizado um estudo para investigar a relevância de *hashtags* contidas em *tweets* e em descrições de perfis de usuários, no cenário de eleições.

Uma das limitações identificadas durante a realização deste trabalho, refere-se aos dados utilizados para avaliar a contribuição de informações contidas nas descrições dos perfis dos usuários da base brasileira, pois para essa base, foi coletada apenas a última atualização da descrição do perfil de cada usuário, antes do dia da votação. Outra limitação consiste na quantidade de bases de dados utilizadas para avaliar o modelo proposto. Devido a singularidade do problema investigado, não são encontradas na literatura bases de dados no cenário político, contendo as duas informações necessárias ao modelo proposto: *tweets* e descrições de perfis de usuários. Por último, outra limitação encontrada consiste na quantidade de candidatos presentes nos *tweets* que o modelo é capaz de avaliar.

Como trabalho futuro, pretende-se avaliar o modelo proposto a partir de outras bases de dados contendo *tweets* no cenário eleitoral. Além disto, os resultados obtidos neste trabalho, usando o cenário político, mostram a viabilidade em estender o estudo proposto para outros domínios, onde há a presença de *hashtags* com características similares aquelas mostradas nos experimentos realizados no Capítulo 5. Uma outra frente de pesquisa a ser investigada, é o desenvolvimento de um aplicativo de software para realizar a seleção automática de *hashtags* políticas. Uma outra proposta de continuidade a este trabalho, consiste em analisar outros algoritmos de aprendizado de máquina supervisionado, uti-

lizados na última etapa do modelo, e investigar outras técnicas para representação de atributos aos classificadores, como *TF-IDF* [73].

Referências

[1] Agarwal, A.; Xie, B.; Vovsha, I.; Rambow, O.; Passonneau, R. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media* (2011), Association for Computational Linguistics, pp. 30–38.

[2] Aisopos, F.; Papadakis, G.; Varvarigou, T. Sentiment analysis of social media content using n-gram graphs. In *Proceedings of the 3rd ACM SIGMM international workshop on Social media* (2011), ACM, pp. 9–14.

[3] Alfina, I.; Sigmawaty, D.; Nurhidayati, F.; Hidayanto, A. N. Utilizing hashtags for sentiment analysis of tweets in the political domain. In *Proceedings of the 9th International Conference on Machine Learning and Computing* (2017), ACM, pp. 43–47.

[4] Allcott, H.; Gentzkow, M. Social media and fake news in the 2016 election. *Journal of Economic Perspectives 31*, 2 (2017), 211–36.

[5] Almatrafi, O.; Parack, S.; Chavan, B. Application of location-based sentiment analysis using twitter for identifying trends towards indian general elections 2014. In *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication* (2015), ACM, p. 41.

[6] Anjaria, M.; Guddeti, R. M. R. A novel sentiment analysis of social networks using supervised learning. *Social Network Analysis and Mining 4*, 1 (2014), 181.

[7] Baccianella, S.; Esuli, A.; Sebastiani, F. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec* (2010), vol. 10, pp. 2200–2204.

[8] Bakliwal, A.; Arora, P.; Madhappan, S.; Kapre, N.; Singh, M.; Varma, V. Mining sentiments from tweets. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis* (2012), pp. 11–18.

[9] Barbosa, L.; Feng, J. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd international conference on computational linguistics: posters* (2010), Association for Computational Linguistics, pp. 36–44.

[10] Bermingham, A.; Smeaton, A. On using twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)* (2011), pp. 2–10.

[11] Bouazizi, M.; Ohtsuki, T. O. A pattern-based approach for sarcasm detection on twitter. *IEEE Access 4* (2016), 5477–5488.

[12] Bruns, A.; Burgess, J. E. # ausvotes: How twitter covered the 2010 australian federal election. *Communication, Politics and Culture 44*, 2 (2011), 37–56.

[13] Burger, J. D.; Henderson, J.; Kim, G.; Zarrella, G. Discriminating gender on twitter. In *Proceedings of the conference on empirical methods in natural language processing* (2011), Association for Computational Linguistics, pp. 1301–1309.

[14] Ceron, A.; Curini, L.; Iacus, S. M. Using sentiment analysis to monitor electoral campaigns: Method matters evidence from the united states and italy. *Social Science Computer Review 33*, 1 (2015), 3–20.

[15] Ceron, A.; Curini, L.; Iacus, S. M.; Porro, G. Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens political preferences with an application to italy and france. *New Media & Society 16*, 2 (2014), 340–358.

[16] Cerón-Guzmán, J. A.; León-Guzmán, E. A sentiment analysis system of spanish tweets and its application in colombia 2014 presidential election. In *Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom), 2016 IEEE International Conferences on* (2016), IEEE, pp. 250–257.

[17] Chaovalit, P.; Zhou, L. Movie review mining: A comparison between supervised and unsupervised classification approaches. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on* (2005), IEEE, pp. 112c–112c.

[18] Choy, M.; Cheong, M.; Laik, M. N.; Shung, K. P. Us presidential election 2012 prediction using census corrected twitter model. *arXiv preprint arXiv:1211.0938* (2012).

[19] Choy, M.; Cheong, M. L.; Laik, M. N.; Shung, K. P. A sentiment analysis of singapore presidential election 2011 using twitter data with census correction. *arXiv preprint arXiv:1108.5520* (2011).

[20] Chung, J. E.; Mustafaraj, E. Can collective sentiment expressed on twitter predict political elections? In *AAAI* (2011), vol. 11, pp. 1770–1771.

[21] Coletta, L. F. S.; da Silva, N. F. F.; Hruschka, E. R.; Hruschka, E. R. Combining classification and clustering for tweet sentiment analysis. In *Intelligent Systems (BRACIS), 2014 Brazilian Conference on* (2014), IEEE, pp. 210–215.

[22] Conover, M. D.; Gonçalves, B.; Ratkiewicz, J.; Flammini, A.; Menczer, F. Predicting the political alignment of twitter users. 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. 192–199.

[23] Contractor, D.; Faruquie, T. A. Understanding election candidate approval ratings using social media data. In *Proceedings of the 22nd International Conference on World Wide Web* (2013), ACM, pp. 189–190.

[24] Dave, K.; Lawrence, S.; Pennock, D. M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web* (2003), ACM, pp. 519–528.

[25] Diakopoulos, N. A.; Shamma, D. A. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010), ACM, pp. 1195–1198.

[26] do Povo, G. Como é feita uma pesquisa eleitoral? `https://especiais.gazetadopovo.com.br/eleicoes/2018/qual-o-passo-passo-de-uma-pesquisa-eleitoral/`, 2018. [Online; accessado 07-Julho-2018].

[27] Dosciatti, M. M.; Ferreira, L. P. C.; Paraiso, E. C. Identificando emoçoes em textos em português do brasil usando máquina de vetores de suporte em soluçao multiclasse. *ENIAC-Encontro Nacional de Inteligência Artificial e Computacional. Fortaleza, Brasil* (2013).

[28] Duwairi, R.; El-Orfali, M. A study of the effects of preprocessing strategies on sentiment analysis for arabic text. *Journal of Information Science 40*, 4 (2014), 501–513.

[29] Feldman, R. Techniques and applications for sentiment analysis. *Communications of the ACM 56*, 4 (2013), 82–89.

[30] Filho, C. A. F. Mineração de opiniões: Um classificador ternário ou dois binários? `http://www.ic.uff.br/PosGraduacao/frontend-tesesdissertacoes/download.php?id=765.pdf&tipo=trabalho`, 2016. [Online; accessado 01-Julho-2018].

[31] Fink, C.; Bos, N.; Perrone, A.; Liu, E.; Kopecky, J. Twitter, public opinion, and the 2011 nigerian presidential election. In *Social Computing (SocialCom), 2013 International Conference on* (2013), IEEE, pp. 311–320.

[32] Flekova, L.; Ferschke, O.; Gurevych, I. Ukpdipf: Lexical semantic approach to sentiment polarity prediction in twitter data. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)* (2014), pp. 704–710.

[33] Gamon, M. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics* (2004), Association for Computational Linguistics, p. 841.

[34] Gaurav, M.; Srivastava, A.; Kumar, A.; Miller, S. Leveraging candidate popularity on twitter to predict election outcome. In *Proceedings of the 7th workshop on social network mining and analysis* (2013), ACM, p. 7.

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

[36] Gayo Avello, D.; Metaxas, P. T.; Mustafaraj, E. Limits of electoral predictions using twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (2011), Association for the Advancement of Artificial Intelligence.

[37] Gilbert, C. H. E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf* (2014).

[38] Gimpel, K.; Schneider, N.; O'Connor, B.; Das, D.; Mills, D.; Eisenstein, J.; Heilman, M.; Yogatama, D.; Flanigan, J.; Smith, N. A. Part-of-speech tagging for twitter: Annotation, features, and experiments. Tech. rep., Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, 2010.

[39] Go, A.; Bhayani, R.; Huang, L. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford 1*, 12 (2009).

[40] Gonçalves, P.; Araújo, M.; Benevenuto, F.; Cha, M. Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks* (2013), ACM, pp. 27–38.

[41] Haddi, E.; Liu, X.; Shi, Y. The role of text pre-processing in sentiment analysis. *Procedia Computer Science 17* (2013), 26–32.

[42] Han, J.; Pei, J.; Kamber, M. *Data mining: concepts and techniques*. Elsevier, 2011.

[43] Han, S.; Kavuluru, R. On assessing the sentiment of general tweets. In *Canadian Conference on Artificial Intelligence* (2015), Springer, pp. 181–195.

[44] Hatzivassiloglou, V.; McKeown, K. R. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics* (1997), Association for Computational Linguistics, pp. 174–181.

[45] Hatzivassiloglou, V.; Wiebe, J. M. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational linguistics-Volume 1* (2000), Association for Computational Linguistics, pp. 299–305.

[46] Hearst, M. A.; Dumais, S. T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. *IEEE Intelligent Systems and their applications 13*, 4 (1998), 18–28.

[47] Ho, C.; Murad, M. A. A.; Kadir, R. A.; Doraisamy, S. C. Word sense disambiguation-based sentence similarity. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (2010), Association for Computational Linguistics, pp. 418–426.

[48] Hssina, B.; Merbouha, A.; Ezzikouri, H.; Erritali, M. A comparative study of decision tree id3 and c4. 5. *International Journal of Advanced Computer Science and Applications 4*, 2 (2014).

[49] Hu, M.; Liu, B. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (2004), ACM, pp. 168–177.

[50] Hu, X.; Zhang, X.; Yoo, I.; Wang, X.; Feng, J. Mining hidden connections among biomedical concepts from disjoint biomedical literature sets through semantic-based association rule. *International Journal of Intelligent Systems 25*, 2 (2010), 207–223.

[51] Huang, Y.; Mitchell, T. M. Text clustering with extended user feedback. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (2006), ACM, pp. 413–420.

[52] Hull, D. A. Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science 47*, 1 (1996), 70–84.

[53] Jiang, L.; Yu, M.; Zhou, M.; Liu, X.; Zhao, T. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (2011), Association for Computational Linguistics, pp. 151–160.

[54] Jivani, A. G., et al. A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl 2*, 6 (2011), 1930–1938.

[55] John, G. H.; Langley, P. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* (1995), Morgan Kaufmann Publishers Inc., pp. 338–345.

[56] Jungherr, A.; Jurgens, P.; Schoen, H. Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im "predicting elections with twitter: What 140 characters reveal about political sentiment". *Social science computer review 30*, 2 (2012), 229–234.

[57] Khabiri, E. Ranking, labeling, and summarizing short text in social media by elham khabiri; with prateek jain as coordinator. *ACM SIGWEB Newsletter*, Autumn (2013), 3.

[58] Kohavi, R., et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (1995), vol. 14, Montreal, Canada, pp. 1137–1145.

[59] Kolchyna, O.; Souza, T. T.; Treleaven, P.; Aste, T. Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *arXiv preprint arXiv:1507.00955* (2015).

[60] Kouloumpis, E.; Wilson, T.; Moore, J. D. Twitter sentiment analysis: The good the bad and the omg! *Icwsm 11*, 538-541 (2011), 164.

[61] Kumar, A.; Sebastian, T. M. Sentiment analysis on twitter. *International Journal of Computer Science Issues (IJCSI) 9*, 4 (2012), 372.

[62] Kumaran, G.; Allan, J. Using names and topics for new event detection. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (2005), Association for Computational Linguistics, pp. 121–128.

[63] Lewis-Beck, M. S. Election forecasting: principles and practice. *The British Journal of Politics and International Relations 7*, 2 (2005), 145–164.

[64] Lin, C.; He, Y. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management* (2009), ACM, pp. 375–384.

[65] Liu, B. Sentiment analysis and subjectivity. *Handbook of natural language processing 2* (2010), 627–666.

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

[67] Liu, B.; Zhang, L. A survey of opinion mining and sentiment analysis. In *Mining text data*. Springer, 2012, pp. 415–463.

[68] Liu, K.-L.; Li, W.-J.; Guo, M. Emoticon smoothed language models for twitter sentiment analysis. In *Aaai* (2012), vol. 12, pp. 22–26.

[69] Lovins, J. B. Development of a stemming algorithm. *Mech. Translat. & Comp. Linguistics 11*, 1-2 (1968), 22–31.

[70] Macropol, K.; Bogdanov, P.; Singh, A. K.; Petzold, L.; Yan, X. I act, therefore i judge: Network sentiment dynamics based on user activity change. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on* (2013), IEEE, pp. 396–402.

[71] Makazhanov, A.; Rafiei, D.; Waqar, M. Predicting political preference of twitter users. *Social Network Analysis and Mining 4*, 1 (2014), 193.

[72] Malouf, R.; Mullen, T. Taking sides: User classification for informal online political discourse. *Internet Research 18*, 2 (2008), 177–190.

[73] Martineau, J.; Finin, T., et al. Delta tfidf: An improved feature space for sentiment analysis. *Icwsm 9* (2009), 106.

[74] Martínez-Cámara, E.; Montejo-Ráez, A.; Martín-Valdivia, M. T.; Ureña-López, L. A. Sinai: Machine learning and emotion of the crowd for sentiment analysis in microblogs. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (2013), vol. 2, pp. 402–407.

[75] Mascaro, C.; Goggins, S. P. Twitter as virtual town square: Citizen engagement during a nationally televised republican primary debate.

[76] McCallum, A.; Nigam, K., et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (1998), vol. 752, Citeseer, pp. 41–48.

[77] Mejova, Y.; Srinivasan, P.; Boynton, B. Gop primary season on twitter: popular political sentiment in social media. In *Proceedings of the sixth ACM international conference on Web search and data mining* (2013), ACM, pp. 517–526.

[78] Metaxas, P. T.; Mustafaraj, E.; Gayo-Avello, D. How (not) to predict elections. 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. 165–171.

[79] Mohammad, S. M.; Kiritchenko, S. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence 31*, 2 (2015), 301–326.

[80] Mohammad, S. M.; Kiritchenko, S.; Zhu, X. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242* (2013).

[81] Mohammad, S. M.; Turney, P. D. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence 29*, 3 (2013), 436–465.

[82] Mullen, T.; Collier, N. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (2004).

[83] Nexo, J. Quanto custam e como são reguladas as pesquisas eleitorais. `https://www.nexojornal.com.br/expresso/2016/09/29/Quanto-custam-e-como-s%C3%A3o-reguladas-as-pesquisas-eleitorais`, 2016. [Online; accessado 05-Setembro-2018].

[84] NLTK, L. Stemmers. `http://www.nltk.org/howto/stem.html`, 2018. [Online; accessado 04-Agosto-2018].

[85] NLTK, S. Source code for nltk.stem.snowball. `https://www.nltk.org/_modules/nltk/stem/snowball.html`, 2018. [Online; accessado 03-Agosto-2018].

[86] Nooralahzadeh, F.; Arunachalam, V.; Chiru, C.-G. 2012 presidential elections on twitter - an analysis of how the us and french election were reflected in tweets. In *Control Systems and Computer Science (CSCS), 2013 19th International Conference on* (2013), IEEE, pp. 240–246.

[87] Notícias, R. Bolsonaro no sul x haddad no nordeste: Ibope por regiões. `https://especiais.gazetadopovo.com.br/eleicoes/2018/graficos/bolsonaro-x-haddad-ibope-regioes/`, 2018. [Online; accessado 25-Setembro-2018].

[88] O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; Smith, N. A., et al. From tweets to polls: Linking text sentiment to public opinion time series. *Icwsm 11*, 122-129 (2010), 1–2.

[89] O'Hare, N.; Davy, M.; Bermingham, A.; Ferguson, P.; Sheridan, P.; Gurrin, C.; Smeaton, A. F. Topic-dependent sentiment analysis of financial blogs. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion* (2009), ACM, pp. 9–16.

[90] Osgood, C. E.; Suci, G. J.; Tannenbaum, P. H. 1975: The measurement of meaning. *Urbana, IL: University of Illinois Press* (1957).

[91] Paice, C. D. An evaluation method for stemming algorithms. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (1994), Springer-Verlag New York, Inc., pp. 42–50.

[92] Pak, A.; Paroubek, P. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (2010), vol. 10, pp. 1320–1326.

[93] Pandey, V.; Iyer, C. Sentiment analysis of microblogs. *CS 229: Machine learning final projects* (2009).

[94] Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (2002), Association for Computational Linguistics, pp. 79–86.

[95] Paula Filho, W.; Garcia, A. C. B. Predição do resultado das eleições presidenciais do brasil baseado em tuítes. In *Anais do XXXV Congresso da Sociedade Brasileira de Computação (CSBC)* (Recife, PE, BRASIL, Julho 2015).

[96] Paula Filho, W.; Garcia, A. C. B. Rotuel: A semi-automated method for labeling political tweets. In *IJCAI* (2015), pp. 4361–4362.

[97] Paula Filho, W.; Garcia, A. C. B. Uma proposta de método semiautomático para rotulação de tweets de cunho político. In *VI Workshop de Teses e Dissertações em Sistemas Colaborativos* (2015), Sociedade Brasileira de Computação.

[98] Paula Filho, W.; Rosseti, I.; Viterbo, J. On tweets, retweets, hashtags and user profiles in the 2016 american presidential election scene. In *Proceedings of the 18th Annual International Conference on Digital Government Research* (2017), ACM, pp. 120–128.

[99] Pavel, A.; Palade, V.; Iqbal, R.; Hintea, D. Using short urls in tweets to improve twitter opinion mining. In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on* (2017), IEEE, pp. 965–970.

[100] Pennacchiotti, M.; Popescu, A.-M. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (2011), ACM, pp. 430–438.

[101] Petrović, S.; Osborne, M.; Lavrenko, V. The edinburgh twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media* (2010), pp. 25–26.

[102] Prasetyo, N. D. *Tweet-based election prediction.* Tese de Doutorado, Citeseer, 2014.

[103] Rao, D.; Yarowsky, D.; Shreevats, A.; Gupta, M. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents* (2010), ACM, pp. 37–44.

[104] Refaee, E. Sentiment analysis for micro-blogging platforms in arabic. In *International Conference on Social Computing and Social Media* (2017), Springer, pp. 275–294.

[105] Ribeiro, F. N.; Araújo, M.; Gonçalves, P.; Benevenuto, F.; Gonçalves, M. A. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *arXiv preprint arXiv:1512.01818* (2015).

[106] Riloff, E.; Wiebe, J. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* (2003), Association for Computational Linguistics, pp. 105–112.

[107] Saif, H. *Semantic Sentiment Analysis of Microblogs.* Tese de Doutorado, The Open University, 2015.

[108] Sanders, E.; van den Bosch, A. Relating political party mentions on twitter with polls and election results.

[109] Sang, E. T. K.; Bos, J. Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the workshop on semantic analysis in social media* (2012), Association for Computational Linguistics, pp. 53–60.

[110] Segaran, T.; Hammerbacher, J. *Beautiful data: the stories behind elegant data solutions.* "O'Reilly Media, Inc.", 2009.

[111] Shao, C.; Ciampaglia, G. L.; Varol, O.; Flammini, A.; Menczer, F. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592* (2017), 96–104.

[112] Sidorov, G.; Miranda-Jiménez, S.; Viveros-Jiménez, F.; Gelbukh, A.; Castro-Sánchez, N.; Velásquez, F.; Díaz-Rangel, I.; Suárez-Guerra, S.; Trevino, A.; Gordon, J. Empirical study of machine learning based approach for opinion mining in tweets. In *Mexican international conference on Artificial intelligence* (2012), Springer, pp. 1–14.

[113] Simeon, C.; Hilderman, R. Evaluating the effectiveness of hashtags as predictors of the sentiment of tweets. In *International Conference on Discovery Science* (2015), Springer, pp. 251–265.

[114] Singh, P.; Sawhney, R. S.; Kahlon, K. S. Predicting the outcome of spanish general elections 2016 using twitter as a tool. In *Advanced Informatics for Computing Research.* Springer, 2017, pp. 73–83.

[115] Sohn, S.; Torii, M.; Li, D.; Wagholikar, K.; Wu, S.; Liu, H. A hybrid approach to sentiment sentence classification in suicide notes. *Biomedical informatics insights 5* (2012), BII–S8961.

[116] Statista. Number of monthly active twitter users in the united states from 1st quarter 2010 to 3rd quarter 2017 (in millions), 2017. Disponível em http://www.statista.com/statistics/274564/monthly-active-twitter-users-in-the-unitedstates/.

[117] Statista. Number of monthly active twitter users worldwide from 1st quarter 2010 to 3rd quarter 2017 (in millions), 2017. Disponível em http://www.statista.com/statistics/282087/number-ofmonthly-active-twitter-users/.

[118] Stehman, S. V. Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment 62*, 1 (1997), 77–89.

[119] Stone, P. J.; Dunphy, D. C.; Smith, M. S. The general inquirer: A computer approach to content analysis.

[120] Student. The probable error of a mean. *Biometrika* (1908), 1–25.

[121] Sulis, E.; Farías, D. I. H.; Rosso, P.; Patti, V.; Ruffo, G. Figurative messages and affect in twitter: Differences between# irony,# sarcasm and# not. *Knowledge-Based Systems 108* (2016), 132–143.

[122] Thelwall, M.; Buckley, K.; Paltoglou, G.; Cai, D.; Kappas, A. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology 61*, 12 (2010), 2544–2558.

[123] Tong, R. M. An operational system for detecting and tracking opinions in on-line discussion. In *Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification* (2001), vol. 1.

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

[125] Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; Welpe, I. M. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsm 10*, 1 (2010), 178–185.

[126] Turney, P. D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics* (2002), Association for Computational Linguistics, pp. 417–424.

[127] Turney, P. D.; Littman, M. L. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS) 21*, 4 (2003), 315–346.

[128] University, P. Wordnet - a lexical database for english: Number of words, synsets and senses, 2018. Disponível em https://wordnet.princeton.edu/documentation/wnstats7wn.

[129] Vapnik, V. *The nature of statistical learning theory*. Springer science & business media, 2013.

[130] Vyrva, N. Sentiment analysis in social media. Master's thesis, 2016.

[131] Wang, D.; Liu, Y. A cross-corpus study of unsupervised subjectivity identification based on calibrated em. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis* (2011), Association for Computational Linguistics, pp. 161–167.

[132] Wiebe, J.; Wilson, T.; Bruce, R.; Bell, M.; Martin, M. Learning subjective language. *Computational linguistics 30*, 3 (2004), 277–308.

[133] Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics bulletin 1*, 6 (1945), 80–83.

[134] Wilson, T.; Wiebe, J.; Hoffmann, P. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (2005), Association for Computational Linguistics, pp. 347–354.

[135] Wilson, T.; Wiebe, J.; Hwa, R. Just how mad are you? finding strong and weak opinion clauses. In *aaai* (2004), vol. 4, pp. 761–769.

[136] Wilson, T.; Wiebe, J.; Hwa, R. Recognizing strong and weak opinion clauses. *Computational intelligence 22*, 2 (2006), 73–99.

[137] Witten, I. H.; Frank, E.; Hall, M. A.; Pal, C. J. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[138] Wong, F.; Tan, C. W.; Sen, S.; Chiang, M. Media, pundits and the us presidential election: Quantifying political leanings from tweets. In *Proceedings of the International Conference on Weblogs and Social Media* (2013).

[139] Yu, H.; Hatzivassiloglou, V. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* (2003), Association for Computational Linguistics, pp. 129–136.

APÊNDICE A – EXEMPLOS DE HASHTAGS UTILIZADAS COM MAIOR FREQUÊNCIA EM CENÁRIO ELEITORAL

Exemplos de *hashtags* utilizadas com maior frequência de uso, identificadas em *tweets* coletados durante os períodos de campanha eleitoral presidencial brasileira e americana, realizadas nos anos de 2014 e 2016, respectivamente. As identificações dos usuários nas menções diretas, contidas nos exemplos abaixo, foram substituídas pelo identificador @USER, com exceção de menções a políticos, empresas, instituições, etc., e os endereços eletrônicos foram substituídos pelo termo URL.

Eleição Brasileira	
Hashtag	Exemplo de *tweet* contendo a *hashtag*
#13rasilTodoComDilma	Dilma assinou compromisso contra trabalho escravo; Aécio não! URL #QueroDilmaTreze #13rasilTodoComDilma
#45AecioConfirma	Dia 26/10/2014 Não Vote Em Branco; Não Anule Teu Voto; VOTE: #45AecioConfirma Quer a verdade sobre o Aécio? ?URL
#Aecio	#EuQueroDebateNaGlobo #Dilma nao quer ir ao debate da Globo! Para nao levar nocaute de #William e #Aecio nojenta falsa e mentirosa
#Aecio45	Entre Dilma e Aécio prefiro Sempree Aécio!! #Aecio45 #Aecio45Presidente
#Aecio45Confirma	@dilmabr O futuro é Aécio Neves 45. Sem duvida o melhor para o Brasil. #Aecio45 #Aecio45Confirma
#Aecio45PeloBrasil	Dilma só entregou 12% do PAC, mas entregou o porto de Cuba! #MudaBrasil #Aecio45PeloBrasil URL

#AecioNever	#AecioNever pelo amor de deus n votem no Aécio
#AecioPelaMudanca	#AecioPelaMudanca Dilma engana vocês.Vai deixar todos s/comida,como já faz seus amiguinhos comunistas na VZLA/Argent URL
#AecioPeloBR45IL	#AecioPeloBR45IL Brasil tinha tudo para entrar no Conselho de Segurança da ONU. Mas apoio de Dilma ao terrorismo, acabou com essa chance.
#Aecioporto	Pesquisa vox populi mostra DILMA com 45% e #Aecioporto com 44%. Vamos vencer pela esquerda. #PorIssoDilma13 #FelizComDilma13
#debatenosbt	@USER @USER isso e feio: #AecioNeves #PSDBsecouSP #DebateNoSBT #MelhorcomDilma13 AECIO BATE EM MULHER URL
#desesperodaveja	@USER #13rasilTodoComDilma Alerta de golpe contra Dilma #DesesperodaVeja #DesesperodaGlobo
#DILMA	#Datafolha/#Ibope: Aprovação de #Dilma e reprovação de #Aécio sobem quatro pontos URL #MenosOdioMaisDilma .....
#Dilma13	@USER Esse é #aécio, o ex plauboy. Diga um NÂO sonoro. Universide pública para todos. Vote #Dilma13 em 26/10.
#Dilma13MaisNordeste	só Dilma *, a estrela 13* para fazer o Brasil andar pra frente ! #Dilma13MaisNordeste #Dilma13PraVencer #Dilma13MudaMais
#Dilma13PraVencer	#Dilma13PraVencer Esta será a atenção que o Gov Tucano dará a Saude no Brasil, é isso Aecio? URL
#Eleicoes2014	Por um #waze com a presidANTA #Dilma narrando...pausadamente como nos debates #Eleicoes2014
#EmTodoBrasilAecio45	URL #QueroDilmaTreze #EmTodoBrasilAecio45 bem que o Aecio avisou ontem.
#ForaDilma	Vaza Dilma, já deu #FORADILMA #FORAPT #DILMACUBANA #DILMANOPAISDASMARAVILHAS
#ForaPT	DILMA TERRORISTA de meia pataca.. (nunca fez nada certo..) continua só pensando em golpe.. URL #ForaDilma #ForaPT

#MelhorcomDilma13	Pelo menos Aécio fez um programa novo. É desrespeitoso repetir sempre as mesmas mentiras. #MelhorcomDilma13 #ProfessoresComDilma
#MenosOdioMaisDilma	#MenosOdioMaisDilma Dilma, coração valente!
#MudaBrasil	Romário assume apoio a Aécio e faz gravação para programa eleitoral! #SouAécio #MudaBrasil URL URL
#PretonoBranco	@OGloboPolitica: #Pretonobranco checa frase de @Dilmabr sobre inflação. #Aecio45PeloBrasil URL URL
#QueroDilmaTreze	Boa noite a todxs. A cada dia, a cada debate mais convicto da vitória da Dilma. #MenosÓdioMaisDilma, #QueroDilmaTreze. Dia 26 é nois.
#SomosTodosDilma	#SomosTodosDilma É Dilma 13 de novo, com a força do povo
#VotoAecioPeloBR45IL	Eu voto Aécio 4?5? ? Chega de PT! ???? #VotoAecioPeloBR45IL ????????????

Eleição Americana	
Hashtag	Exemplo de *tweet* contendo a *hashtag*
#AmericaFirst	Hillary Clinton calls Arabs sand ni**ers! #WikiLeak #PodestaEmails2 #MAGA #NeverEverQuit? #AmericaFirst? URL
#Breaking	#BREAKING Early vote numbers in Florida should spook Hillary Clinton #news #Tampa
#Clinton	#Hillary #Clinton Expresses Support For #Fracking In #Wikileaks Document URL
#CrookedHillary	The Dangers Of Hillary Clinton. #TrumpPence16 #Maga #DrainTheSwamp #Neverhillary #CrookedHillary URL
#debate	Donald Trump in the town hall debate stocked Hillary Clinton like a sexual predator! #debate
#debatenight	Hillary underpaid female employees of the Clinton Foundation BIG-LY #debatenight
#DrainTheSwamp	FBI Just Reopened Hillary Clinton Investigation! #DrainTheSwamp URL

#Election2016	Melania Trump to give campaign speech in swing state Pennsylvania ? live: Follow along for the? URL | #Election2016
#FollowTheMoney	Hillary Clinton, two faces of Eve,Scary she will say anything to get elected #debat #followthemoney #Trump URL
#HillaryClinton	#HillaryClinton Begs Forgiveness From #Rothschilds In Leaked #PodestaEmails #Debatenight #debates #debate #JillStein URL
#ImWithHer	sounds fantastic!! #ImWithHer #StrongerTogether Cher campaigns for Hillary Clinton in Flint URL
#MAGA	What Does It Take To Bring Hillary Clinton To Justice? #PodestaEmails28 #NeverHillary #MAGA URLE
#NeverHillary	Hillary Clinton = Awful, Nasty, Repugnant, Vile, Vulgar, Foul, Shameful, Revolting, Scandalous, Dirty, Hateful, Lying, Despicable. #NeverHillary #MAGA
#NeverTrump	#NeverTrump: Hillary Clinton supporters trade votes with third party voters in swing states URL #imwithher
#PodestaEmails	Mysterious Hillary Clinton sign exposes #PodestaEmails and dark secrets URL #tcot
#SpiritCooking	I really HOPE Hillary Clinton is a witch. #spiritcooking
#tcot	Terrorist Murderers Support Hillary Clinton URL #TeaParty #tcot
#Trump	#ChangeYourVote These states allow early voters to #TakeItBack #Hillary to Donald #Trump #MAGA #PodestaEmails25 URL
#TrumpTrain	Hillary Clinton is not the victim. End of story. #MAGA #Trump2016 #Trump #TrumpTrain @Realdonaldtrump @USER URL
#USA	Donald and Hillary in love? Have a look... #USA #election URL
#VOTETRUMP	FBI WIKILEAKS JUST TOOK A DUMP ON HILLARY CLINTON #Wikileaks #LyingClintonCamp #VoteTrump URL

#Wikileaks	New #Wikileaks Emails: #Clinton Campaign Insiders Fear Bill?s Sex Life Could Sink #Hillary URL via @USER

APÊNDICE B – EXEMPLOS DE HASHTAGS POLÍTICAS E NÃO-POLÍTICAS

Exemplos de *hahstags* políticas identificadas em *tweets* coletados durante os períodos de campanha eleitoral presidencial brasileira e americana, realizadas nos anos de 2014 e 2016, respectivamente. As identificações dos usuários nas menções diretas, contidas nos exemplos abaixo, foram substituídas pelo identificador @USER, com exceção de menções a políticos, empresas, instituições, etc., e os endereços eletrônicos foram substituídos pelo termo URL.

Eleição Brasileira		
Categoria de palavras contidas em *hashtag*	Ex. de *hashtag* política	Exemplo de *tweet* contendo *hashtag* a categoria de *hashtag* política
Primeiro nome e/ou sobrenome do candidato	#Aecio45	Entre Dilma e Aécio prefiro Sempree Aécio!! #Aecio45 #Aecio45Presidente
	#Dilma13	Sem Aécio e Lobão, o Brasil vai bater um bolão! #Dilma13 URL
Nome do partido político do candidato	#PT	#PT #Dilma13 URL Conquistas femininas retrocederão com Aécio, diz defensora de direitos humanos
Slogan de campanha do candidato	#MudaBrasil	Dilma só entregou 12% do PAC, mas entregou o porto de Cuba! #MudaBrasil #Aecio45PeloBrasil URL
	#QueroDilmaTreze	Aécio perde batalha da verdade #QueroDilmaTreze URL via @folha_com
Expressão de repúdio contra candidato	#AecioNever	Dilma venceu onde reside e onde nasceu. Aécio perdeu onde nasceu e onde reside. #Dilma13PraVencer #QuemTeConheceNãoVotaJamais #AecioNever

	#ForaDilma	Mais uma do Governo Dilma! E o Brasil só perde com isso! #ForaDilma URL

Eleição Americana		
Categoria de palavras contidas em *hashtag*	Ex. de *hashtag* política	Exemplo de *tweet* contendo *hashtag* a categoria de *hashtag* política
Primeiro nome e/ou sobrenome do candidato	#HillaryClinton	What?s scarier than Donald Trump? Hillary Clinton?s plans to gut Social Security URL #ImWithHer #HillaryClinton
	#Trump	Hillary Clinton is not the victim. End of story. #MAGA #Trump2016 #Trump #TrumpTrain @USER @USER URL
Slogan de campanha do candidato	#MAGA	The truth about #Hillary Clinton... URL #maga
	#ImWithHer	sounds fantastic!! #ImWithHer #StrongerTogether Cher campaigns for Hillary Clinton in Flint URL
Expressão de repúdio contra candidato	#NeverTrump	Hillary Clinton is amply qualified to be president URL #Elections2016 #Hillary #Clinton #ImWithHer #NeverTrump
	#Wikileaks	Bill Clinton?s lover says Hillary schemed to get her to LIE URL #ripjournalism #wikileaks #maga #trump