



UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE ENGENHARIA ELÉTRICA MECÂNICA E DE COMPUTAÇÃO
ENGENHARIA DE COMPUTAÇÃO

GABRIEL RIOS LEMES COSTA

PREDIÇÃO DO RESULTADO DAS ELEIÇÕES PRESIDENCIAIS BRASILEIRAS
UTILIZANDO ANÁLISE DE SENTIMENTOS EM TWEETS

GOIÂNIA
2018
GABRIEL RIOS LEMES COSTA

**TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR
VERSÕES ELETRÔNICAS DE TESES E DISSERTAÇÕES
NA BIBLIOTECA DIGITAL DA UFG**

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

1. Identificação do material bibliográfico: ☐ Dissertação ☐ Tese

2. Identificação da Tese ou Dissertação:

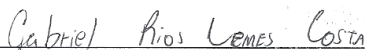
Nome completo do autor: GABRIEL RIOS LEMES COSTA

Título do trabalho: PREDIÇÃO DO RESULTADO DAS ELEIÇÕES PRESIDENCIAIS BRASILEIRAS
UTILIZANDO ANÁLISE DE SENTIMENTOS EM TWEETS

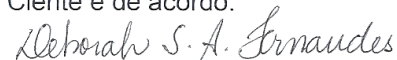
3. Informações de acesso ao documento:

Concorda com a liberação total do documento ☒ SIM ☐ NÃO¹

Havendo concordância com a disponibilização eletrônica, torna-se imprescindível o envio do(s) arquivo(s) em formato digital PDF da tese ou dissertação.


Assinatura do(a) autor(a)²

Ciente e de acordo:


Assinatura do(a) orientador(a)²

Data: 19 / 12 / 18

¹ Neste caso o documento será embargado por até um ano a partir da data de defesa. A extensão deste prazo suscita justificativa junto à coordenação do curso. Os dados do documento não serão disponibilizados durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

² A assinatura deve ser escaneada.

PREDIÇÃO DO RESULTADO DAS ELEIÇÕES PRESIDENCIAIS BRASILEIRAS UTILIZANDO ANÁLISE DE SENTIMENTOS EM TWEETS

Projeto Final de curso, apresentado à
Universidade Federal de Goiás, como parte
das exigências para a obtenção do título de
Bacharel em Engenharia de Computação.

Orientadora: Profa. Dra. Deborah S. A.
Fernandes

GOIÂNIA
2018

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Costa, Gabriel Rios Lemes

PREDIÇÃO DO RESULTADO DAS ELEIÇÕES PRESIDENCIAIS
BRASILEIRAS UTILIZANDO ANÁLISE DE SENTIMENTOS EM
TWEETS [manuscrito] / Gabriel Rios Lemes Costa. - 2018.
50 f.: il.

Orientador: Profa. Dra. Deborah Silva Alves Fernandes.

Trabalho de Conclusão de Curso (Graduação) - Universidade
Federal de Goiás, Escola de Engenharia Elétrica, Mecânica e de
Computação (EMC), Engenharia de Computação, Cidade de Goiás,
2018.

Bibliografia.

Inclui gráfico, tabelas, lista de figuras.

1. Análise de Sentimentos. 2. Machine learning. 3. Twitter. 4.
Predição. 5. Eleições. I. Fernandes, Deborah Silva Alves, orient. II. Título.

CDU 004

ATA DE AVALIAÇÃO DE PROJETO FINAL II

Aos 19 dias do mês de dezembro do ano de 2018, foi apresentado e defendido o Projeto Final II, intitulado Predição do resultado das eleições brasileiras utilizando análise de sentimentos em tweets.

_____ perante a banca examinadora composta pelos membros:

1. Deborah Silva Alves Fernandes (INF), orientador e presidente;
2. Saudreyky Ramos Pires (EMC);
3. Nádia Félix Felipe da Silva (INF).

Após a exposição do trabalho por parte do(s) autor(es), aluno(s) do curso de Engenharia de Computação, foram lhe(s) atribuídas as seguintes notas pelos membros da banca:

Nome do(a) Aluno(a)	Membro 1	Membro 2	Membro 3
Gabriel Rios Leemes Costa	9,0	9,0	9,0
	—	—	—

Nada mais havendo a registrar, eu, Deborah S. A. Limaudes, designado secretário "ad hoc" da banca examinadora, lavrei a presente Ata do ocorrido, a qual, lida e considerada conforme, vai assinada por mim e pelos membros da banca.

Goiânia, 19 de dezembro de 2018.

20 10.

Deborah S. A. Fernandes
Zanduy ten
Nadine Lilise F. da Silve

Agradeço a todos que me acompanharam
nesta jornada.

Dedico a minha família, amigos e à
Professora Deborah por toda colaboração,
apoio e paciência durante o
desenvolvimento deste projeto.

RESUMO

O tema central deste trabalho é a aplicação de análise de sentimentos em dados provenientes de redes sociais online, especificamente o Twitter, com o intuito de prever o resultado do primeiro turno das eleições presidenciais do Brasil de 2018. Utilizar esta aproximação permite a obtenção de análises do cenário político necessitando de tempo e gastos menores quando comparados a meios tradicionais (e.g. entrevista boca de urna). Diferentes de classificadores de sentimento foram avaliados e utilizando aquele apresentou melhores resultados fez-se uma análise quantitativa da distribuição de sentimentos nos documentos coletados, comparando-os com meios tradicionais de pesquisa.

Foi suposto que os dados advindos de redes sociais se comportariam como bons indicadores para predição do resultado da eleição, visto que os usuários e os próprios candidatos se encontram cada vez mais engajados neste meio. Esta suposição foi confirmada uma vez que foi observado que um dos classificadores utilizados no experimento conseguiu acertar a ordem correta em que os três primeiros candidatos ficariam.

Conclui-se que esta abordagem para análise de cenários políticos trata-se de uma área de estudo muito frutífera e com grande importância, no entanto ainda são necessários estudos mais aprofundados em como tratar certos tipos de opiniões que muitas vezes são descartadas para a análise e também identificar termos que podem ser excluídos da análise de uma maneira automática.

LISTA DE IMAGENS

Figura 1: Métodos automatizados para realização de análise de sentimentos.....	16
Figura 2: Representação gráfica simplificada de um algoritmo KNN.....	17
Figura 3: Representação gráfica simplificada de um algoritmo SVM.....	18
Figura 4: Representação simplificada de uma rede neural.....	19
Figura 5: Representação da equação <i>sigmoid</i>	20
Figura 6: Estrutura do projeto.....	21
Figura 7: Esquema de coleta e classificação dos tweets.....	21
Figura 8: Interesse sobre candidatos à eleições presidenciais pelo <i>Google Trends</i> ..	25
Figura 9: Termos removidos do Conjunto de Dados.....	26
Figura 10: Quantia de Tweets coletados por dia.....	27
Figura 11: Quantia de Tweets únicos coletados por dia.....	28
Figura 12: Tipos de classificadores disponíveis na ferramenta Orange.....	31
Figura 13: Nuvem de palavras do arquivo de treinamento.....	32
Figura 14: Representação de validação cruzada.....	33
Figura 15: Representação de uma matriz de confusão.....	34
Figura 16: Estrutura para avaliação de cada classificador.....	35
Figura 17: Distribuição de menções por perfil de candidato.....	36
Figura 18: Interesse sobre os candidatos.....	37
Figura 19: Matrizes de confusão de cada classificador.....	39
Figura 20: Curvas ROC de cada classificador.....	39
Figura 21: Matrizes de confusão de cada classificador.....	41
Figura 22. Curvas ROC de cada classificador.....	41
Figura 23: Parâmetros do classificador de regressão logística e rede neural.....	42
Figura 24: Parâmetros do classificador SVM e KNN.....	42
Figura 25: Estrutura para predição de classe.....	43

SUMÁRIO

1. INTRODUÇÃO.....	8
2. REVISÃO BIBLIOGRÁFICA.....	11
3. FUNDAMENTAÇÃO TEÓRICA.....	13
3.1 Análise de Sentimentos.....	13
3.2 Classificadores.....	15
4. EXPERIMENTO.....	21
4.1 Conjunto de Dados.....	21
4.1.1 Coleta.....	22
4.1.2 Pré-processamento.....	25
4.1.3 Rotulação.....	26
4.2. Classificação de dados.....	31
5. RESULTADOS.....	36
5.1 Contagem ingênua.....	36
5.2 Google Trends.....	37
5.3 Classificação automática de tweets.....	38
5.4 Resultados obtidos <i>versus</i> Institutos de pesquisa.....	45
CONCLUSÕES.....	47
REFERÊNCIAS.....	49

1. INTRODUÇÃO

As percepções da realidade, crenças, atividades e comportamentos cotidianos de uma pessoa são, até certo ponto, condicionadas por como aqueles que a cercam pensam, veem e interagem com o mundo. Isso faz com que opiniões tenham um papel primordial na generalidade das relações humanas e também são elementos-chave na definição de nossos comportamentos [Liu 2012].

Opiniões, sentimentos, emoções, atitudes e outros conceitos relacionados são o objeto de estudo da área de Análise de Sentimentos também chamada de Mineração de Opiniões. Trata-se de um campo de estudo que prospera à medida que as Redes Sociais Online (RSO) tornam-se cada vez mais populares e integradas à sociedade [Liu 2012].

Pode-se definir uma RSO como um site que possibilita que um usuário (1) crie um perfil público ou semipúblico dentro de um ambiente definido, (2) conceba uma lista de outros usuários com os quais este tem alguma conexão e (3) visualize e explore sua lista de conexões e também a de outros presentes no sistema [Boyd 2007].

Antes da popularização de RSO, para obter a opinião de alguém seria necessário conversar com esta pessoa ou então receber esta opinião de forma escrita através de livros, cartas ou questionários. Após este passo, para realizar análises computacionais destes dados seria necessário digitalizá-los. Atualmente, as diversas RSO disponíveis na internet contam com uma grande quantidade de usuários - Tabela 1, que a todo instante publicam mensagens contendo opiniões e sentimentos em formato já digital.

Rede Social	Número de perfis ativos em outubro de 2018 (em milhões), segundo Statista, acesso em 12/2018
Facebook	2.234
YouTube	1.900
Instagram	1.000
Twitter	335
LinkedIn	303

Tabela 1: Número de perfis ativos em RSO populares da internet em outubro de 2018, Statista¹.

Coletar e analisar o pensamento do público é objeto de estudo tanto na academia quanto fora dela. Há variadas aplicações para o estudo de opiniões e tendências. Sempre que alguém necessita tomar uma decisão este procura saber qual a crença dos outros. Empresas querem saber o que a coletividade pensa sobre seus produtos, consumidores têm interesse na opinião de pessoas que já possuem um produto antes de realizar uma compra similar e na esfera política, saber o que os outros pensam antes de se posicionar [Liu 2012].

Em [Liu 2007] foi proposto um modelo de análise de sentimentos utilizando dados de blogs para predizer sucessos de vendas. [O'Connor 2010] relacionou sentimento do Twitter com pesquisas de opinião pública. Utilizando dados de blogs [Gruhl 2005] concluiu que o volume do burburinho na web pode ser utilizado para prever picos de interesse em produtos. Os pesquisadores de [Asur and Huberman 2010] utilizaram menções no Twitter para prever sucessos de bilheteria. Em [Lui 2011] índices de interesse por termos utilizados em pesquisas no google foram empregados para predizer os resultados das eleições americanas de 2008 e 2010.

Em [Alves 2015] foram levantados indicadores para compra e vendas de ações, no mercado brasileiro de bolsa de valores, utilizando sentimentos presentes em dados oriundos do Twitter, [Das et al 2018] fizeram o mesmo no mercado estadunidense de bolsa de valores.

1 "Statista." <https://www.statista.com/>. Acesso em 07 Out. 2018.

Observadas as aplicações e importância dos dados de RSO, este projeto propõe o estudo e experimentação de coleta e análise de dados do Twitter para previsão dos resultados das eleições presidenciais brasileiras de 2018. No capítulo 2 será apresentada uma revisão bibliográfica sobre predição de eventos utilizando dados de RSO. No capítulo 3 é apresentada a fundamentação teórica. O experimento realizado é mostrado no capítulo 4. Finalmente o capítulo 5 apresenta e discute os resultados obtidos, enquanto as conclusões e observações finais serão apresentadas no capítulo 6.

2. REVISÃO BIBLIOGRÁFICA

Embora a área de Processamento de Linguagens Naturais (PLN) tenha uma longa história, até o ano 2000 pouca pesquisa havia sido feita na área das opiniões das pessoas [Liu 2012]. Mas, desde então esta tem se tornado produtiva e isso se dá devido a diversos fatores. Primeiramente, trata-se de um campo interessante e com vários desafios para pesquisas que nunca haviam sido abordados. Em segundo lugar, com a popularização das RSO a disponibilidade de informações havia aumentado de maneira estrondosa, grandes volumes de opiniões passaram a ser gerados a cada segundo facilitando a coleta e análise. E em terceiro lugar, houve a atração de investidores devido a grande gama de aplicações para estas pesquisas e soluções.

No ramo de entretenimento [Asur & Huberman 2010] utilizaram cerca de 3 milhões de *tweets* para comprovar que existe uma correlação entre a quantia de atenção que um determinado título de cinema recebe e o quão aclamado este será quando for a público.

Em [Bollen et al. 2011] observa-se uma aplicação para a economia. Neste foram coletados diversos *tweets* e utilizando de dois classificadores de sentimentos verificou-se que o sentimento geral do Twitter se relaciona com o índice *Dow Jones Industrial Average (DJIA)* - baseia-se nas cotações das 30 maiores empresas dos Estados Unidos. Os pesquisadores conseguiram prever alterações no índice *DJIA* com 6 dias de antecedência em alguns casos.

Em aplicações políticas, os autores [Yano & Smith, 2010] analisaram blogs de política para descobrir o motivo de uma postagem atrair maior ou menor atenção e engajamento dos usuários. Além do desenvolvimento de um modelo preditivo, foram levantados alguns tópicos que quando mencionados trazem maior destaque para a postagem.

Coleta e análise de *tweets* que faziam referência a um dos principais partidos ou a um candidato proeminente durante o período das eleições nacionais da Alemanha foram feitos por [Tumasjan et al. 2010]. Os pesquisadores verificaram se o volume de mensagens aludindo a um determinado partido ou candidato se

relaciona diretamente com a quantia de votos recebidos por este, com resultados próximos ao de pesquisas eleitorais tradicionais.

Em [Filho e Garcia 2014] e [Yaqub et al. 2017] os autores reproduziram o experimento de [Tumasjan et al. 2010] mas nos contextos das eleições presidenciais do Brasil de 2014 e presidenciais americanas de 2016, respectivamente. Utilizando contagem simples de *tweets* ambos obtiveram sucesso ao encontrarem porcentagens de menções próximas às percentagens de votos recebidos, confirmando a relação entre o volume de interesse por um candidato ou partido e quantia de votos que este recebe no dia da votação. [Yaqub et al. 2017] também apresenta como os autores delinearão a opinião da população em relação ao cenário político utilizando análise de sentimentos, ao perceber que a maioria dos *tweets* coletados possuíam uma conotação negativa os autores concluíram que a população se encontrava descrente com a política, em [*Business Insider*, 2016] foi feito um levantamento de sentimentos utilizando meios tradicionais de pesquisa de opiniões e o mesmo fenômeno foi reportado.

Os pesquisadores de [Hamling & Agrawal 2017], analisando os dados da eleição presidencial americana de 2016, utilizaram a análise de sentimentos em *tweets* para mapear, por estado, qual era o candidato que recebeu a maior porcentagem de *tweets* positivos, e foi percebido que na maioria dos casos em que isto ocorreu o candidato também venceu em quantia de votos naquele estado. No artigo [Karami et al 2018] os pesquisadores utilizaram *tweets* mencionando diferentes tópicos relacionados a economia, para descobrir qual o sentimento da população com relação aos candidatos das eleições presidenciais americanas de 2012. Os resultados obtidos foram comparados com pesquisas tradicionais feitas pelo *Pew Research Center*² e foram obtidos resultados semelhantes em 80% dos casos abordados.

2 "Pew Research Center." <http://www.pewresearch.org/>. Acesso em 15 Nov. 2018

3. FUNDAMENTAÇÃO TEÓRICA

O foco de PLN é o estudo das capacidades e limitações de uma máquina compreender a linguagem humana, uma subárea deste campo de estudo trata-se da análise de sentimentos ou mineração de opiniões, onde o objetivo é classificar um documento, frase, *post* ou outra expressão de opinião como positiva ou negativa [Liu 2012].

3.1 Análise de Sentimentos

O estudo de mineração de opiniões trabalha, geralmente, com expressões de opinião que se encaixam em um dos três seguintes níveis: documento, sentença ou entidade e aspecto [Liu 2012]. Utilizando o nível de documento, assume-se que cada documento expressa uma opinião dirigida a somente um objeto e a tarefa a ser realizada é concluir se o sentimento total do documento é positivo ou negativo.

Em nível de sentença, separa-se o documento em diversas sentenças, seguindo a estrutura da linguagem utilizada neste, cada sentença então é analisada separadamente como positiva ou negativa, este processo está extremamente relacionado com classificação de subjetividade que distinguem frases em que são expressadas informações factuais (sentenças objetivas) de frases onde são expressadas opiniões e sentimentos (sentenças subjetivas) [Liu 2012].

O terceiro nível de análise não tem ligação com a estrutura da língua ou do documento, a análise é feita em cima da opinião pura. Para [Liu 2012], uma opinião é definida a partir de uma quintupla (e, c, s, d, t) , onde (e) é o nome de uma entidade, (c) é uma característica de (e) , (s) é o sentimento sobre a característica (c) da entidade (e) , (d) trata-se do dono deste sentimento e (t) o tempo em que a opinião é expressada. O sentimento (s) é positivo, negativo ou neutro ou então expressado através de diferentes intensidades (e.g. uma nota de 1-10 estrelas). A junção de (e) e (c) definem o alvo da opinião.

Uma vez que todo o documento é destrinchado em opiniões é possível dar início a análise definindo se cada quintupla encontrada representa um sentimento positivo ou negativo sobre seu alvo.

É importante observar que estas maneiras de análise são efetivas somente quando trata-se de opiniões simples (Alvo + Sentimento), opiniões comparativas possuem análises mais complexas, e.g. “Coca-Cola é melhor que Pepsi”, nesta frase existem dois alvos da opinião e o sentimento expresso depende de qual alvo está sendo considerado. Tendo isto em vista, este trabalho abordará somente opiniões simples.

Segundo [Liu 2012], para identificar a polaridade e intensidade de sentimentos utiliza-se como indicadores principais, as chamadas palavras de opinião (PO) ou palavras de sentimentos - termos e expressões comumente utilizadas para demonstrar positividade (e.g. Bom, Ótimo, Fantástico) ou negatividade (e.g. Ruim, Péssimo, Horrível). Uma lista destas palavras é chamada de *Lexicon* de Sentimentos, e embora sejam importantes e necessários para análise de sentimentos, utilizar somente *lexicons* para classificar opiniões é insuficiente pois existem dificuldades inerentes à PLN que se manifestam nesta tarefa.

1. Inversão da orientação de uma PO dependendo do contexto, por exemplo a palavra “ruim” geralmente indica uma opinião negativa, e.g. “Este celular é muito ruim”, mas também pode ser utilizada para expressar algo positivo. e.g. “Não consegui encontrar nada de ruim nesse filme”.
2. Frases contendo PO podem não expressar nenhuma opinião, este efeito ocorre bastante em frases interrogativas (e.g. “Onde posso comprar um celular bom?”) ou condicionais (e.g. “Se o lanche estiver bom eu vou repetir”).
 Importante notar que existem frases interrogativas e condicionais que ainda sim expressam opinião, e.g. “Alguém sabe o que há de errado com esse celular horrível?” e “Se você quiser comer bem, vá no outro restaurante”.
3. Presença de sarcasmo e ironia, com ou sem PO, são frases de difícil classificação, e.g. “Que celular maravilhoso, quebrou com 2 dias de

uso”, parte da estrutura da frase pode ser identificada como um sentimento positivo, mas a opinião expressada é negativa.

4. Frases sem nenhuma PO podem também representar opiniões, sentenças com esta característica geralmente são utilizadas para expressar informações factuais, e.g. “Depois de comer naquele restaurante eu passei mal”, mesmo que o que foi exposto seja um fato e não a opinião do interlocutor, observa-se que existe um sentimento negativo com relação ao restaurante.

3.2 Classificadores

A classificação de sentimentos é um processo que pode ser realizado manualmente sem grandes dificuldades. Porém, para realizar predições, reconhecer padrões e tendências da população é necessário analisar documentos provenientes de diversos autores. Como a quantidade de análises necessárias para que a predição seja feita em tempo hábil é grande, faz-se imprescindível o uso de sistemas computacionais para tal tarefa.

Sistemas classificadores podem ser divididos em duas categorias, segundo [Medhat 2014]:

1. Os que utilizam lexemas. Um lexema é uma palavra ou parte de uma palavra que serve de base para seu sentido expresso [Dicionário Priberam da Língua Portuguesa 2018]. Para realizar classificação de sentimentos é utilizada uma lista de lexemas, chamada de *lexicon*, onde além do termo também existe uma nota atribuída a cada um dos lexemas, indicando a polaridade e intensidade do sentimento que aquela expressão representa. O classificador realiza a pesquisa das palavras presentes no documento a ser classificado dentro do *lexicon* e, caso sejam encontradas, somam-se as notas associadas, criando uma nota geral do documento que quanto mais próxima ao mínimo mais negativa é a opinião expressada e quanto mais próxima ao máximo, mais positiva.

2. Os que utilizam aprendizado de máquina. Trata-se de uma ampla categoria que é subdividida em “Aprendizado Supervisionado” (AS) e “Aprendizado Não Supervisionado” (ANS). Algoritmos de AS dependem da existência de exemplos pré-rotulados para realizar o treinamento e com base neste conhecimento classificam novas entradas, os algoritmos que podem realizar esta tarefa são diversos, alguns estão descritos em [Liu 2012] e [Medhat 2014]. Algoritmos de ANS por sua vez são preferíveis quando tem-se dificuldade em obter modelos pré-rotulados, ou então para obter percepções em padrões presentes nos dados disponíveis, estes sistemas não necessitam de exemplos, somente a quantia de classes em que o conjunto de dados devem ser separados, e então procuram por similaridades no conteúdo dos documentos a serem classificados

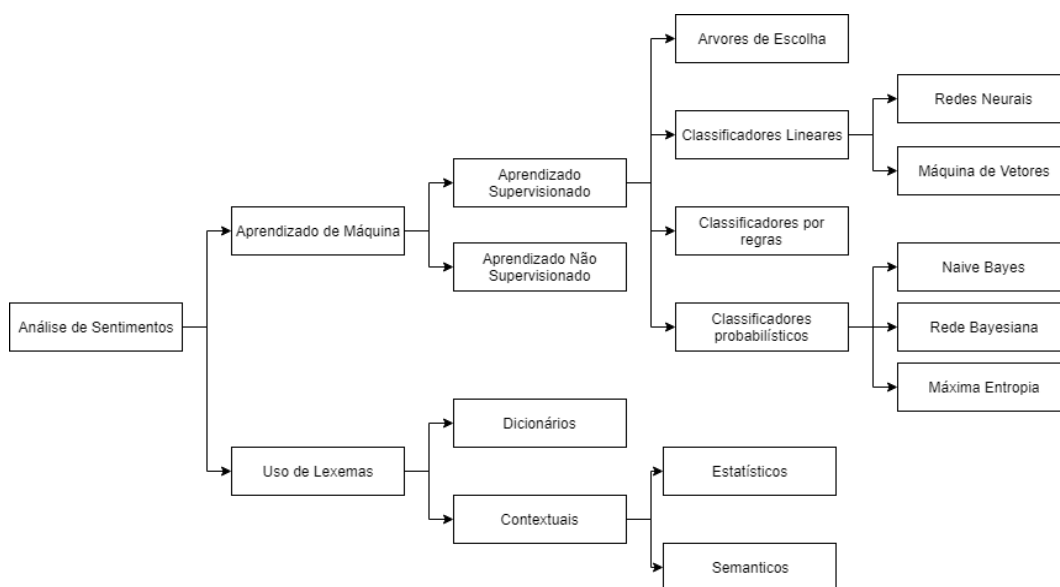


Figura 1: Métodos automatizados para realização de análise de sentimentos..

[Medhat 2014]

Os métodos de análise de sentimentos adotados nos experimentos deste trabalho fazem parte da classe de algoritmos de aprendizado de máquina supervisionado e serão comentados a seguir.

● K-Nearest Neighbors (KNN)

Este algoritmo é de simples entendimento e implementação mas ainda sim produz bons resultados [Daumé III 2012]. Para cada entrada no conjunto treinamento é atribuído um conjunto de coordenadas em um espaço N -dimensional com base nas variáveis independentes que são informadas. Para classificar uma nova entrada não rotulada, o sistema calcula as coordenadas que representam aquela entrada no espaço e classifica esta entrada com base nos K vizinhos mais próximos a ela.

No exemplo exposto na Figura 2, o número de vizinhos é um parâmetro que influencia no resultado da classificação. Neste caso se forem considerados somente 3 vizinhos o ponto vermelho deve fazer parte da classe B, se forem considerados 6 vizinhos o ponto vermelho será rotulado como classe A. O cálculo da distância pode ser feito por qualquer fórmula que calcule distância entre dois pontos no espaço (e.g, Distância Euclidiana, de Manhattan, de Hamming e de Mahalanobis).

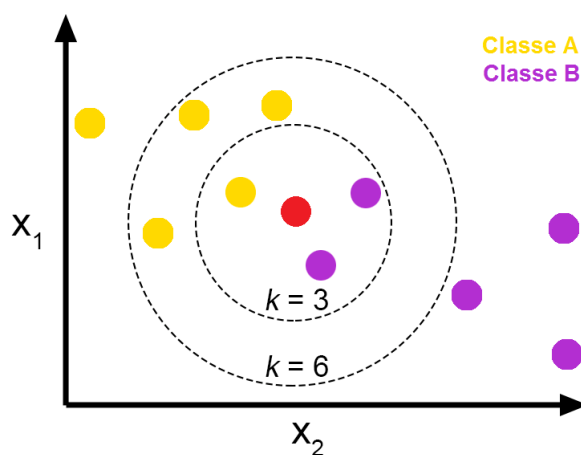


Figura 2: Representação gráfica simplificada de um algoritmo KNN.

[Dugué, 2015]

- **Support Vector Machine (SVM)**

Similar ao algoritmo KNN, segundo [Medhat 2014], as entradas são projetadas em um espaço N -dimensional, mas a função do algoritmo é identificar onde pode ser traçado o melhor hiperplano para separar as classes, como podemos observar na Figura 3. SVMs foram projetadas para lidar com problemas envolvendo duas classes alvo. Mas, existem soluções que expandem o número de classes possíveis e.g. o uso de m -SVMs onde m se refere ao número de classes desejado, cada uma verifica se pertence ou não a classe que ela representa, realizando assim uma separação binomial.

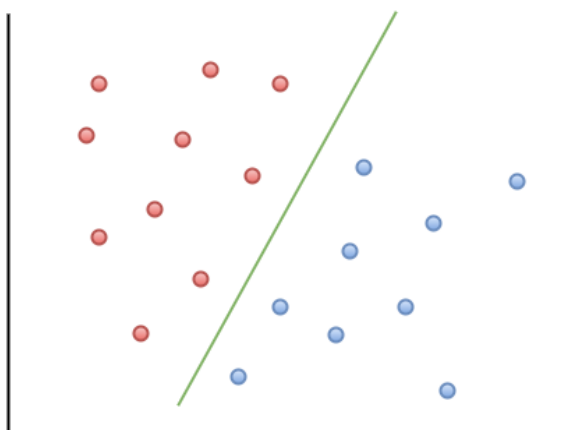


Figura 3: Representação gráfica simplificada de um algoritmo SVM.

[K, 2017]

● Rede Neural (RN)

Sistemas que utilizam RNs funcionam a partir de conexões entre os chamados Neurônio Virtuais [Medhat 2014]. Cada neurônio recebe um conjunto de valores e os multiplica por coeficientes e cada coeficiente indica qual a relevância da variável que ele multiplica para o resultado final.

É feita a soma destas multiplicações e o resultado é aplicado em uma expressão não linear, a saída deste neurônio então é conectada a entrada de outro neurônio e o processo se repete, como está ilustrado na Figura 4.

Para refinar a precisão de sistemas que utilizam RN, são realizados testes nos exemplos já rotulados, com base no erro da predição os valores dos coeficientes são modificados procurando o conjunto que apresenta o melhor resultado possível na predição [Medhat 2014].

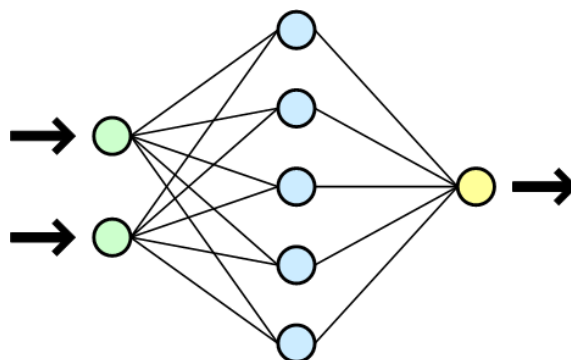


Figura 4: Representação simplificada de uma rede neural.

[Wikipedia 2010]

● Regressão Logística (RL)

Este modelo, segundo [Daumé III 2012], estima a probabilidade de ocorrência de um determinado evento, neste caso a probabilidade de pertencer ou não pertencer a uma determinada classe. Similar às SVMs, RL é um algoritmo desenvolvido para trabalhar com problemas envolvendo duas classes alvo, porém existem soluções para aumentar o número de classes possíveis (múltiplas RLs encadeadas ou interpretar diferentemente a probabilidade, estabelecendo faixas para cada classe).

Para criar um classificador que siga este modelo, assume-se que existe uma relação entre as variáveis independentes e a variável alvo e que esta pode ser representada por uma equação. Atribui-se um peso para cada variável, assim como é feito em RNs, e estes valores são aplicados na equação desejada. O resultado da equação é aplicado em uma função não linear (*sigmoid*, Figura 5), de modo a mapear os valores entre 0 e 1 representando a probabilidade de pertencer ou não a uma determinada classe.

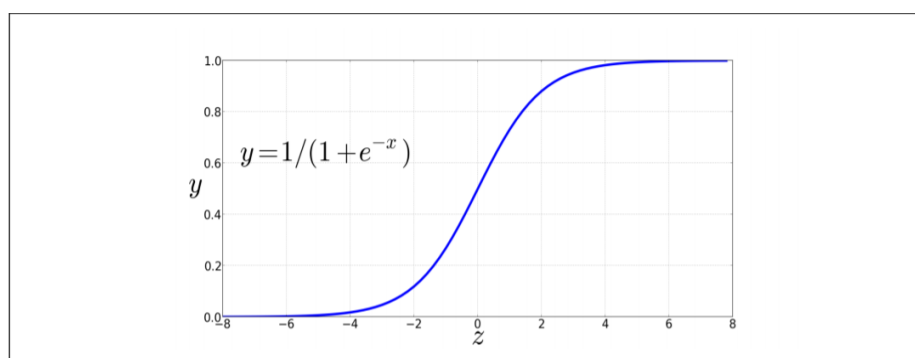


Figura 5: Representação da equação *sigmoid*.

4. EXPERIMENTO

A arquitetura desenvolvida para a realização do experimento pode ser observada na Figura 6. As atividades que compõem os blocos apresentados na figura serão descritas nas subseções abaixo.

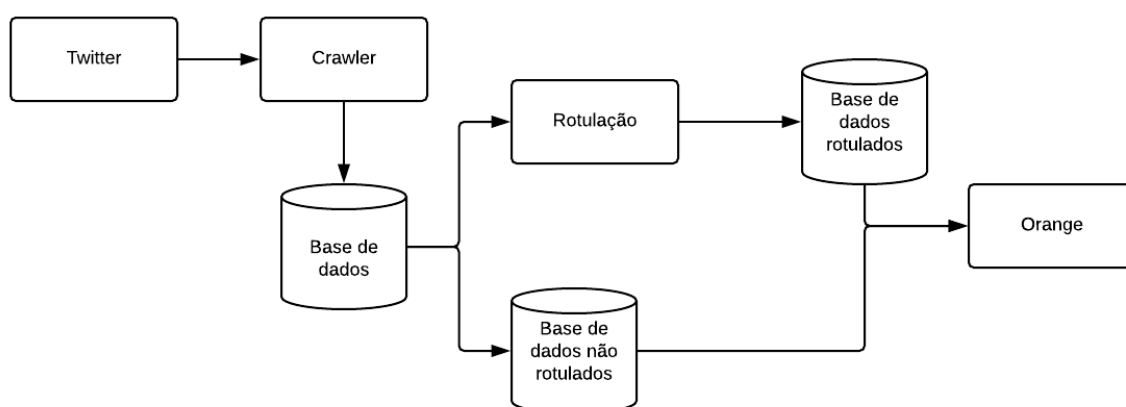


Figura 6: Estrutura do projeto

4.1 Conjunto de Dados

A Figura 7 apresenta os passos realizados para a composição do conjunto de dados utilizado no experimento.

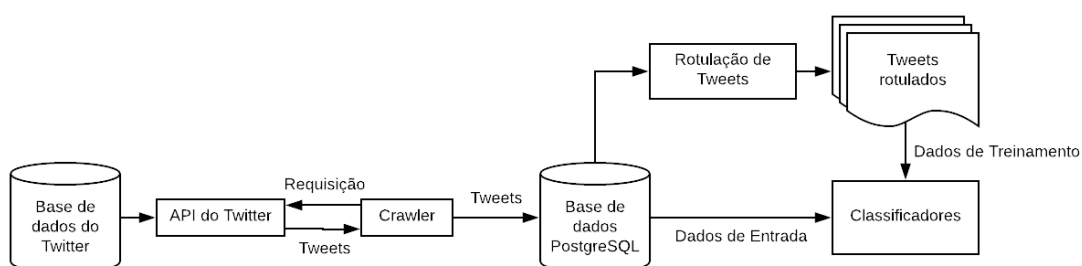


Figura 7: Esquema de coleta e classificação dos tweets.

4.1.1 Coleta

Com o objetivo de realizar uma predição do resultado das eleições presidenciais brasileiras de 2018 foram utilizados dados da RSO Twitter e também de busca do *Google Trends*.

A escolha pelo Twitter em vez de outras RSO com quantias maiores de usuários ativos (vide tabela 1) como Facebook ou Instagram, se deu pelas seguintes características da plataforma:

- ◆ Mensagens de tamanho limitado. O Twitter permite publicação com textos de no máximo 280 caracteres (chamados *tweets*), o usuário então deve ser breve e direto no que quer dizer tornando a análise do sentimento expresso uma tarefa mais simples;
- ◆ Os perfis são, em sua maioria, abertos, permitindo que todos visualizem os *tweets* publicados sem a necessidade de entrar em contato com o usuário;
- ◆ Existe grande atividade na rede quando se trata de assuntos de cunho político, como foi relatado em [G1 2014], [The New York Times 2016], [USA Today 2016], [Twitter Blog 2016] e [Twitter Blog 2018];
- ◆ A academia já se mostrou interessada em utilizar o Twitter combinado a mineração de opiniões para predizer eventos, e.g. [Tumasjan et al. 2010], [O'Connor et al 2010], [Birmingham and Smeaton 2011], [Filho e Garcia 2014] e [Yaqub et al. 2017].

Para coleta de *tweets* foi utilizado um programa denominado de *crawler*³. Este programa realiza requisições para uma API de pesquisa disponibilizada pelo próprio Twitter [Developer Twitter 2018]. Nele é informando um conjunto de *hashtags*, termos e usuários que faziam referência aos candidatos à presidência do Brasil, a partir desses dados o coletor encontra os tweets. As mensagens foram armazenadas em um banco de dados PostgreSQL⁴.

Assim como em [Tumasjan et al 2010] e [O'Connor et al 2010], os termos utilizados nas coletas do Twitter para este experimento continham o nome dos

3 Desenvolvido por participantes do projeto "Observatório da Dinâmica humana nas redes sociais online em língua portuguesa" Pode ser acessado em:

<https://github.com/ProfaDeborah/ObservatorioODH>

4 "PostgreSQL." <https://www.postgresql.org/>. Acesso em 01 Dez. 2018.

candidatos proeminentes [Veja, 2018], as siglas dos partidos os quais estes são filiados e o nome de perfil do Twitter de cada um. Também foram consultadas *hashtags* que demonstravam apoio ou rejeição a algum usuário, e.g. “#Bolsonaro17” e “#elenao”. Na Tabela 2 está disponível a listagem completa de termos utilizados.

Tweets que apresentavam termos referentes a dois candidatos diferentes foram desconsiderados, isto foi feito devido a maior complexidade em analisar opiniões com múltiplos alvos, como dito anteriormente.

O crawler foi iniciado no dia 05/09/2018 e manteve-se ativo até o dia 07/10/2018, dia da votação do primeiro turno, totalizando 2.512.906 *tweets* capturados. Dentre estes, uma quantia de 2.340.943 mensagens (93.31% do total) menciona diretamente o perfil de algum dos 8 candidatos principais da disputa.

Nome do Candidato	Perfil do Twitter	Hashtags Associadas
Alvaro Dias	@alvarodias_	#Alvaro, #alvaro2018, #alvaropresidente
Ciro Gomes	@cirogomes	#CiroPresidente12, #Ciro12, #EuQueroCiro
Fernando Haddad	@Haddad_Fernando	#Haddad, #haddad2018, #haddad13, #haddadELula, #LulaEHaddad
Geraldo Alckmin	@geraldoalckmin	#Alckmin2018, #Alckmin, #GeraldoAlckmin
Guilherme Boulos	@GuilhermeBoulos	#boulos, #boulos2018, #vamoscomBouloseSônia
Henrique Meirelles	@meirelles	#Meirelles, #MeirellesPresidente, #henriquemeirelles
Jair Bolsonaro	@jairbolsonaro	#bolsonaro, #jairbolsonaro, #BolsonaroPresidente17, #elenao
Marina Silva	@MarinaSilva	#marinasilva, #marinapresidente, #marina2018, #elasim

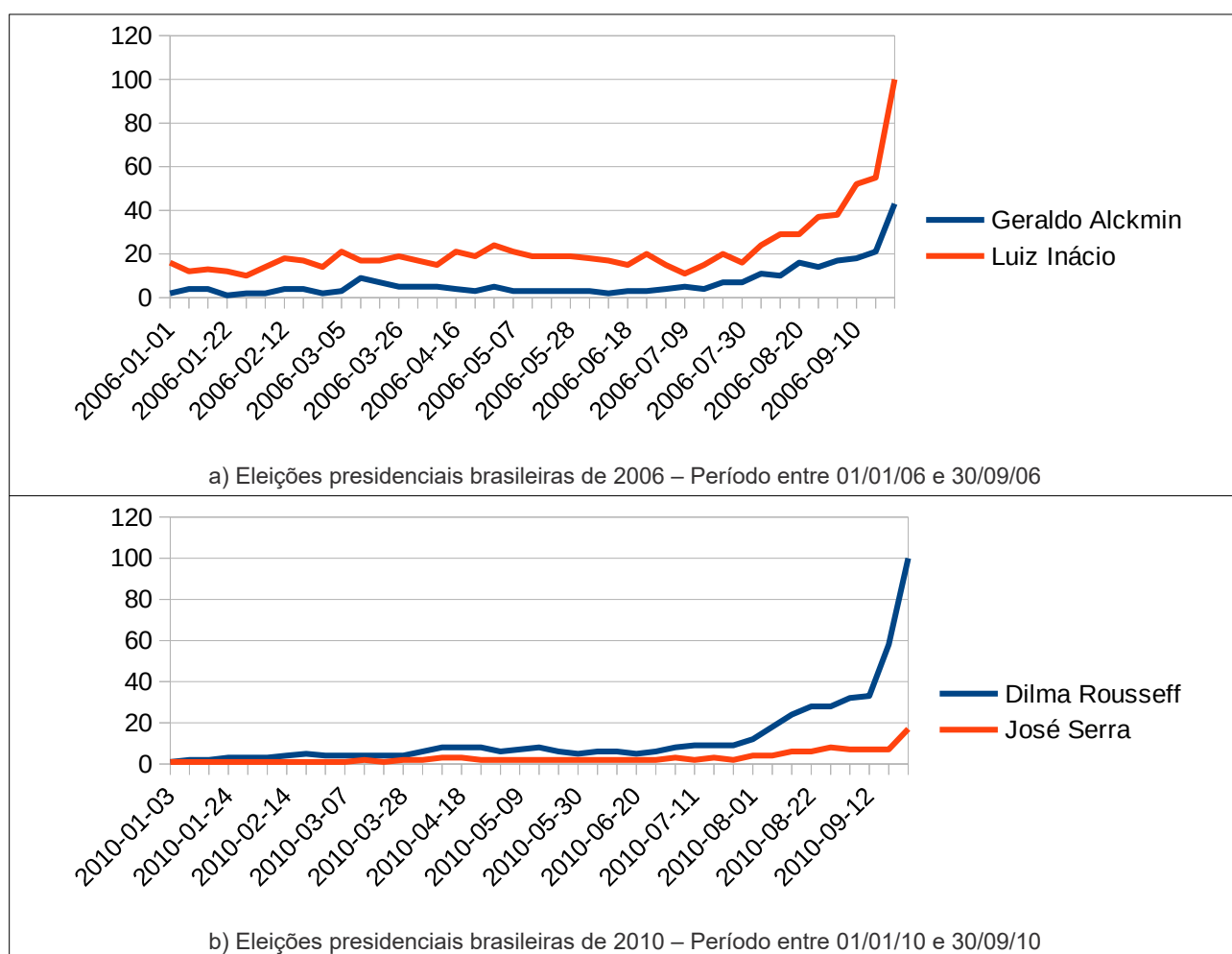
Tabela 2: Lista de candidatos e termos.

Além da avaliação de sentimentos em tweets foram realizados alguns testes com a ferramenta *Google Trends*⁵ a fim de verificar sua performance com relação à eleições passadas e a atual. Criado em 2006, o site disponibiliza acesso às taxas de buscas feitas no *Google* que refletem o interesse sobre determinados termos. É possível comparar entre dois ou mais termos qual foi o mais pesquisado em um determinado intervalo de tempo. Em [Preys et al. 2010] os autores adotaram esta ferramenta para correlacionar o sucesso de uma empresa com o interesse pela mesma.

Uma amostra de como funciona o *Google Trends* e que seus dados podem ser relevantes para pesquisa pode ser visualizada através da Figuras 8 (a)-(d).

5 "Google Trends." <https://trends.google.com/trends/>. Accessed 14 Dec. 2018.

Nestas os nomes dos dois principais candidatos de cada eleição dentre as seguintes: eleições presidenciais do Brasil de 2006, 2010, 2014 e eleições presidenciais dos Estados Unidos de 2012 e 2016 foram pesquisados. O período considerado foi desde o dia primeiro de janeiro do ano da eleição até o último dia do mês anterior ao que ocorreria a votação. Percebe-se, pelas figuras apresentadas que todos os candidatos vencedores naquela época possuíam um índice de interesse médio maior, durante o ano, em relação a seu oponente. Além disso, é possível perceber que existe um aumento no interesse sobre os candidatos à medida que o dia da votação se aproxima.



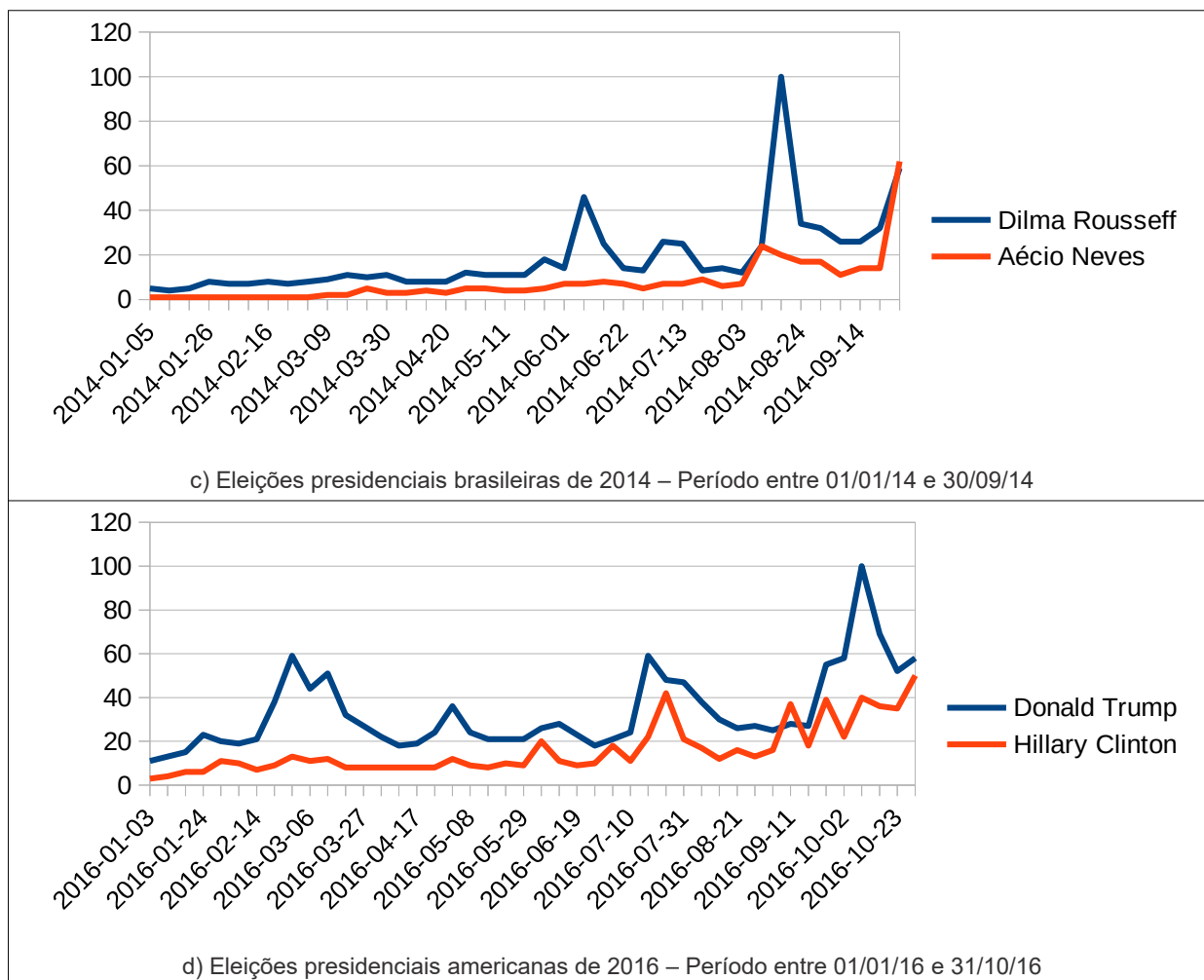


Figura 8: Interesse sobre candidatos à eleições presidenciais pelo *Google Trends*.

4.1.2 Pré-processamento

Assim como adotado em [Prasetyo 2014], para melhorar o desempenho dos classificadores foi realizado um pré-processamento dos *tweets*. Devido à natureza informal das RSO, muitas regras e padrões da língua portuguesa não são considerados. Tal característica, além de outras, polui os dados tendo haja vista que diversas expressões escritas de maneira diferente possuem o mesmo significado. Sob essa perspectiva, os seguintes passos foram adotados para pré-processamento:

- Colocar todas as letras como minúsculas, e.g., “Bom” e “bom” são a mesma palavra, então faz-se com que o classificador receba apenas uma palavra – “bom”.

- Remoção de URLs, não é objeto de estudo acessar o conteúdo das páginas referenciadas pelos endereços, de forma que para o experimento são informações inúteis.
- Remoção de pontuação, acentos e caracteres especiais como “ç”. Poucos usuários utilizam estes recursos e quando isto ocorre, não seguem padrões linguísticos. Para evitar confusão no sistema decidiu-se por eliminar todas as ocorrências. No caso de acentos e caracteres especiais sempre que possível foi feita a troca para a versão simples da letra, e.g. “ç” e “é” são trocados, respectivamente, por “c” e “e”.
- Remoção de *stopwords*, palavras que possuem aparecem em documentos pertencentes a diversas classes e devido a isso podem não ser relevantes para a análise [John Wilbur & Sirotkin 1992]. Este processo também foi adotado em [Asur & Huberman 2010] e [Alves 2015]. A lista das palavras removidas pode ser observada na Figura 9.

<p>11, 6, a, ao, ai, aquela, aquele, aquilo, as, ato, coisas, com, da, de, do, e, ela, elas, ele, eles, em, eu, esse, estao, isso, isto, ja, na, nas, ne, neh, no, nos, o, os, ou, por, pra, q, que, se, ser, so, sua, ta, te, ti, to, tu, tem, um, uma, vai, vem, vc, vcs, voce, vocês</p>

Figura 9: Termos removidos do Conjunto de Dados.

É importante destacar que mesmo com este pré-processamento, os símbolos conhecidos como *Emoticons* (ou *Emojis*) foram mantidos no texto, visto que são comumente utilizados para expressar sentimentos.

4.1.3 Rotulação

Classificadores com aprendizado supervisionado necessitam de um conjunto de dados que já esteja rotulado corretamente para aprender as características que definem cada classe. Para montar o conjunto de treinamento foi identificado quantos tweets foram capturados por dia desde o início da execução do crawler (05/09/2018) até o dia da votação do primeiro turno (07/10/2018), este período compreende aproximadamente 5 semanas.

Foi feito um levantamento para descobrir em quais dias houveram a maior quantia de capturas de cada semana, é possível ver a distribuição das quantias na Figura 12 e os dias com maior quantia de *tweets* estão destacados na Tabela 4. De cada um dos 5 dias com mais publicações na semana foram escolhidos, aleatoriamente, 400 mensagens sem repetições, totalizando 2000 exemplos.

Como estes *tweets* seriam utilizados para montar a base de treinamento, é interessante que fossem exemplos os mais diversos possíveis evitando ao máximo repetições. Para diminuir aumentar a disponibilidade de exemplos diferentes para o treinamento, foi feito o levantamento de quais os dias em que houve maior coleta de *tweets* únicos, a distribuição da quantia de *tweets* únicos por dia pode ser visualizada na Figura 10 e os dias com maior quantia de *tweets* únicos estão destacados na Tabela 3.

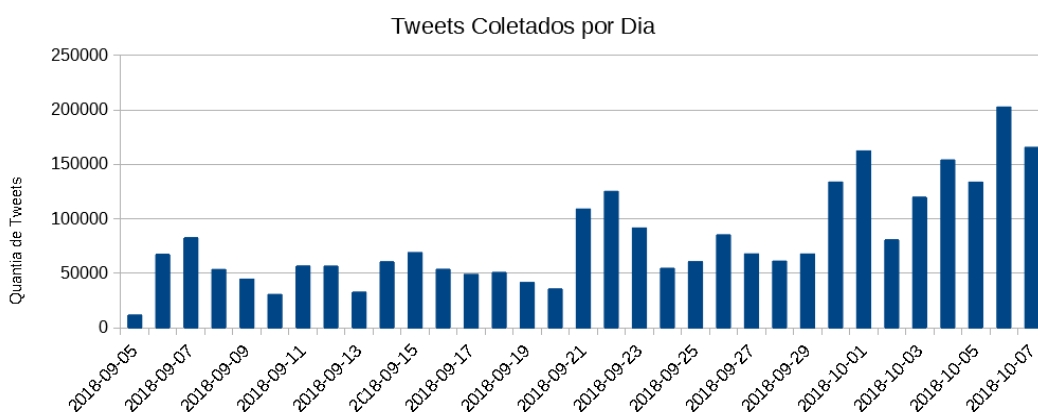


Figura 10: Quantia de Tweets coletados por dia.

Dia	Quantia de Tweets Coletados
07/09/2018 - Sexta	82.875
15/09/2018 - Sábado	69.196
22/09/2018 - Sábado	125.327
23/09/2018 - Domingo	91.705
06/10/2018 - Sábado	202.741

Tabela 4: Dia de cada semana da coleta no qual obteve-se maior quantia de tweets.

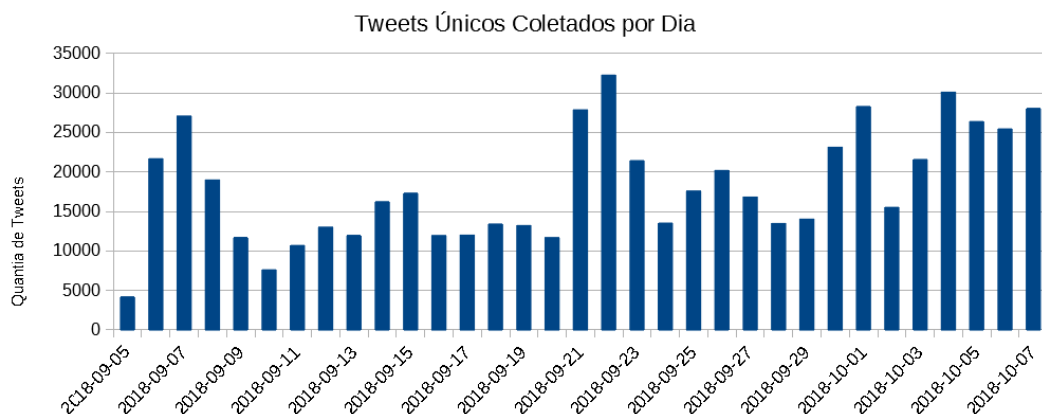


Figura 11: Quantia de Tweets únicos coletados por dia.

Dia	Quantia de Tweets Coletados
07/09/2018 - Sexta	27.091
15/09/2018 - Sábado	17.327
22/09/2018 - Sábado	32.265
23/09/2018 - Domingo	21.437
04/10/2018 - Quinta	30.105

Tabela 5: Dia de cada semana que teve maior quantia de tweets únicos.

Comparando as tabelas 4 e 5 percebe-se que, na maioria das vezes, o dia com maior quantia de tweets também foi o dia com a maior quantia de tweets únicos, a anormalidade identificada no dia 04/10/2018 pode ser justificada pelo fato de que naquela noite estava ocorrendo o último debate presidencial transmitido pela rede Globo [G1 2018].

Decidiu-se pela utilização dos dias em que houve a captura maior quantia de *tweets* únicos para gerar o arquivo de treinamento, mas manteve-se a lógica, foram escolhidos 400 *tweets* aleatórios de cada dia, sem repetições, totalizando 2000 mensagens.

Quatro voluntários realizaram a rotulação, cada um classificando manualmente 500 *tweets* entre as classes “-1” para *tweets* negativos, “0” para *tweets* neutros e “1” para *tweets* positivos.

Para este trabalho, definiu-se que para um *tweet* ser considerado como positivo o usuário poderia:

- Expressar apoio ou encorajamento claro a um candidato, e.g. “@cirogomes Arrasa!!! Não tenha medo de defender seus ideais.”;
- Concordar com suas propostas, ideais ou falas, e.g. “@Haddad_Fernando Melhor definição, @MichelTemer é fragil no ponto de vista de caráter”;
- Desejar melhoras em questões de saúde, devido ao incidente ocorrido no dia 06 setembro de 2018 (o presidenciável Jair Bolsonaro foi esfaqueado enquanto participava em um ato de campanha na cidade de Juiz de Fora (MG) [Folha de São Paulo 2018]) tweets com este padrão tornaram-se muito comuns, e.g. “Oremos mais ainda por nosso Presidente #JairMessiasBolsonaro Está estável agora, graças a Deus! #ForçaBolsonaro”.

Um tweet negativo por sua vez deveria conter alguma das seguintes características:

- Discordância clara com alguma ideia apresentada, e.g. “@cirogomes @ccexplicit A culpa então é do jornal o globo? Me poupe nos poupe”;
- Ataques pessoais a uma pessoa ou grupo, e.g. “@jairbolsonaro Esses imbecis nem sabe porque que a bandeira é verde e amarela”;
- Perguntas em tom de chacota, e.g. “@geraldoalckmin Kd a merenda?”.

A definição de neutralidade foi a mesma utilizada em [Go et al 2017], caso a sentença pudesse aparecer como um título de matéria em jornal ou revista, ou então em um artigo na Wikipédia, o sentimento apresentado é neutro.

Classe	Quantia de Tweets
-1	853
0	183
1	964

Tabela 6: Distribuição dos tweets rotulados entre as 3 classes.

É possível observar, segundo a Tabela 6, que existe uma diferença entre as quantidades de dados de cada classe. Tal diferença configura um problema de balanceamento entre as classes, enquanto os tweets rotulados como positivos e negativos compunham, respectivamente, 48,2% e 42,65% da base de treinamento a neutra obteve somente 9,15%. Essa situação expõe um ambiente político muito polarizado, como foi noticiado em [Veja 2018] e [BBC 2018].

Uma vez que foi identificada esta discrepância na representação de cada classe e também notado, como relatado em [Huang et al 2014], o fato de que problemas de classificação envolvendo 3 classes são mais complexos em comparação a problemas que utilizam somente 2, foi decidido progredir o trabalho sem utilizar a classe neutra e adotar somente duas classes para a classificação, 1 (sentimento positivo) e -1 (sentimento negativo), como feito em diversos outros trabalhos, e.g. [O'Connor et al 2010], [Speriosu et al 2011], [Huang et al 2014] e [Go et al 2017].

Os exemplos neutros foram removidos do arquivo, e então utilizou-se o método *oversampling* [Schiavoni 2010], para rebalancear o conjunto de treinamento, adicionando 147 *tweets* negativos e 36 *tweets* positivos ao documento, desta forma cada classe apresentou exatamente 1000 exemplares.

4.2. Classificação de dados

Os classificadores foram configurados através do uso da plataforma Orange⁶, [Demsar et al. 2013] a descrevem como um conjunto de ferramentas para análise e mineração de dados que disponibiliza uma interface gráfica baseada na conexão de componentes. O software foi feito utilizando a linguagem Python⁷ de programação e apresenta uma variedade de classificadores para serem utilizados, Figura 12. Neste estudo os adotados são: KNN, SVM, Rede Neural e Regressão Logística.

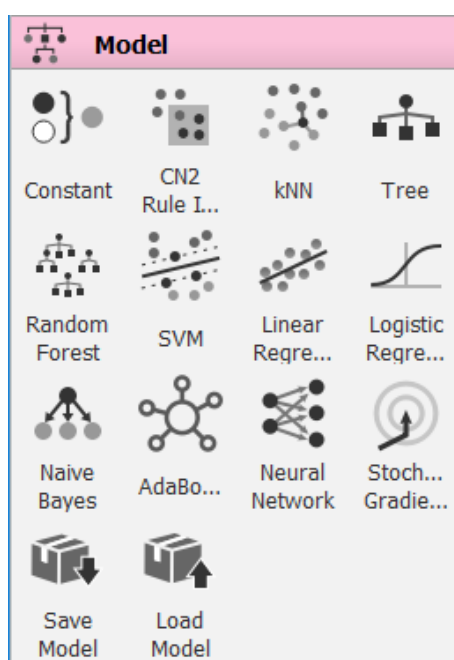


Figura 12: Tipos de classificadores disponíveis na ferramenta Orange.

Para definir qual classificador seria o mais adequado para realizar a predição foi necessário comparar a taxa de acerto de cada um, para isto foram realizados os seguintes procedimentos:

1. Pré-processamento dos *tweets* que compunham a base de treinamento, seguindo o processo discutido no item 4.1.2;

6 "Orange Data Mining." <https://orange.biolab.si/>. Acesso em 01 Dez. 2018.

7 "Python" <https://www.python.org/>. Acesso em 01 Dez. 2018.

3. Os unigramas e bigramas foram utilizados como variáveis independentes, as entradas para o aprendizado de máquina. O treino e teste foram executados em cima da mesma base, aplicando validação cruzada, uma técnica que permite, segundo [Jurafsky & Martin 2018], que todos os exemplos sejam utilizados no treinamento de maneira que uma parte aleatória é utilizada como treino e o restante é utilizado para validação, este processo é repetido um número determinado de vezes (chamado de *folds*), a Figura 14 trata-se de uma representação de uma validação cruzada utilizando 10 *folds*;

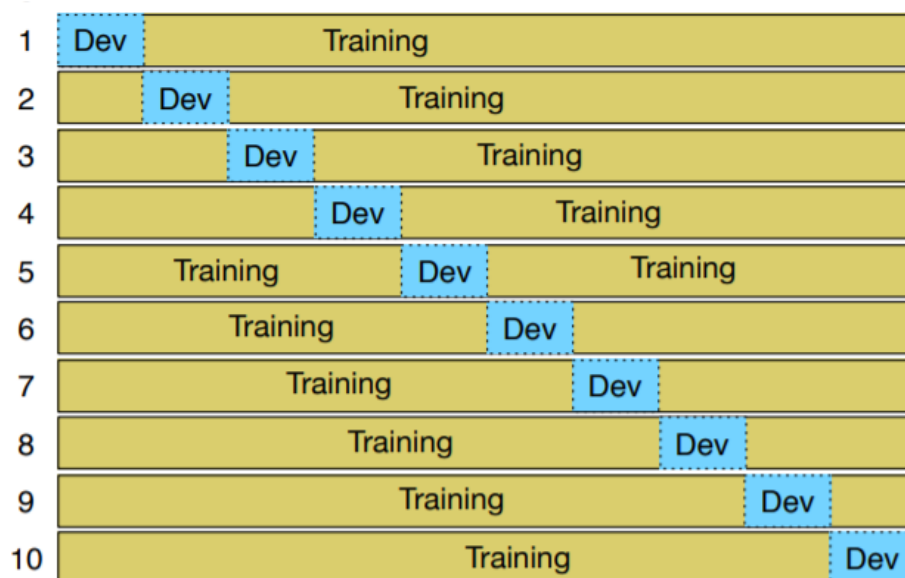


Figura 14: Representação de validação cruzada.

[Jurafsky & Martin 2018].

4. Comparação das métricas Precisão, Revocação, *F-score*, Área sob a curva Roc (AUC) e Acurácia, que segundo [Powers 2007] podem ser definidas utilizando as notações da Figura 15.

	+R	-R	
+P	tp	fp	pp
-P	fn	tn	pn
	rp	rn	1

Figura 15: Representação de uma matriz de confusão.

- Precisão – A relação entre a quantia de verdadeiros positivos (tp) e o total de itens classificados como positivos, (pp) :

$$\left(\frac{tp}{pp}\right) ;$$

- Revocação – A relação entre a quantia de verdadeiros positivos (tp) e o total de itens que realmente são positivos (rp) :

$$\left(\frac{tp}{rp}\right)=tpr ;$$

- F-score* – A média harmônica entre Precisão e Revocação:

$$\frac{2*(Precisão*Revocação)}{(Precisão+Revocação)} ;$$

- AUC – A área de um gráfico formado pela plotagem das porcentagens

de verdadeiros positivos (tpr) pela de falsos positivos ($\frac{fp}{rp}=fpr$), o

valor desta área pode ser definida por:

$$\left(\frac{1+tpr-fpr}{2}\right) ;$$

- Acurácia – A relação entre a quantia de itens corretamente classificados ($tp+tn$) e a quantia total de itens ($tp+tn+fp+fn=N$) :

$$\left(\frac{tp+tn}{N}\right) .$$

Para realizar os passos descritos acima foi montada uma estrutura no Orange (Figura 16) de modo que foi possível comparar o desempenho dos quatro classificadores de maneira simultânea.

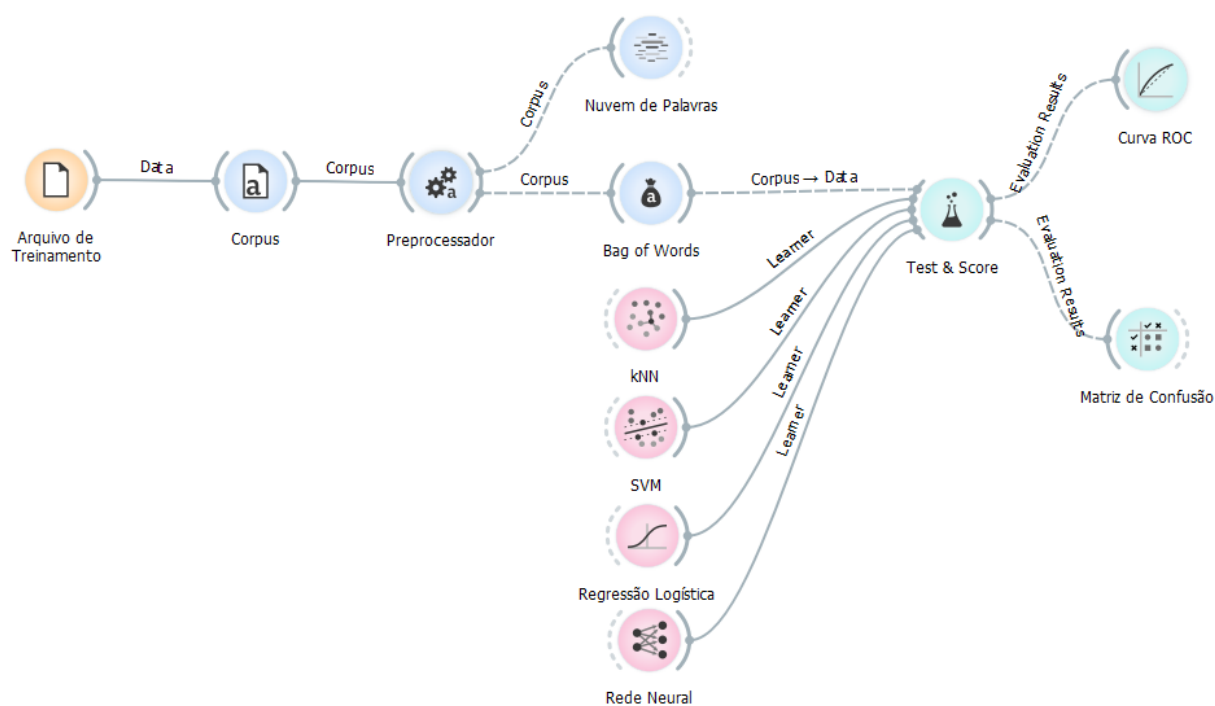


Figura 16: Estrutura para avaliação de cada classificador.

5. RESULTADOS

5.1 Contagem ingênua

Assim como em [Tumasjan et al 2010], neste trabalho adotou-se uma contagem simples de *tweets* para averiguar se a porcentagem de menções dos candidatos se relaciona a quantia de votos recebidos por este. Para tal procedimento foram consideradas somente menções diretas ao perfil do candidato. Dos 2.357.533 de mensagens que faziam menção a algum candidato, o perfil do candidato Jair Bolsonaro foi o mais mencionado com 1.178.641 (o equivalente a 49.9% de todas as menções), em segundo colocado encontra-se o candidato Ciro Gomes com 426.556 menções (18.1%) e em terceiro lugar encontra-se Fernando Haddad com 248.166 menções (10.5%), Figura 17. De acordo com os resultados obtidos da contagem, o candidato Jair Bolsonaro poderia obter a vitória ainda no primeiro turno, e se fosse a segundo turno estaria contra o candidato Ciro Gomes.

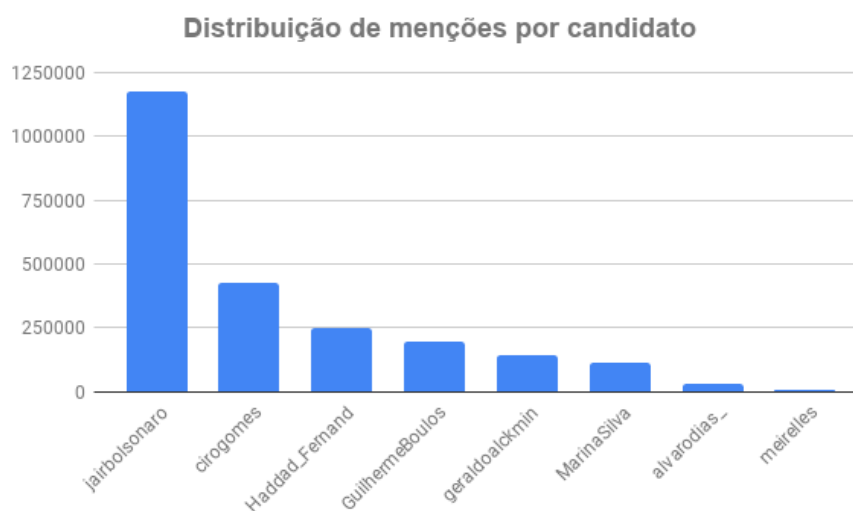


Figura 17: Distribuição de menções por perfil de candidato.

5.2 Google Trends

A fim de realizar a predição do resultado das eleições utilizando o *Google Trends* foram inseridos na ferramenta os nomes dos cinco maiores entres os oito candidatos com maior destaque segundo [VEJA 2018] e considerou-se o período compreendido entre os dias 01/01/2018 e 30/09/2018.

Percebe-se que, de acordo com a Figura 18, Jair Bolsonaro liderou durante todo o ano em quantia de interesse, e que após o incidente ocorrido no dia 06/09/2018 esta vantagem ficou ainda mais evidente, o candidato que se encontra em segundo lugar em relação a quantia de interesse foi Ciro Gomes novamente, nota-se no entanto que a medida que o dia da votação foi aproximando-se o interesse pelo candidato Fernando Haddad ultrapassou o interesse por Ciro Gomes, mas de acordo com a média de todo o ano Ciro continuou em segundo.

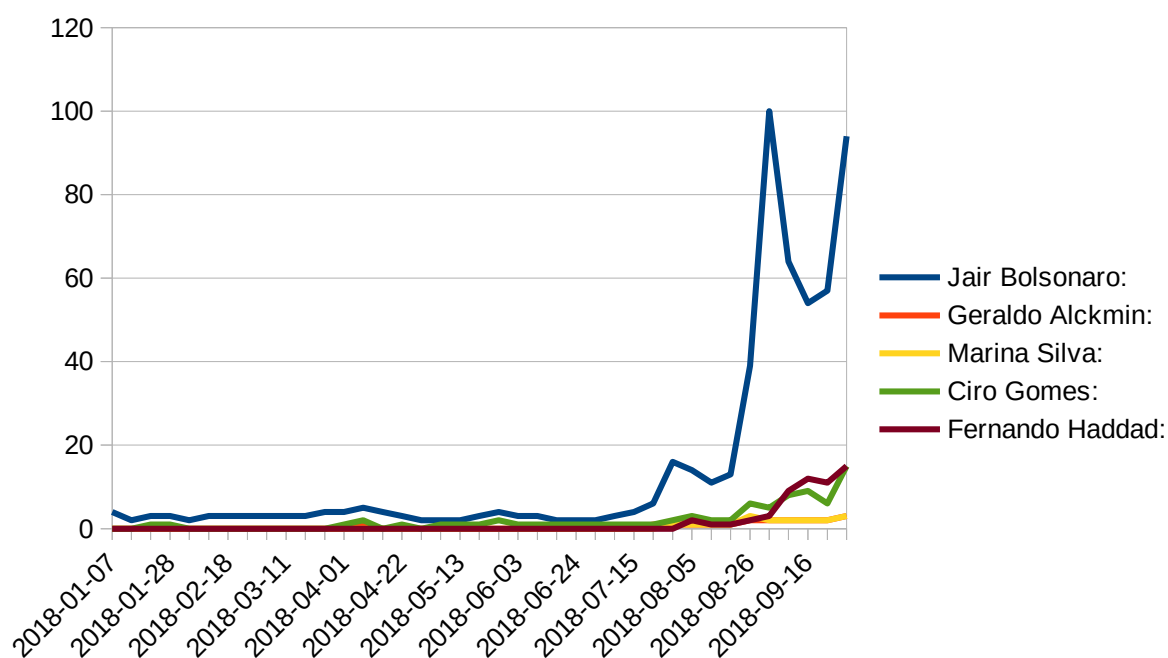


Figura 18: Interesse sobre os candidatos.

5.3 Classificação automática de tweets

Para definir qual classificador seria utilizado, foram aplicados os passos definidos na seção 4.2 deste trabalho. Em iterações iniciais realizadas o classificador que apresentou melhores resultados foi o que utilizou RL, como pode-se observar na Tabela 7.

Método	AUC	Acurácia	F1-score	Precisão	Revocação
Regressão Logística	0.893	0.818	0.816	0.826	0.806
Rede Neural	0.876	0.795	0.797	0.792	0.802
SVM	0.819	0.648	0.723	0.596	0.918
kNN	0.553	0.517	0.673	0.509	0.993

Tabela 7: Métricas das primeiras iterações.

Em [Wilson et al. 2005] define-se que a porcentagem média de concordância em classificação de opiniões entre duas pessoas está entre 82% e 84%, tomou-se como objetivo então encontrar um resultado que possuísse acurácia entre estes valores. Como pode-se observar, a RL obteve um resultado próximo mas ainda não aceitável.

O KNN mostrou-se como o classificador com piores resultados para essa aplicação. Se a métrica Precisão for observada pode-se notar que a quantia de verdadeiros positivos está próxima da quantia de documentos classificados como positivos, isto indica que o classificador está com uma grande taxa de falsos positivos.

A Revocação de todos os classificadores está acima de 0.8, o que mostra que nenhum destes está com uma taxa alta de falsos negativos.

Para auxiliar a análise também foram observadas a matriz de confusão dos classificadores e a plotagem da curva ROC, Figuras 19 e 20 respectivamente.

<div>Predicted</div> <div><div><div>-1</div><div>1</div><div>Σ</div></div><div><div><div>-1</div><div>1</div></div><div><div>830</div><div>170</div></div><div><div>1000</div><div>1000</div><div>2000</div></div></div></div>				<div>Predicted</div> <div><div><div>-1</div><div>1</div><div>Σ</div></div><div><div><div>-1</div><div>1</div></div><div><div>789</div><div>211</div></div><div><div>1000</div><div>1000</div><div>2000</div></div></div></div>			
Regressão Logística				Rede Neural			
<div>Predicted</div> <div><div><div>-1</div><div>1</div><div>Σ</div></div><div><div><div>-1</div><div>1</div></div><div><div>379</div><div>621</div></div><div><div>1000</div><div>1000</div><div>2000</div></div></div></div>				<div>Predicted</div> <div><div><div>-1</div><div>1</div><div>Σ</div></div><div><div><div>-1</div><div>1</div></div><div><div>41</div><div>959</div></div><div><div>1000</div><div>1000</div><div>2000</div></div></div></div>			
SVM				KNN			

Figura 19: Matrizes de confusão de cada classificador.

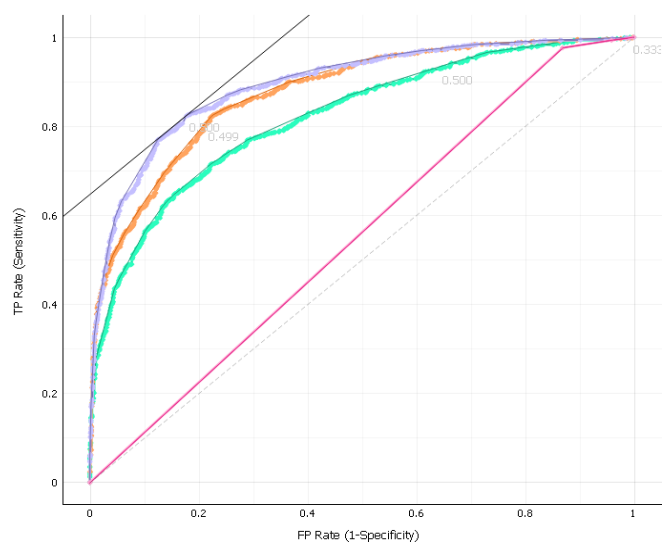


Figura 20: Curvas ROC de cada classificador.

Sendo RL(Azul), RN(Laranja), SVM(Verde) e KNN(Rosa).

Após diversas iterações obteve-se resultados mais satisfatórios, como os indicados na Tabela 8. O classificador RN atingiu acurácia de 84%, e além disto ao observar a métrica AUC é possível perceber que esta se encontra muito próxima a 1, e segundo [Powers 2007] este seria o comportamento de um classificador ideal.

Método	AUC	Acurácia	F1-score	Precisão	Revocação
Rede Neural	0.914	0.840	0.842	0.828	0.857
Regressão Logística	0.908	0.832	0.834	0.821	0.848
SVM	0.863	0.737	0.768	0.686	0.871
kNN	0.816	0.618	0.717	0.569	0.970

Tabela 8: Métricas finais.

É possível notar que houve uma melhora em todos os classificadores, embora disso o KNN continuasse sendo o que possui a pior eficácia, pois manteve os problemas apontados anteriormente, grandes quantias de falsos positivos. Nas Figuras 21 e 22 podem ser observadas a matriz de confusão e a curva ROC, respectivamente. Os parâmetros finais adotados em todos os classificadores podem ser visualizados nas Figuras 23 e 24.

		Predicted			
		-1	1	Σ	
Actual	-1	815	185	1000	
	1	152	848	1000	
	Σ	967	1033	2000	
		Predicted			
		-1	1	Σ	
Actual	-1	822	178	1000	
	1	143	857	1000	
	Σ	965	1035	2000	
		Predicted			
		-1	1	Σ	
Actual	-1	602	398	1000	
	1	129	871	1000	
	Σ	731	1269	2000	
		Predicted			
		-1	1	Σ	
Actual	-1	266	734	1000	
	1	30	970	1000	
	Σ	296	1704	2000	

Regressão Logística

Rede Neural

SVM

KNN

Figura 21: Matrices de confusão de cada classificador.

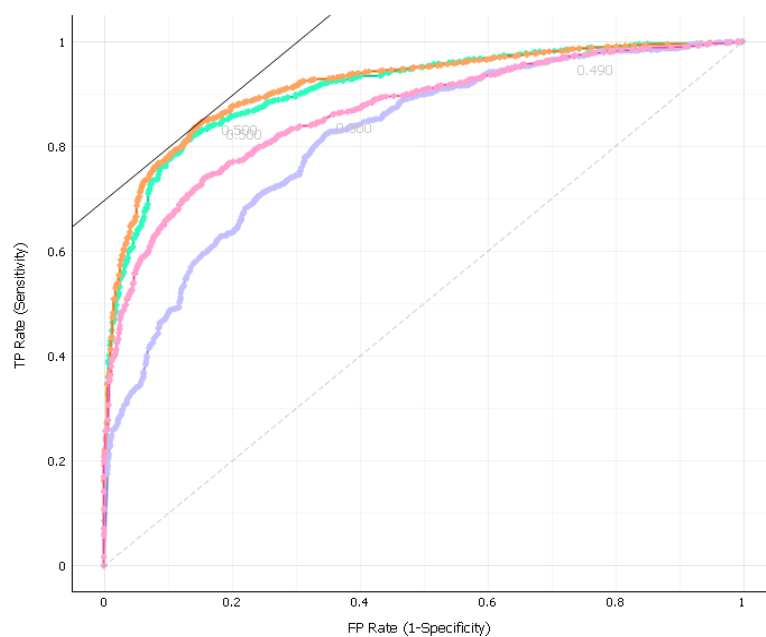


Figura 22. Curvas ROC de cada classificador.

Sendo RL(Verde), RN(Laranja), SVM(Rosa) e KNN(Azul).

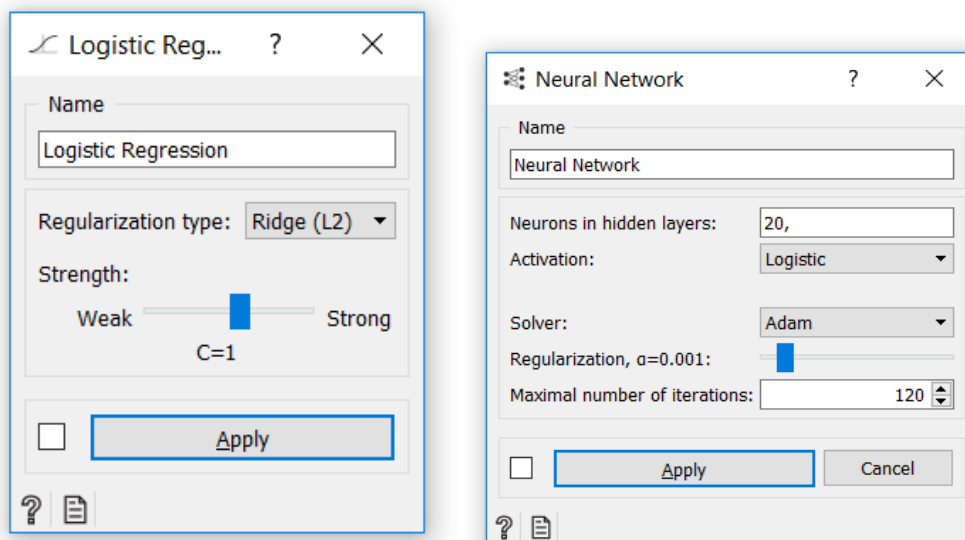


Figura 23: Parâmetros do classificador de regressão logística e rede neural

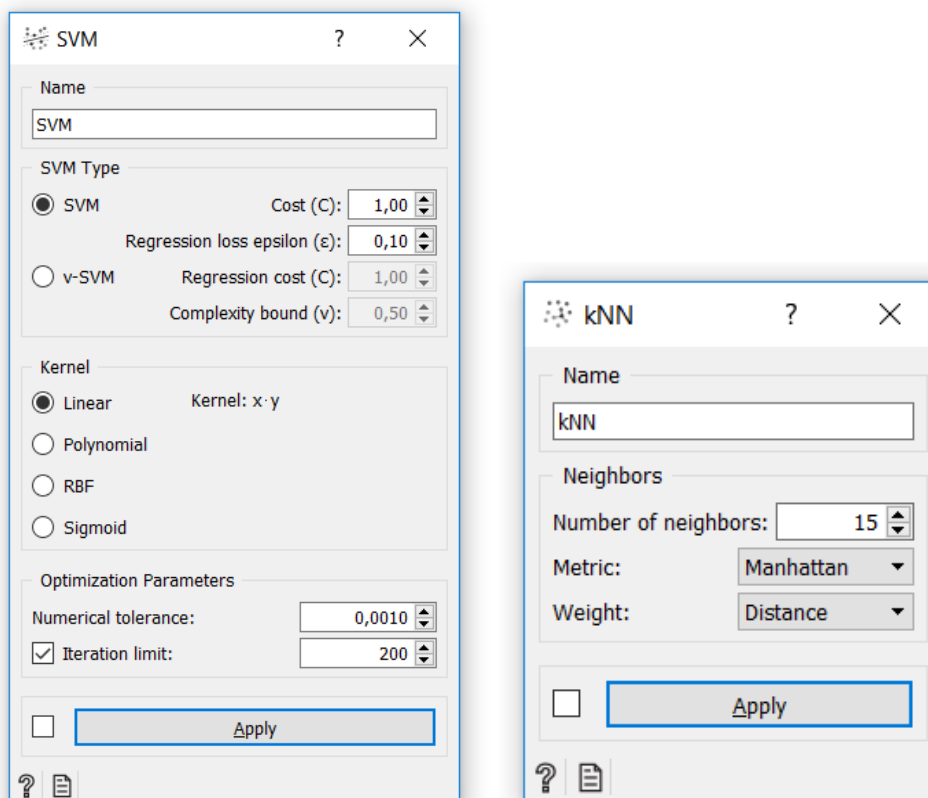


Figura 24: Parâmetros do classificador SVM e KNN.

Uma vez definido que o classificador a ser utilizado seria o RL com os parâmetros exibidos acima, foi montada uma segunda estrutura no Orange (Figura 25) para que pudessem ser realizadas as predições e acessar quantos tweets foram alocados para cada classe.

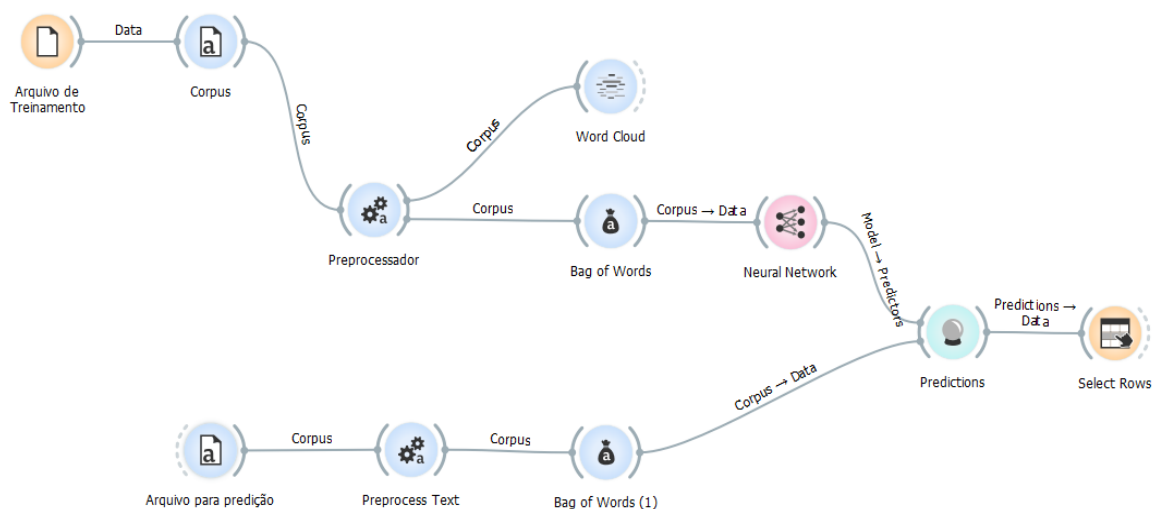


Figura 25: Estrutura para predição de classe.

O mesmo classificador foi utilizado para todos os candidatos, porém a base de dados foi separada por candidato para que fossem obtidos os índices de *tweets* positivos e negativos de forma independente, estas relações podem ser observadas na Tabela 9.

Candidato	Quantia de positivos	Quantia de negativos	Porcentagem de positivos	Porcentagem de negativos
Jair Bolsonaro	912.385	266.255	77,41%	22,59%
Ciro Gomes	308.530	118.026	72,33%	28,67%
Fernando Haddad	184.436	63.729	74,32%	25,68%
Marina Silva	66.449	48.692	57,71%	43,29%
Henrique Meirelles	5.610	5.019	52,78%	47,22%
Alvaro Dias	17.193	16.342	51,27%	48,73%
Guilherme Boulos	98.897	100.332	49,64%	50,36%
Geraldo Alckmin	70.698	89.926	48,58%	52,42%

Tabela 9: Resultados do classificador.

Estes resultados foram ordenados de duas maneiras diferentes, utilizando a quantia de *tweets* positivos (Tabela 10) e também a porcentagem de *tweets* positivos (Tabela 11).

Posição	Candidato	Quantia de <i>tweets</i> positivos
1	Jair Bolsonaro	912.385
2	Ciro Gomes	308.530
3	Fernando Haddad	184.436
4	Guilherme Boulos	98.897
5	Geraldo Alckmin	70.698
6	Marina Silva	66.449
7	Alvaro Dias	17.193
8	Henrique Meirelles	5.610

Tabela 10: Resultados ordenados por quantia de *tweets* positivos.

Posição	Candidato	Porcentagem de <i>tweets</i> positivos
1	Jair Bolsonaro	77,41%
2	Fernando Haddad	74,32%
3	Ciro Gomes	72,33%
4	Marina Silva	57,71%
5	Henrique Meirelles	52,78%
6	Alvaro Dias	51,27%
7	Guilherme Boulos	49,64%
8	Geraldo Alckmin	48,58%

Tabela 11: Resultados ordenados por porcentagem de *tweets* positivos.

5.4 Resultados obtidos *versus* Institutos de pesquisa

Com a finalidade de validar os resultados de predição obtidos com as ferramentas tecnológicas, foram utilizadas pesquisas eleitorais dos institutos Datafolha⁸ e Ibope⁹. Tais agências realizam pesquisas via meio tradicional, através de questionamentos boca a boca feitos dias antes das eleições e também de boca de urna e em períodos próximos à data de votação do primeiro turno. Esses institutos são tidos como confiáveis por empresas jornalísticas e governamentais.

O instituto Datafolha publicou no dia 07/10/2018 o seguinte *ranking*, em ordem decrescente de intenção de votos: Jair Bolsonaro (40%), Fernando Haddad (25%), Ciro Gomes (15%), Geraldo Alckmin (8%), Marina Silva (3%), João Amoêdo (3%), Henrique Meirelles (2%), Alvaro Dias (2%), Cabo Daciolo (1%) e Guilherme Boulos (1%) [Datafolha 2018].

O Ibope por sua vez liberou em [Ibope 2018] a seguinte ordem de classificação, também orientada por ordem decrescente de intenção de votos: Jair Bolsonaro (45%), Fernando Haddad (28%), Ciro Gomes (14%), Geraldo Alckmin (4%), João Amoêdo (3%), Marina Silva (2%), Alvaro Dias (1%), Henrique Meirelles (1%), Cabo Daciolo (1%) e Guilherme Boulos (1%).

O resultado real do primeiro turno pode ser encontrado em [Estadão 2018], e deu-se a seguinte ordenação: Jair Bolsonaro (46%), Fernando Haddad (29.28%), Ciro Gomes (12.47%), Geraldo Alckmin (4.76%), João Amoedo (2,5%), Cabo Daciolo (1,26%), Henrique Meirelles (1,20%), Marina Silva (1%), Alvaro Dias (0.8%) e Guilherme Boulos (0.58%).

8 "Datafolha." <http://datafolha.folha.uol.com.br/>. Acessado em 5 de Dez. 2018.

9 "Ibope." <http://www.ibope.com.br/>. Acessado em 5 de Dez. 2018.

	Real	Datafolha	Ibope	Contagem simples	RN contagem	RN porcentagem	Google Statistics
1	Jair B.	Jair B.	Jair B.	Jair B.	Jair B.	Jair B.	Jair B.
2	Fernando H.	Fernando H.	Fernando H.	Ciro G	Ciro G.	Fernando H.	Ciro G.
3	Ciro G.	Ciro G.	Ciro G.	Fernando H.	Fernando H.	Ciro G.	Fernando H.
4	Geraldo A.	Geraldo A.	Geraldo A.	Guilherme B.	Guilherme B.	Marina S.	Geraldo A.
5	João A.	João A.	Marina S.	Geraldo A.	Geraldo A.	Henrique M.	Marina S.
6	Cabo D.	Marina S.	Alvaro D.	Marina S	Marina S.	Alvaro D.	-
7	Henrique M.	Alvaro D.	Henrique M.	Alvaro D.	Alvaro D.	Guilherme B.	-

Tabela 12: Comparação dos resultados de cada meio de predição.

De acordo com a Tabela 12 onde os resultados foram colocados lado a lado, os dois institutos considerados apresentaram listas bem similares, sendo que o Datafolha foi o que obteve o melhor resultado, acertando até a 5ª posição. Quanto aos métodos alternativos explorados neste trabalho, todos conseguiram acertar que o candidato Jair Bolsonaro ficaria em primeiro lugar, e o classificador Rede Neural utilizando a ordenação por percentual de *tweets* positivos foi a classificação que apresentou os melhores resultados, acertou até a 3ª posição.

CONCLUSÕES

O objetivo do projeto era prever o resultado das eleições com base no Twitter, validando assim se esta é uma fonte de dados que pode ser consultada para realizar pesquisas eleitorais em prazos menores e com menor gasto de recursos. Pode-se afirmar que houve um sucesso parcial neste ponto, as 4 previsões realizadas acertaram os dois seguintes pontos:

- Jair Bolsonaro ficaria em primeiro lugar;
- Jair Bolsonaro, Ciro Gomes e Fernando Haddad iriam compor as três primeiras posições.

Como discutido anteriormente, dos métodos de classificação adotados o de melhor resultado foi a rede neural (na lista ordenada por porcentagem total de tweets positivos) e o pior o KNN. Mesmo possuindo um banco de treinamento balanceado grande parte das entradas eram direcionadas para a classe de sentimento positivo.

O resultado da predição utilizando o *Google Trends* possui uma peculiaridade interessante. É possível observar pela Figura 18 que Fernando Haddad ultrapassou Ciro Gomes à medida que o dia da votação foi se aproximando. Caso o experimento considerasse um período menor de tempo (e.g. somente o mês anterior à votação), esta predição teria acertado as primeiras 4 posições e passaria a ser o método com melhor desempenho. Pode-se supor então que este método é mais efetivo quando utilizado para prever um futuro imediato e levando em consideração um horizonte de dados menor.

Outro ponto interessante a ser observado trata-se da situação dos candidatos Geraldo Alckmin e Guilherme Boulos. Enquanto Alckmin foi o presidenciável que recebeu a maior quantia de *tweets* negativos o mesmo terminou em quarto lugar, já Boulos que recebeu uma quantidade considerável de *tweets* e uma porcentagem de negativos menor que de Geraldo Alckmin, obteve uma quantia pífia de votos. A suposição feita para explicar esta divergência está no público-alvo de cada candidato. O partido de Boulos se identifica com eleitores mais jovens, e Alckmin busca apoio de pessoas mais maduras. Desta maneira, a rejeição no Twitter afetou muito mais a Boulos, visto que foi rejeitado em um ambiente dominado por pessoas

que compõem seu público-alvo, enquanto Alckmin recebeu votos de pessoas que não são tão envolvidas com RSO.

Em questão de dificuldades observadas no projeto, foi notado que a ferramenta *Orange* apesar de extremamente intuitiva e didática possui um grave limitante no que se refere a quantia de documentos que podem ser carregados. Mesmo configurando um acesso a banco de dados este programa necessita de baixar todas as entradas na memória do computador para conseguir proceder com a classificação. Tal situação fez com que diversas vezes ocorressem erros de execução devido ao volume de dados. Para aplicações onde faz-se necessário o trabalho com quantias maiores que 50.000 entradas recomenda-se que seja feita a classificação de outra maneira.

Também foi notada a falta de uma boa lista de *stopwords* da variação da língua portuguesa que é utilizada na internet, as listas encontradas por nós seriam efetivas caso os *tweets* fossem escritos seguindo a norma-padrão, o que não é o fato.

Por fim também foi observado o problema com amostragem e classificação quando utilizada a opinião neutra. No campo político muitas vezes encontra-se um ambiente muito polarizado, com baixa densidade de opiniões neutras ou frases informativas, no entanto o descarte destes exemplos não é a resolução ideal para este problema.

Trabalhar com extração de opiniões a partir de textos publicados em RSOs foi uma experiência fascinante. Para trabalhos futuros nesta área a exploração de como tratar a classe neutra e qual o melhor classificador para esta tarefa é um tópico em aberto e de extrema importância.

A detecção automática de *stopwords* em textos também é uma pesquisa que merece atenção. Devido aos diversos ambientes presentes na internet, a linguagem não seguir regras bem definidas, as gírias, abreviações e expressões não há como criar uma lista absoluta de *stopwords* e existência de uma ferramenta que permitisse uma análise de exemplos procurando por *stopwords* facilitaria diversas outras pesquisas.

REFERÊNCIAS

Alves, D. S. (2015). Uso de técnicas de Computação Social para tomada de decisão de compra e venda de ações no mercado brasileiro de bolsa de valores. Disponível em: <<http://repositorio.unb.br/handle/10482/19345>>

Asur, S., & Huberman, B. A. (2010). Predicting the Future With Social Media. *Comparative Strategy*, 12(2), 141–165.
Disponível em: <<https://doi.org/10.1080/01495939308402915>>

BBC. (2018). Brazil elections: Bolsonaro and Haddad choices before voters. Disponível em: <<https://www.bbc.com/news/world-latin-america-46008907>>.
Acesso em 08 dez. de 2018.

Bermingham, A., & Smeaton, A. F. (2011). On Using Twitter to Monitor Political Sentiment and Predict Election Results. *Psychology*, 2–10.

Bollen, Johan, Huina Mao and Xiao-Jun Zeng. “Twitter mood predicts the stock market.” *J. Comput. Science* 2 (2011): 1-8.

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 210–230.
Disponível em: <<https://doi.org/10.1111/j.1083-6101.2007.00393.x>>

Business Insider. (2016). Almost 80% of Americans say they are angry with the federal government.
Disponível em: <<https://www.businessinsider.com/ap-poll-americans-angry-with-federal-government-happy-at-home-2016-4>>.
Acesso em 10 nov. 2018.

Das, S., Behera, R. K., Kumar, M., & Rath, S. K. (2018). Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction. *Procedia Computer Science*, 132(Iccids), 956–964. Disponível em: <<https://doi.org/10.1016/j.procs.2018.05.111>>

DataFolha. (2018). Bolsonaro tem 40% dos válidos na véspera do 1º turno. Disponível em: <<http://datafolha.folha.uol.com.br/eleicoes/2018/10/1983037-bolsonaro-tem-40-dos-validos-na-vespera-do-1-turno.shtml>>.

Acesso em 08 nov. 2018.

Daumé III, H. (2012). *A Course in Machine Learning*.

Disponível em: <http://ciml.info/dl/v0_8/ciml-v0_8-all.pdf>

Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research* 14(Aug):2349–2353.

Disponível em: <<http://jmlr.org/papers/volume14/demsar13a/demsar13a.pdf>>.

Dicionário Priberam. (n.d.). Definição de lexema.

Disponível em: <<https://dicionario.priberam.org/lexema>>. Acesso em 13 out. 2018.

Dugué, Nicolas (2015). Rechercher une aiguille dans une botte d'un milliard de tweets : les challenges de la fouille du réseau Twitter.

Estadão (2018). Apuração 1º turno

Disponível em: <<https://politica.estadao.com.br/eleicoes/2018/cobertura-votacao-apuracao/primeiro-turno>>.

Acesso 09 out 2018.

Filho, W. D. P., Cristina, A., & Garcia, B. (2014). Predição do resultado das eleições presidenciais do Brasil baseado em tuítes.

Folha de São Paulo. (2018). Bolsonaro leva facada durante ato de campanha em Juiz de Fora.

Disponível em: <<https://www1.folha.uol.com.br/poder/2018/09/bolsonaro-leva-facada-durante-ato-de-campanha-em-juiz-de-fora.shtml>>.

Acesso em 13 de set. de 2018.

G1. (2014). Eleições brasileiras geraram quase 40 milhões de tuítes, diz Twitter.

Disponível em: <<http://g1.globo.com/politica/eleicoes/2014/noticia/2014/10/eleicoes-brasileiras-geraram-quase-40-milhoes-de-tuites-diz-twitter.html>>.

Acesso em 11 nov. de 2018.

G1. (2018). Leia e veja a íntegra do Debate na Globo.

Disponível em: <<https://g1.globo.com/politica/eleicoes/2018/noticia/2018/10/05/veja-a-integra-do-debate-na-globo.ghtml>>.

Acesso em 03 dez. de 2018.

Go, A., Bhayani, R., & Huang, L. (2017). Twitter Sentiment Classification using Distant Supervision.

Gruhl, D., Road, H., Jose, S., Guha, R., Kumar, R., Road, H., ... Tomkins, A. (2005). The Predictive Power of Online Chatter Categories and Subject Descriptors.

Huang, M., Ye, B., Wang, Y., Chen, H., Cheng, J., & Zhu, X. (2014). New Word Detection for Sentiment Analysis, 531–541.

Hamling, T., & Agrawal, A. (2017). Sentiment Analysis of Tweets to Gain Insights into the 2016 US Election. Columbia Undergraduate Science Journal, 11, 34–42.

Ibope (2018). Boca de Urna aponta Jair Bolsonaro e Fernando Haddad no 2º turno. Disponível em: <<http://datafolha.folha.uol.com.br/eleicoes/2018/10/1983037-bolsonaro-tem-40-dos-validos-na-vespera-do-1-turno.shtml>>.

Acesso em 08 nov. 2018.

John Wilbur, W., & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of Information Science*, 18(1), 45–55.

Disponível em: <<https://doi.org/10.1177/016555159201800106>>

Jurafsky, D., & Martin, J. H. (2018). Speech and Language Processing.

Disponível em: <<https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>>

K. (2017) A gentle introduction to support vector machines using R.

Disponível em: <<https://eight2late.wordpress.com/2017/02/07/a-gentle-introduction-to-support-vector-machines-using-r/>>

Karami, A., Bennett, L. S., & He, X. (2018). Mining Public Opinion about Economic Issues: Twitter and the U.S. Presidential Election. *International Journal of Strategic Decision Sciences (IJSDS)*, 9(1), 18-28.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. AACL-2011 Tutorial.

Disponível em: <<https://doi.org/10.2200/S00416ED1V01Y201204HLT016>>

Lui, C., Metaxas, P. T., & Mustafaraj, E. (2011). On the Predictability of the U.S. Elections Through Search Volume Activity. Proceedings of the IADIS International Conference on E-Society, c, 1–9.

Disponível em: <<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.220.7184>>

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications : A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113.

Disponível em: <<https://doi.org/10.1016/j.asej.2014.04.011>>

New York Times. (2016). For Election Day Influence, Twitter Ruled Social Media. Disponível em: <<https://www.nytimes.com/2016/11/09/technology/for-election-day-chatter-twitter-ruled-social-media.html>>.

Acesso em 11 nov. de 2018.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From Tweets to Polls : Linking Text Sentiment to Public Opinion Time Series. Proceedings of the International AAAI Conference on Weblogs and Social Media.

Powers, D. M. W. (2007). Evaluation : From Precision , Recall and F-Factor to ROC , Informedness , Markedness & Correlation.

Prasetyo, N. D. (2014). Tweet-Based Election Prediction, (December), 67.

Schiavoni, A. S. (2010). UM ESTUDO COMPARATIVO DE MÉTODOS PARA BALANCEAMENTO DO CONJUNTO DE TREINAMENTO EM APRENDIZADO DE REDES NEURAIS ARTIFICIAIS.

Speriosu, M., Sudan, N., Upadhyay, S., & Baldridge, J. (2011). Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph, 53–63.

Statista. (2018). Most famous social network sites worldwide as of October 2018, ranked by number of active users (in millions).

Disponível em: <<https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>>.

Acesso em 08 nov. de 2018.

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 178–185.

Twitter Blog. (2016). How #Election2016 was Tweeted so far.

Disponível em: <https://blog.twitter.com/official/en_us/a/2016/how-election2016-was-tweeted-so-far.html>.

Acesso em 11 nov. de 2018.

Twitter Blog. (2018). Twitter e as #Eleições2018 no Brasil.

Disponível em: <https://blog.twitter.com/official/pt_br/topics/company/2018/twitter-e-as-eleicoes-2018-no-brasil.html>.

Acesso em 11 nov. de 2018.

USA TODAY. (2016). Forget Trump: Election's big winner was Twitter. Disponível em:

<<https://www.usatoday.com/story/tech/news/2016/11/08/election-winner-twitter/93509896/>> . Acesso em 11 nov. de 2018.

VEJA. (2018). Bolsonaro leva facada em atentado durante campanha em Juiz de Fora.

Disponível em: <<https://veja.abril.com.br/politica/bolsonaro-leva-facada-em-atentado-durante-campanha-em-juiz-de-fora/>>.

Acesso em 18 out. de 2018.

VEJA. (2018). Quem são os treze candidatos à Presidência da República em 2018.

Disponível em: <<https://veja.abril.com.br/politica/quem-sao-os-13-candidatos-a-presidencia-da-republica-em-2018/>> . Acesso em 08 out. de 2018.

VEJA. (2018). Em eleição polarizada, país vai às urnas para escolher seu 38o presidente.

Disponível em: <<https://veja.abril.com.br/politica/em-eleicao-polarizada-pais-vai-as-urnas-para-escolher-seu-38o-presidente/>>.

Acesso em 11 nov. de 2018.

Wikipédia (2010). Neural Network.

Disponível em: <https://commons.wikimedia.org/wiki/File:Neural_network.svg?uselang=ru>

Wilson, T., Hoffmann, P., & Wiebe, J. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, P 347–354.

Yano, T., & Smith, N. A. (2010). What's Worthy of Comment ? Content and Comment Volume in Political Blogs.

Yaqub, U., Chun, S. A., Atluri, V., & Vaidya, J. (2017). Analysis of political discourse on twitter in the context of the 2016 US presidential elections. Government Information Quarterly, 34(4), 613–626.

Disponível em: <<https://doi.org/10.1016/j.giq.2017.11.001>>