Universidade Federal do ABC
Programa de Iniciação Científica - UFABC

Análise de Sentimento no Twitter nas Eleições de 2018

Iniciação Científica - Modalidade bolsista

Edital N° 01/2018

Aluno: Bruno Sanches Rodrigues
RA: 11201721076
Bacharelado em Ciência & Tecnologia
rodrigues.bruno@aluno.ufabc.edu.br

Orientador: José Artur Quilici-Gonzalez
Centro de Matemática, Computação e Cognição
jose.gonzalez@ufabc.edu.br

Este projeto conta com o auxílio financeiro do CNPq (Bolsa de IC)

Santo André, 30 de Agosto de 2019

Resumo

A Análise de Sentimento é um campo de estudo no qual técnicas de aprendizado de máquina e processamento de linguagem natural são aplicadas em dados textuais para extrair informação subjetiva contida no texto, com o objetivo de classificar a opinião do autor sobre o assunto, pessoa ou entidade em questão. Esse tipo de estudo analítico pode trazer informações sobre a opinião de um vasto público, quando aplicado em redes sociais. O Twitter tem mecanismos que facilitam a coleta de dados, e por isso é uma plataforma onde esse tipo de análise é muito comum. Técnicas similares têm sido utilizadas para a predição de resultados de eleições em trabalhos e pesquisas acadêmicas, e algumas delas serão detalhadas neste trabalho. Entretanto, recentemente verificou-se que há um crescimento no número de perfis falsos no Twitter com objetivo de favorecer certos candidatos ou partidos. Com isso em mente, este projeto propõe um estudo comparativo entre Análise de Sentimento aplicada ao Twitter e pesquisas eleitorais de fontes reconhecidas. Assim, utilizando a linguagem de programação R, foram coletados mais de 4 milhões de tweets durante os dois últimos meses das eleições presidenciais de 2018, e foram reunidos 50 relatórios de pesquisa de opinião política de 4 fontes diferentes. Taxas como aprovação, rejeição e intenção de voto foram comparadas ao longo do tempo com resultados da Análise de Sentimento de tweets dos principais candidatos às eleições presidenciais de 2018. O estudo comparativo procurou estabelecer eventuais correlações e validação entre as tendências dessas duas fontes. A Análise de Sentimento dos tweets foi feita com base no método léxico, utilizando também a remoção de *stopwords* e aplicação de *stemming* para o português. Os resultados não demonstraram forte correlação com as pesquisas eleitorais, e tampouco verificou-se uma clara preferência dos usuários do Twitter por um dos candidatos.

Palavras-chave: Twitter, Análise de Sentimento, Eleições, Linguagem R.

Sumário

1 Introdução

A Análise de Sentimento é um campo de estudo no qual tenta-se extrair opiniões, emoções, sentimentos ou posicionamentos sobre alguma entidade, pessoa, marca, etc. Seu principal objetivo é determinar se a opinião de certa audiência é positiva ou negativa em relação a determinado assunto. Esse conhecimento pode ser útil, por exemplo, para empresas acompanharem a recepção a seus produtos por meio da análise de avaliações na Internet. Essa tarefa envolve a classificação de grandes quantidades de texto disponível na web, e por isso a Análise de Sentimento está diretamente relacionada a mineração de texto, processamento de linguagem natural e técnicas de aprendizado de máquinas [1]. Trata-se de um campo com muitas aplicações, como na análise automática de pesquisas de opinião, melhoria em sistemas de recomendação e processamento de mensagens e e-mails [2].

Uma aplicação importante da Análise de Sentimento é a extração de sentimentos em redes sociais, o que é especialmente conveniente visto que estas são importantes meios de circulação de opiniões. O Twitter é uma rede social que funciona como microblog, cujo foco é o compartilhamento em tempo real de curtas mensagens de texto [2]. A plataforma é um importante veículo de informação, notícias e opiniões, especialmente em eventos de escala nacional, como olimpíadas e eleições [3]. A análise de dados no Twitter é bastante atrativa em razão da facilidade de se recolher quantidades moderadas de tweets por meio de sua interface de programação de aplicações (API) [4]. Dentre os diferentes softwares e métodos que podem ser utilizados para esse propósito, a linguagem de programação R possui dois pacotes para trabalhar com tweets, e uma série de ferramentas interessantes para a análise, avaliação e representação desses dados [5, 6]. Assim, é possível recolher em tempo real milhares de tweets relacionados a certa palavra-chave, e, desses, extrair o sentimento médio associado, seja ele negativo, positivo, neutro ou ambíguo [3].

Esse conhecimento também pode ser usado no espectro político. Diversos trabalhos têm sido realizados utilizando a Análise de Sentimento na predição dos resultados de eleições [7, 8, 9]. De fato, eleitores cada vez mais expressam sua aprovação ou crítica aos candidatos por meio do Twitter, especialmente durante debates ou nas vésperas das eleições, gerando uma grande quantidade de dados.

Entretanto, há de se notar que, como recentemente noticiado [10, 11], existe um crescente número de perfis falsos com o objetivo de influenciar o pensamento da população em benefício de algum político em particular. Políticos com suficiente influência podem agora se beneficiar de serviços ilegais que consistem em numerosos perfis falsos, parcial ou completamente automatizados, que navegam no Twitter, postando tweets a favor ou contra algum político em questão. Esse fenômeno que tem se intensificado Brasil afora nos últimos anos pode vir a alterar o resultado de análises políticas feitas

com base no Twitter. É relevante notar também que há uma polarização entre os ideais políticos que é acentuado pelas mídias sociais como o Twitter [12].

Diante dessa realidade, a proposta deste projeto é fazer uma análise comparativa entre o sentimento de tweets coletados durante Setembro, mês anterior ao primeiro turno das eleições presidenciais de 2018, e os resultados de pesquisas eleitorais de fontes confiáveis como Ibope, Datafolha e Vox Populi. Essa análise será feita ao longo do tempo, de forma que as tendências de ambas as fontes possam ser comparadas, e assim, possam ser estabelecidas eventuais correlações entre elas. Considerando o cenário político fortemente polarizado desta eleição presidencial de 2018, e um número não desprezível de perfis falsos na rede social, ou mantidos por "influenciadores" obscuramente remunerados, este projeto se diferenciaria das fontes citadas pela eventual comprovação de que talvez o Twitter esteja perdendo sua capacidade de representar a opinião do eleitorado.

2 Objetivos

O objetivo do projeto é realizar um estudo comparativo entre técnicas de Análise de Sentimento aplicada em tweets relacionados às eleições presidenciais do Brasil/2018 e fontes jornalísticas, procurando estabelecer eventuais correlações e validação entre as tendências dessas duas fontes. Para isso, foram delimitados os seguintes objetivos específicos:

- Estudo prático visando o entendimento e manejo das diferentes formas de coleta, pré-processamento e análise automática de sentimento em tweets;
- Coleta de tweets relacionados aos políticos selecionados ao longo do mês anterior à primeira fase das eleições presidenciais;
- Pré-processamento dos tweets coletados para a formação de uma base de dados estruturada;
- Implementação de técnicas de análise automática de sentimento para extrair o sentimento médio ao longo do tempo em relação a cada político escolhido como de interesse para este trabalho;
- Adaptação do formato dos resultados da análise com os dados de pesquisas eleitorais;
- Comparação, análise e validação dos resultados.

3 Cronograma

De acordo com o cronograma proposto no projeto inicial, mostrado na Figura 1, a primeira fase de coleta de dados foi completada com êxito, e um número considerável de tweets foi coletado a cada dia. Assim, até o a entrega do relatório parcial, os resultados consistiam nos tweets coletados e alguns dados obtidos a partir destes.

Já na segunda etapa, foi realizada a Análise de Sentimento, representação dos resultados em si e interpretação dos mesmos.

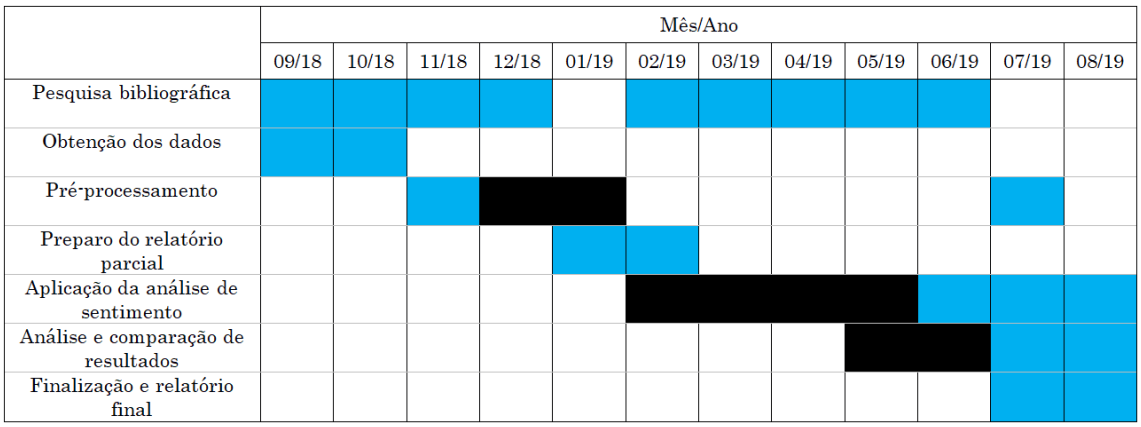


Figura 1: *Cronograma das atividades (planejamento em preto e realização em azul).*

4 Revisão Bibliográfica

Neste capítulo é feito um embasamento sobre os objetos de estudo e o processo de Análise de Sentimento.

4.1 Twitter

O Twitter é uma rede social no formato de microblog. Trata-se de uma plataforma de aspecto público e permite que opiniões sejam expressadas rapidamente através de mensagens curtas de no máximo 280 caracteres.

Sua função *retweet* permite que seja compartilhado um tweet de outro usuário. A função permite ainda que seja adicionado um comentário ao tweet original. É possível também mencionar outros usuários através do uso de '@', incluindo tal usuário na discussão.

Além do texto, tweets podem conter variados tipos de informações como imagens, links para outros websites e localização. Tais dados podem ser extraídos e utilizados de forma a melhorar o entendimento

dos processos envolvendo o Twitter [13].

4.2 Análise de Sentimento

A Análise de Sentimento é uma técnica bastante atual e muito relevante tendo em vista a quantidade de dados textuais desestruturada que é gerada atualmente. Essa análise automática reduz o tempo e custo necessários para a obtenção de resultados, possibilitando até uma análise em tempo real da opinião pública em relação a determinado assunto.

Trata-se de uma técnica com diversas áreas de aplicação. Monitoramento de mídias sociais, monitoramento de marcas em relação ao produto ou prestação de serviço, pesquisa de mercado, etc.

O resultado de uma Análise de Sentimento aplicada sobre um texto (frase, mensagem ou documento maior) pode ter a forma de uma polaridade (positivo, negativo ou neutro), expressar um sentimento básico (raiva, felicidade, nojo), ou até mesmo identificar intenção (interessado, não interessado).

O idioma do texto é importante para a Análise de Sentimento, visto que as estruturas de cada idioma e as palavras e expressões são completamente diferentes [1, 14].

4.2.1 Análise Léxica

O método léxico consiste na análise de tweets com base em um léxico, que relaciona cada palavra contida no mesmo com um sentimento. Nesta técnica, cada palavra do documento é comparada com o léxico, e o sentimento resultante é atribuído ao documento. Esse tipo de análise difícilmente leva em consideração a estrutura do texto, e seu desempenho está diretamente relacionado à qualidade do léxico.

O sentimento associado às palavras pode ser binário (+1 para palavras positivas e -1 para negativas), o que simplifica o problema de classificação, à custa de seu desempenho. Podem também ser atribuidos graus para a positividade ou negatividade das palavras, como por exemplo: *ruim* -2 e *péssimo* -4.

Tal método de Análise de Sentimento pode enfrentar dificuldades com o passar do tempo, pois na medida que novas palavras são adicionadas ao vocabulário, ou novas gírias são utilizadas, o léxico pode ficar desatualizado. Além disso, cada assunto de documento tratado pode requerir léxicos específicos, observando que diferentes palavras podem ter conotações diferentes em cada contexto. A disponibilidade de léxicos em idiomas como o Português é limitada [1].

4.2.2 Análise Supervisionada

A Análise de Sentimento feita através de algoritmos supervisionados consiste na utilização de um conjunto de dados previamente rotulado, para que o modelo possa ser treinado. Uma vez treinado, o modelo pode ser aplicado ao restante dos dados, e também avalidado em um conjunto de teste previamente rotulado.

A Análise de Sentimento consiste em um problema de classificação, no qual as classes são os sentimentos em relação ao documento, e a entrada é alguma característica do documento. A entrada mais simples consiste no uso das próprias palavras e suas frequências para a atribuição do sentimento, porém a ordem e relação entre as palavras pode ser o atributo também. Assim, algoritmos de *Machine Learning* como o *Naive Bayes*, *Support Vector Machines* e *Random Forests* têm sido utilizados [1, 14].

4.3 Trabalhos relacionados

Nesta seção serão apresentados trabalhos acadêmicos relevantes.

Zimbra et al. [14] realizaram uma análise dos 28 sistemas de Análise de Sentimento mais bem sucedidos, tanto na área acadêmica quanto comercial. Essa análise revela as tendências do campo de estudo, detalhando os desafios, abordagens e técnicas atuais de Análise de Sentimento no Twitter. Ainda, cada sistema foi avaliado utilizando um conjunto rotulado de tweets segmentado em diversos assuntos. Os resultados dessa análise mostram que mesmo sistemas do estado da arte ainda exibem acurácias relativamente baixas, com uma média de 61%.

O *Unilex* [15] é um método de Análise de Sentimento léxico para textos em português. Esse trabalho descreve a criação de um dicionário com base em um conjunto rotulado de tweets políticos. Por fim, o método apresentado foi comparado ao *ifeel2.0* [16], um sistema web de Análise de Sentimento em documentos que reúne diversas técnicas atuais. Os resultados mostraram que a acurácia do *Unilex* superou a acurácia obtida em todas as técnicas presentes no *ifeel2.0*.

5 Materiais e Métodos

A seguir serão detalhados os procedimentos realizados em cada etapa do projeto.

5.1 Obtenção dos Dados

No período de 01 de Setembro de 2018 até 10 de Novembro de 2018, realizou-se a coleta de dados no Twitter, possibilitada através da API do Twitter [17], em conjunto com a biblioteca da linguagem R *rtweet* [5].

Inicialmente, foi feita uma rápida pesquisa preliminar para determinar quais e quantos candidatos seriam acompanhados. Uma pesquisa de popularidade simples da época [18] permitiu que fossem escolhidos os seguintes candidatos: Jair Bolsonaro, Ciro Gomes, Marina Silva, Fernando Haddad e Geraldo Alckmin.

Para se trabalhar com a API do Twitter, era necessário utilizar uma conta do Twitter para obter as credenciais necessárias para obter acesso aos tweets. Primeiramente, foi criado um aplicativo no Twitter para obter as chaves de acesso para autenticação. Uma vez com as chaves de acesso gravadas em variáveis no ambiente de programação R, não era mais necessária qualquer interação grande com o browser.

Há mais de uma biblioteca para trabalhar com a API do Twitter no R, sendo as mais populares *rtweet* e *twitteR* [5, 6]. A biblioteca escolhida para este projeto foi *rtweet* por ser mais recente, funcionar melhor com as mudanças na forma de autenticação exigidas pela API do Twitter, e pela sua flexibilidade em suas funções principais.

Através de alguns exemplos disponíveis na documentação da biblioteca, foi possível escrever um algoritmo que coletava tweets passados utilizando a função de busca de tweets. A função permitia que fosse feita uma busca por tweets contendo um conjunto de palavras, podendo determinar um intervalo de tempo no passado em que a busca era feita, e ainda um número esperado de tweets retornados. Uma captura de tela do algoritmo em funcionamento é mostrada na Figura 2. Porém, a versão grátis da API do Twitter sendo utilizada limitava a pesquisa em no máximo 10 dias no passado, e um número máximo de 18,000 tweets a cada 15 minutos.

A função de busca retorna tweets em uma base de dados, como pode ser visto na Figura 3, contendo 88 colunas de informações, incluindo texto do tweet, data de criação, dados do usuário, dados do *retweet*, geolocalização quando presente, etc.

Começando no dia 08 de Setembro de 2019 até o fim da primeira fase das eleições, no dia 07 de Outubro de 2019, o algoritmo foi executado manualmente uma vez por dia, fazendo buscas em separado para cada um dos 5 candidatos por vez. Os tweets retornados foram então gravados em arquivos no formato CSV. A partir do dia 22 de Setembro de 2019, passou-se a ser feita também a

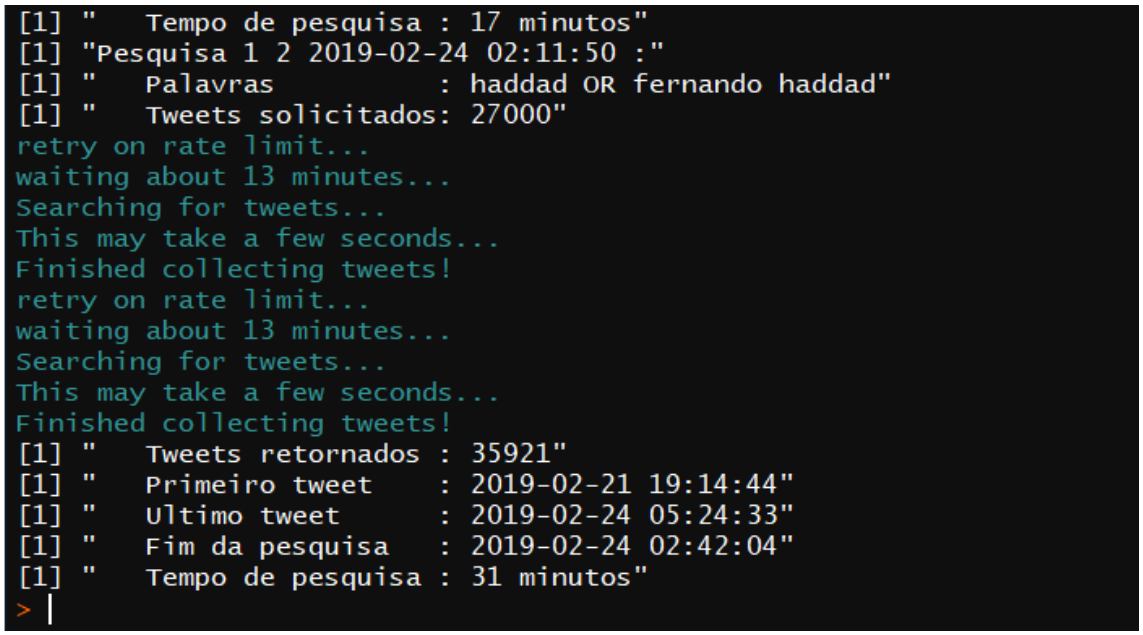
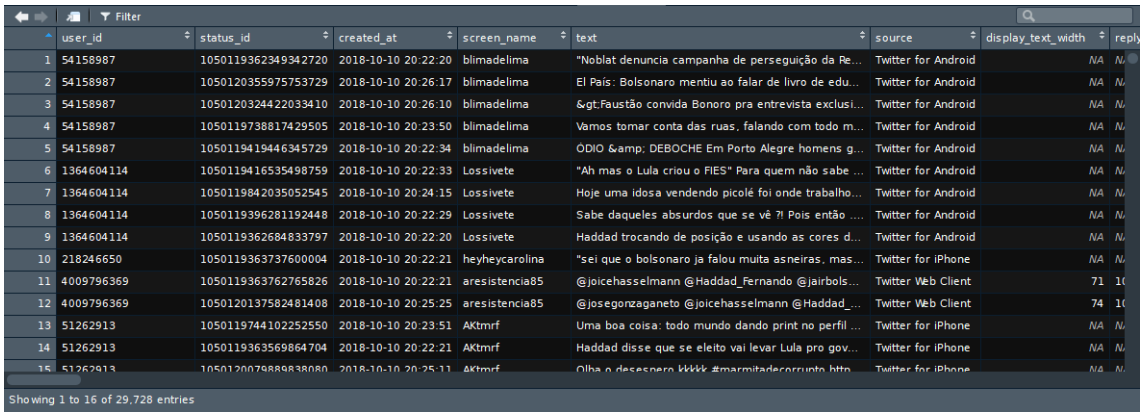


Figura 2: *Captura de um exemplo de execução (tela) do algoritmo principal de coleta.*



Filter

	user_id	status_id	created_at	screen_name	text	source	display_text_width	reply
1	54158987	1050119362349342720	2018-10-10 20:22:20	blimadelima	"Noblat denuncia campanha de perseguição da Re...	Twitter for Android	NA	N
2	54158987	1050120355975753729	2018-10-10 20:26:17	blimadelima	El País: Bolsonaro mentiu ao falar de livro de edu...	Twitter for Android	NA	N
3	54158987	1050120324422033410	2018-10-10 20:26:10	blimadelima	>Faustão convida Bonoro pra entrevista exclusi...	Twitter for Android	NA	N
4	54158987	1050119738817429505	2018-10-10 20:23:50	blimadelima	Vamos tomar conta das ruas, falando com todo m...	Twitter for Android	NA	N
5	54158987	1050119419446345729	2018-10-10 20:22:34	blimadelima	ÓDIO & DEBOCHE Em Porto Alegre homens g...	Twitter for Android	NA	N
6	1364604114	1050119416535498759	2018-10-10 20:22:33	Lossivete	"Ah mas o Lula criou o FIES" Para quem não sabe ...	Twitter for Android	NA	N
7	1364604114	1050119842035052545	2018-10-10 20:24:15	Lossivete	Hoje uma idosa vendendo picolé foi onde trabalho...	Twitter for Android	NA	N
8	1364604114	1050119396281192448	2018-10-10 20:22:29	Lossivete	Sabe daqueles absurdos que se vê ?! Pois então	Twitter for Android	NA	N
9	1364604114	1050119362684833797	2018-10-10 20:22:20	Lossivete	Haddad trocando de posição e usando as cores d...	Twitter for Android	NA	N
10	218246650	1050119363737600004	2018-10-10 20:22:21	heyheycarolina	"sei que o bolsonaro ja falou muita asneiras, mas...	Twitter for iPhone	NA	N
11	4009796369	1050119363762765826	2018-10-10 20:22:21	aresistencia85	@joicehasselmann @Haddad_Fernando @jairbols...	Twitter Web Client	71	1
12	4009796369	1050120137582481408	2018-10-10 20:25:25	aresistencia85	@josegonzaganeto @joicehasselmann @Haddad_...	Twitter Web Client	74	1
13	51262913	1050119744102252550	2018-10-10 20:23:51	AKtmrf	Uma boa coisa: todo mundo dando print no perfil ...	Twitter for iPhone	NA	N
14	51262913	1050119363569864704	2018-10-10 20:22:21	AKtmrf	Haddad disse que se eleito vai levar Lula pro gov...	Twitter for iPhone	NA	N
15	51262913	1050120079889838080	2018-10-10 20:25:11	AKtmrf	Olha o desespero kkkkk #marmitadecorrupto http...	Twitter for iPhone	NA	N

Showing 1 to 16 of 29,728 entries

Figura 3: *Captura de um exemplo de tabela de tweets exibido no R.*

pesquisa pelo nome de todos os candidatos em conjunto, com o propósito de analisar a frequência relativa do número de tweets de cada candidato.

Do dia 08 de Outubro de 2018 até uma semana após o fim da segunda fase, no dia 04 de Novembro de 2018, o algoritmo passou a ser executado pesquisando apenas pelos candidatos Jair Bolsonaro e Fernando Haddad. Contudo, decidiu-se utilizar outra função da biblioteca *rtweet* para a coleta de dados em conjunto com a que já vinha sendo utilizada. Esta outra função procurava os tweets sendo publicados em tempo real, sem número máximo de retornos, e executava por um período de tempo pré-determinado, gravando os tweets em arquivos do tipo JSON.

Visando otimizar o funcionamento das técnicas de Análise de Sentimento, as bases de dados coletadas foram convertidas em um formato de dados nativo ao R, o *.rds*.

A ferramenta *GetOldTweets* [19], escrita em Python, foi testada como uma maneira de coletar tweets no passado, de dias em que o número de tweets foi muito baixo, ou tenha ocorrido um problema com os dados. Foi feito um algoritmo em R para chamar as funções em Python da ferramenta, adequando os dados à forma como a biblioteca *rtweet* os mostra. Todavia, não se viu necessidade de sua utilização pois o número de tweets coletados forneceu um panorama amplo das eleições.

Durante todo o período eleitoral, pesquisas eleitorais de diversas fontes foram reunidas para a comparação com a Análise de Sentimento. Foram reunidas, ao todo, 42 relatórios de pesquisas de intenção de voto das seguintes fontes: Ibope, Datafolha e XP/Ipespe [20, 21, 22]. Para que fossem representados da melhor maneira possível, os gráficos de resultados das pesquisas de opinião Ibope, Datafolha e XP/Ipespe foram copiados e reproduzidos no R.

5.2 Pré-processamento e Análise de Sentimento

O pré-processamento consiste na remoção de ruído de documentos de forma a facilitar o processamento dos documentos pelo algoritmo aplicado posteriormente. Dados do tweet como fonte, número de seguidores, número de *retweets* e geolocalização são descartados para que o foco seja a Análise de Sentimento em texto.

Tome como exemplo um tweet coletado no dia 2018-11-04:

```
"@jornalextra Parece um vicio q sempre tem q jogar política no meio de um assunto.
\"Vai chover hoje\" CULPA DO PT ISSO, \"meu café veio sem açúcar\" bem feito aposto
que votou no Bolsonaro. Gente ?????"
```

A primeira etapa é a remoção de ruído do texto como *hashtags*, links e caracteres especiais:

```
"jornalextra parece um vicio q sempre tem q jogar política no meio de um assunto
vai chover hoje culpa do pt isso meu café veio sem açúcar bem feito aposto
que votou no bolsonaro gente"
```

Depois, é feita a remoção de *stopwords*, palavras que funcionam como conectivos, sem relevância para a Análise de Sentimento. O conjunto de *stopwords* utilizado foi formado pela união de diversos conjuntos disponíveis online ou originários dos próprios pacotes de processamento de linguagem natural do R.

```
"jornalextra parece vicio q q jogar política meio assunto vai chover hoje culpa
pt café veio açúcar bem aposto votou gente"
```

Finalmente, as palavras flexionadas ou derivadas são reduzidas para suas formas mais básicas. Essa técnica é chamada de *stemming*, e o algoritmo de *stemming* para português utilizado é originário da linguagem de processamento de texto *Snowball* [23].

```
"jornalextr parec vici q q jog polít mei assunt vai chov hoj culp
pt caf vei açúc bem apost vot gent"
```

Uma vez com o texto tratado, parte-se para a Análise de Sentimento. O método utilizado foi a Análise de Sentimento léxica, e o algoritmo criado foi feito com base em um tutorial exibido por um blog [24]. A técnica utilizada não envolve aprendizado de máquina, e consiste em transformar cada termo do tweet tratado em um item de uma lista. Essa lista será então checada com o léxico selecionado para o trabalho. O léxico utilizado consiste em 2554 palavras rotuladas como negativas e 1399 palavras rotuladas como positivas [25]. A seguir, um trecho do mesmo:

```
            palavra  valor  polaridade
1            rachar   -1    negativo
2          punitivo   -1    negativo
3             pecar   -1    negativo
4         insalubre   -1    negativo
5         tolerável    1    positivo
6    insubstituível    1    positivo
7         espumante    1    positivo
8       espirituoso    1    positivo
```

Através do léxico, é atribuída uma polaridade (positivo ou negativo) para cada termo que aparece no tweet. As polaridades para cada candidato por cada dia são então somadas para obter os gráficos da seção de Resultados a seguir.

6 Resultados

6.1 Tweets Coletados

Com as bases de dados contendo todos os candidatos, um dos primeiros resultados foi um gráfico obtido com o R, como pode ser visto na Figura 5. A frequência relativa indicada é calculada dividindo o número de tweets cujo texto contém o nome do candidato sobre o número total. Alguns tweets fazem menção à mais de um candidato, e por isso, a soma das frequências relativas pode resultar em mais que 100%.

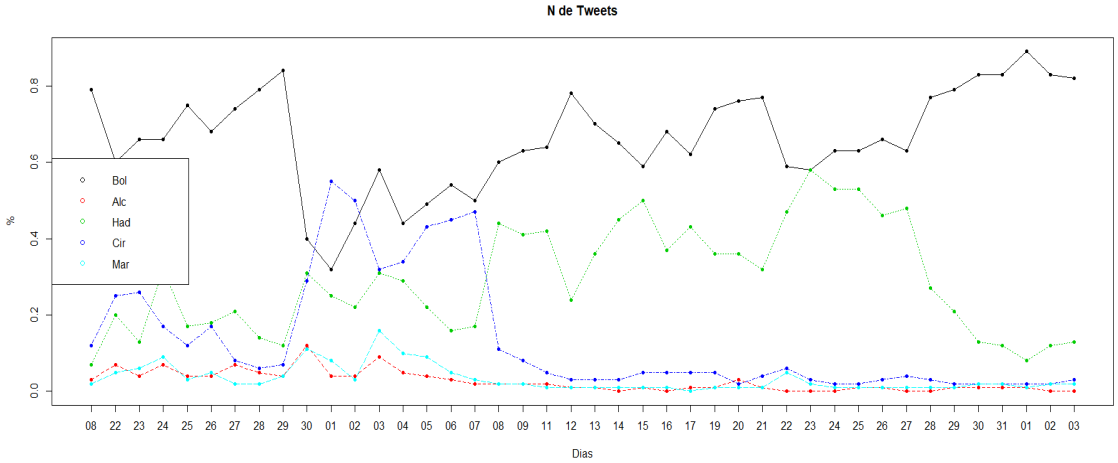


Figura 4: *Gráfico de frequência relativa dos candidatos.*

Além disso, após o fim do período de coleta, foi feito mais um algoritmo para que o número de tweets total fosse contado, excluindo-se os repetidos ou os que não continham texto algum. Gerou-se então um gráfico da quantidade total de tweets por dia coletado, como mostra a Figura 5. Ao todo, foram coletados em torno de 4,000,000 tweets divididos em 830 arquivos. Estes foram posteriormente unidos, resultando em 15 arquivos *.rds*.

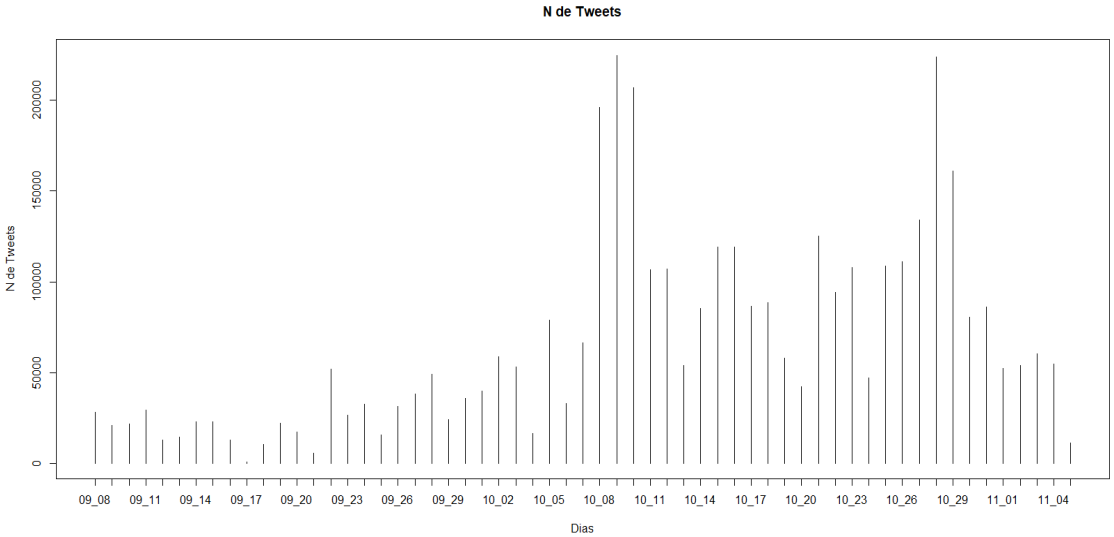


Figura 5: *Gráfico do número total de tweets coletados por dia.*

6.2 Dados Contextuais

Para a contextualização dos resultados da Análise de Sentimento, e para explicar possíveis flutuações na quantidade de tweets e popularidade dos candidatos, foi elaborada uma linha do tempo de acontecimentos relevantes envolvendo os candidatos [26, 27]. Nesta, estão situados também os dias em que os diferentes algoritmos desenvolvidos começaram a operar. As datas a seguir têm formato mês-dia, sendo referentes ao ano de 2018:

- **08-31 Sexta-feira** - O Tribunal Superior Eleitoral (TSE) negou o registro da candidatura do ex-presidente Luiz Inácio Lula da Silva (PT). Início da propaganda eleitoral gratuita no rádio e televisão do primeiro turno.
- **09-06 Quinta-feira** - Jair Bolsonaro (PSL) sofreu um golpe de faca quando participava de uma caminhada pelas ruas de Juiz de Fora (MG).
- **09-08 Sábado** - Início da coleta de tweets pesquisando por cada candidato em separado.
- **09-09 Domingo** - Debate TV Gazeta e Estadão.
- **09-11 Terça-feira** - A Executiva Nacional do PT confirmou o nome de Fernando Haddad como o candidato do partido à Presidência da República. Lula cumpre pena de 12 anos e 1 mês.
- **09-18 Terça-feira** - Debate Poder 360/Youtube/Piauí.
- **09-20 Quinta-feira** - Debate TV Aparecida.
- **09-22 Sábado** - Início da pesquisa por todos os candidatos em conjunto. Candidatos não podem mais ser presos.
- **09-26 Quarta-feira** - Debate SBT, Folha e UOL
- **09-29 Sábado** - Bolsonaro recebe alta após 23 dias internado. No mesmo dia, o movimento #elenão ganhou as ruas do País.
- **09-30 Domingo** - Debate Record.
- **10-04 Quinta-feira** - Debate Globo, Fim da propaganda eleitoral gratuita do primeiro turno.
- **10-06 Sábado** - Término das últimas pesquisas de intenção de voto antes do primeiro turno.
- **10-07 Domingo** - Votação da primeira fase. Bolsonaro e Haddad disputam a segunda fase. Bolsonaro (PSL) - 46,03%; Haddad (PT) - 29,28%; Ciro Gomes (PDT) - 12,47%; Geraldo Alckmin (PSDB) - 4,76%; João Amoêdo - 2,50%; Marina Silva (Rede) - 1%.
- **10-08 Segunda-feira** - Início da coleta em tempo real e mudança da pesquisa para apenas os candidatos concorrendo à segunda fase. Haddad visitou Lula em sua cela na sede da Polícia Federal em Curitiba pela última vez.
- **10-11 Quinta-feira** - Debate Band.

- **10-12 Sexta-feira** - Início da propaganda eleitoral gratuita para o segundo turno.
- **10-14 Domingo** - Debate TV Gazeta.
- **10-15 Segunda-feira** - Debate RedeTV!
- **10-17 Quarta-feira** - Debate SBT.
- **10-21 Domingo** - Circulação de vídeo polêmico envolvendo o filho de Bolsonaro. Debate Record.
- **10-26 Sexta-feira** - Debate Globo. Fim da propaganda eleitoral gratuita do segundo turno.
- **10-27 Sábado** - Término das últimas pesquisas de intenção de voto antes do segundo turno.
- **10-28 Domingo** - Votação da segunda fase. Bolsonaro eleito presidente com 56% dos votos válidos.
- **11-05 Domingo** - Fim do período de coleta de dados.

As pesquisas de opinião [20, 21, 22] foram representadas de forma a incluir as datas mais relevantes para o projeto. As Figuras 6, 7 e 8 são referentes ao dado *"Intenção de voto estimulada para presidente 2018 (total de votos)"* obtido a partir da pergunta *"Se as eleições fossem hoje, em qual desses candidatos você votaria? (Escolha uma opção)"*.

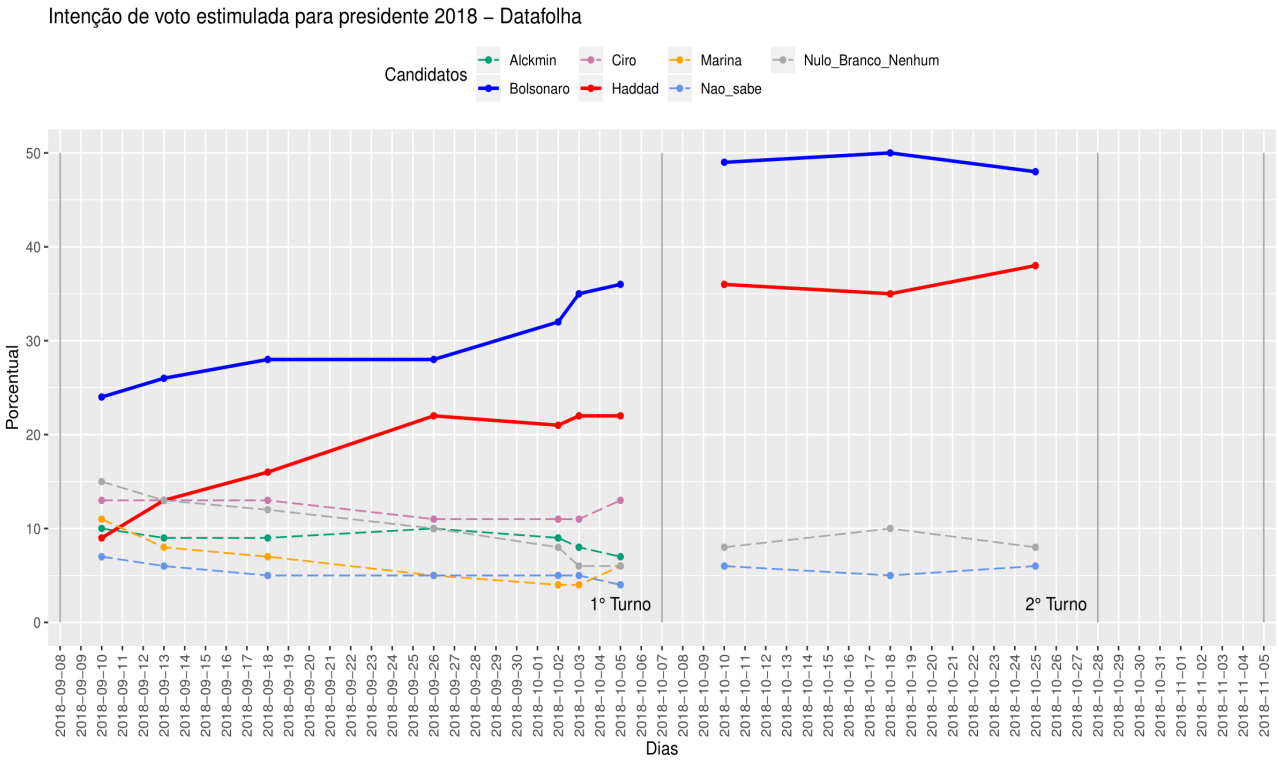


Figura 6: *Pesquisa de intenção de voto Datafolha [21].*

Em todas as pesquisas, observa-se a prevalência do candidato Jair Bolsonaro nas intenções de voto, com o candidato Fernando Haddad se aproximando nos períodos do fim do mês de Setembro e no fim do segundo turno. Há também uma diminuição nas intenções de votos não válidos em geral (brancos ou nulos).

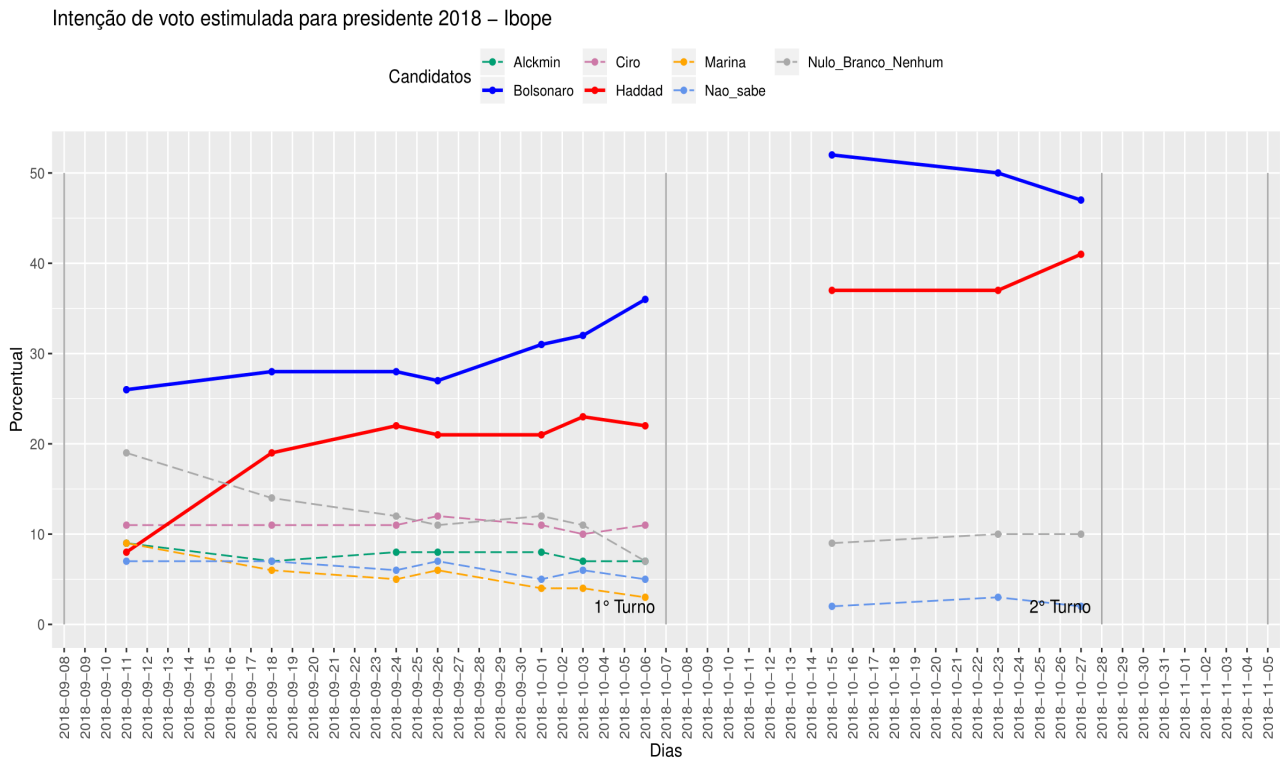


Figura 7: *Pesquisa de intenção de voto Ibope [20].*

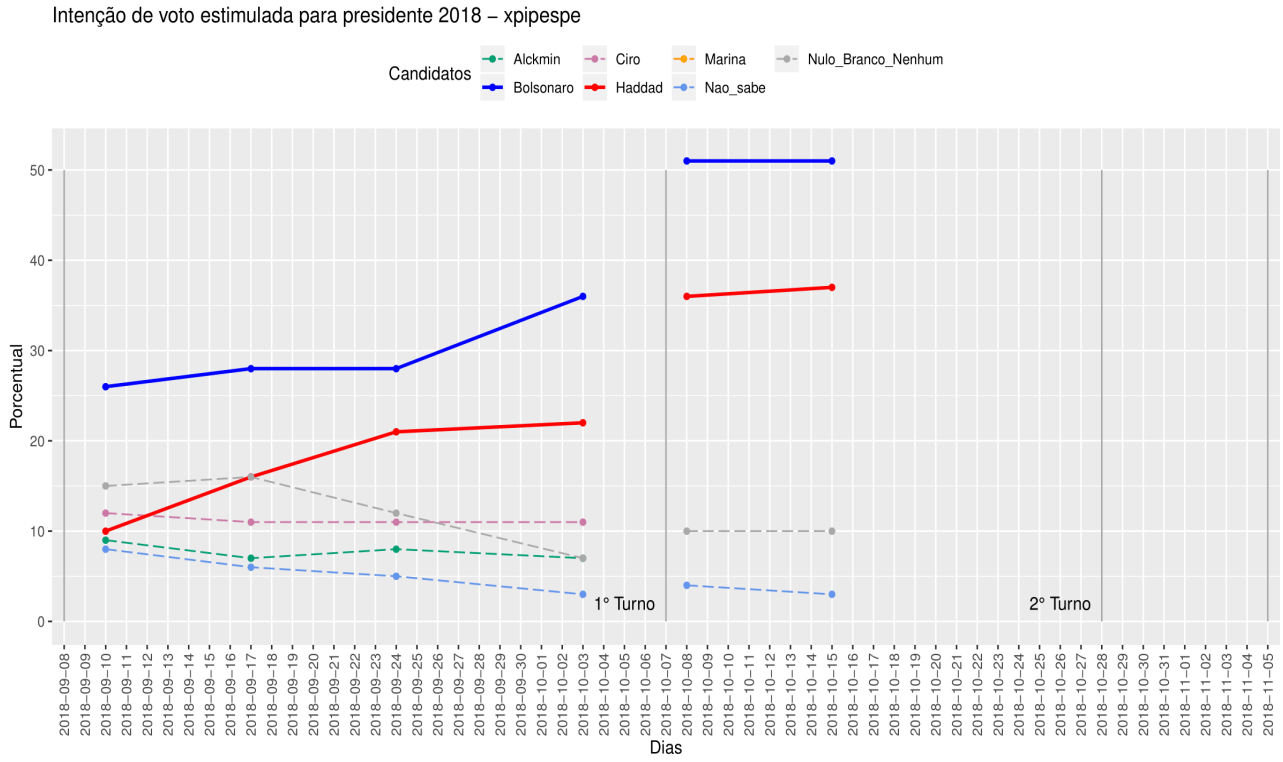


Figura 8: *Pesquisa de intenção de voto XP/Ipespe [22].*

6.3 Análise de Sentimento

O processo de Análise de Sentimento descrito no capítulo anterior foi aplicado no conjunto inteiro de tweets, do dia 2018-09-08 até o dia 2018-11-05. Os tweets foram separados por dia, e por candidato. Tweets contendo mais de um candidato em seu texto apresentavam um desafio para a classificação. Decidiu-se então que tais tweets seriam incluídos mais de uma vez na classificação.

Para os gráficos de sentimento ao longo do tempo obtidos, foi utilizado um máximo de 5000 tweets por candidato por dia, escolhidos de forma aleatória. Essa decisão teve como objetivo evitar a introdução de um desbalanceamento indesejável na análise, visto que tweets contendo os nomes Bolsonaro e Haddad compõem a maior parte das bases de dados.

A Figura 9 mostra a porcentagem de palavras com sentimento positivo em relação ao total de palavras que foram atribuídas sentimento, para cada dia e candidato. Observa-se que as porcentagens dos candidatos Jair Bolsonaro e Fernando Haddad não são díspares nem variam na mesma proporção que o observado nas pesquisas de opinião das Figuras 6 7 8. Ainda, é possível perceber que, apesar da evolução das porcentagens ser parecida para todos os candidatos, a candidata Marina Silva se destaca, acima dos demais. Contrariamente, o candidato Ciro Gomes tem porcentagens ligeiramente mais baixas, principalmente durante o primeiro turno.

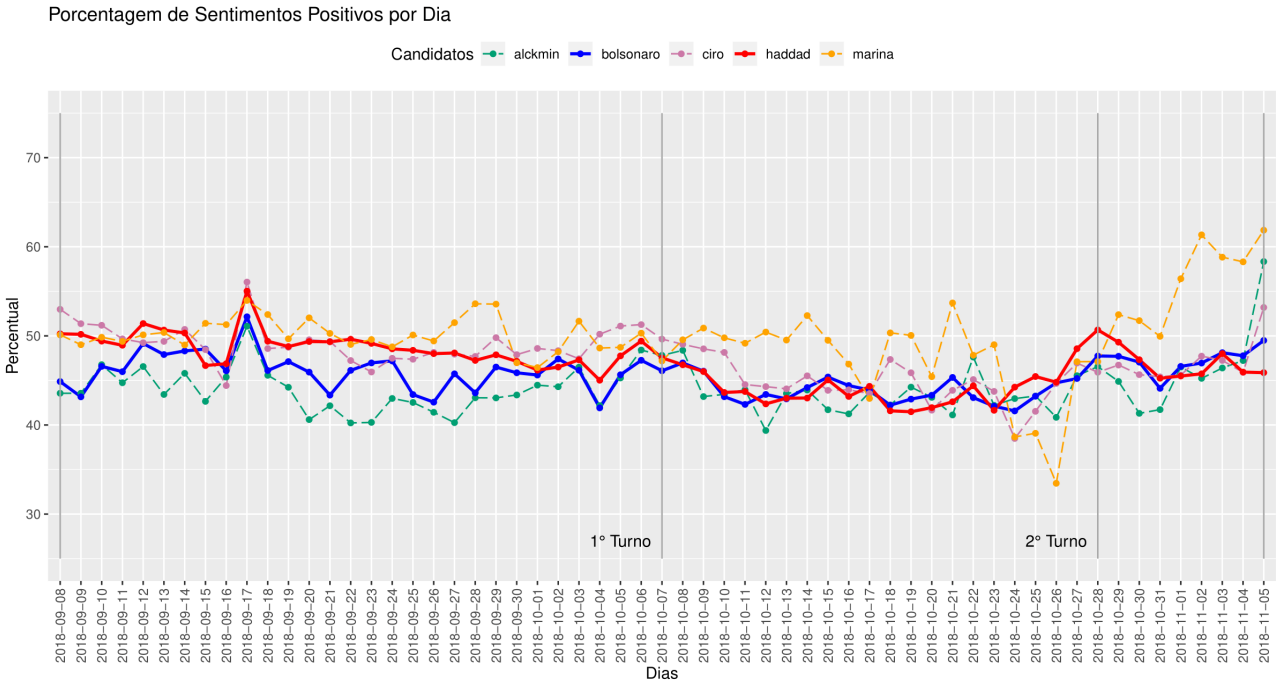


Figura 9: *Gráfico da porcentagem de sentimentos positivos por dia.*

A Figura 10 mostra o número de palavras cujo sentimento atribuído foi positivo, e, analogamente, a Figura 11 para os sentimentos negativos. Nota-se que a quantidade de sentimentos dos candidatos Jair Bolsonaro e Fernando Haddad foi a de menor variação ao longo da campanha, o que pode ser

explicado pela abundância de tweets contendo seus nomes.

Como pode ser observado na Figura 4, os dias 2018-09-17 e 2018-09-21 tiveram poucos tweets coletados. Isso levou a resultados de sentimento que diferem dos demais dias. Estes dias podem ser vistos como *outliers*, ou seja, valores inconsistentes para esta análise.

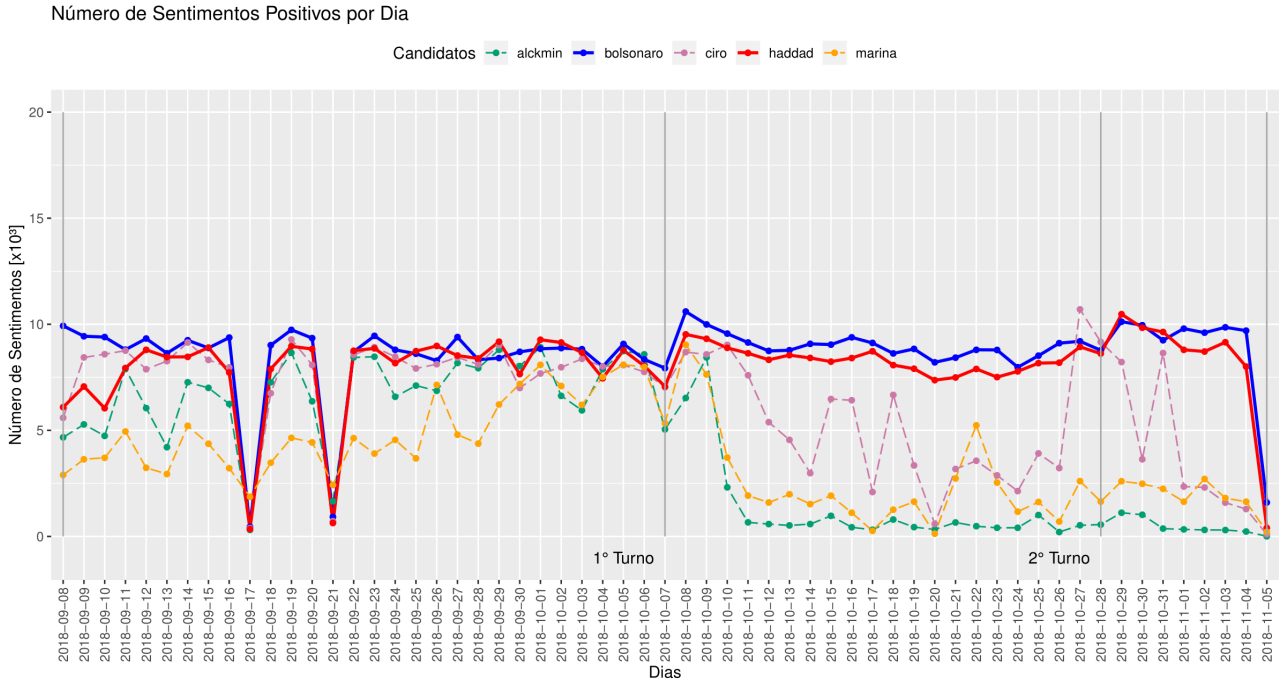


Figura 10: *Gráfico do número de sentimentos positivos por dia.*

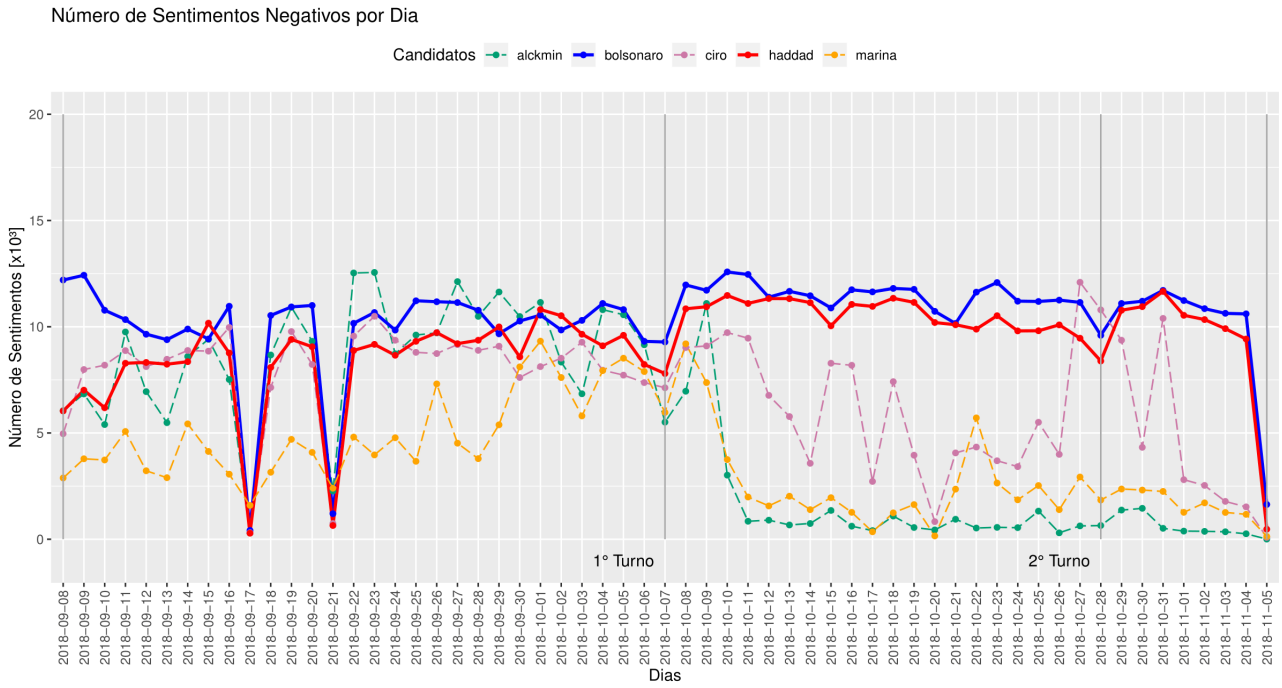


Figura 11: *Gráfico do número de sentimentos negativos por dia.*

O último gráfico, na Figura 12 mostra a soma dos sentimentos positivos decrescido dos negativos. Essa última forma de visualização do sentimento ao longo do tempo destaca as variações dos candidatos Jair Bolsonaro e Fernando Haddad.

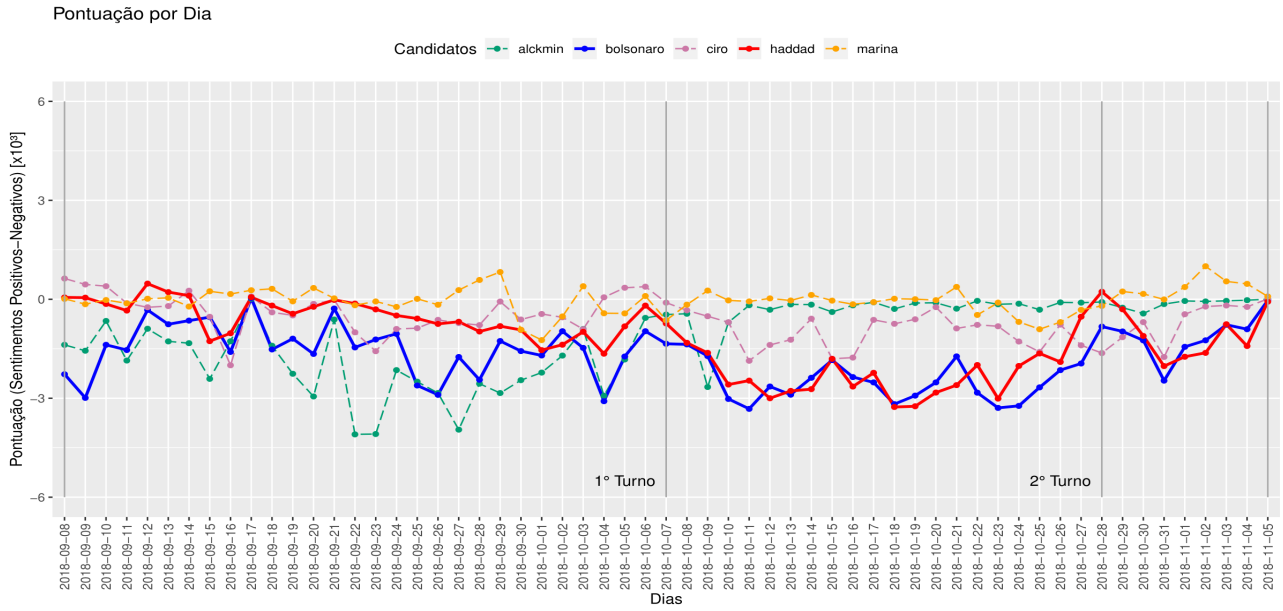


Figura 12: *Gráfico da pontuação por dia.*

6.4 Validação do Método

Para a validação do método utilizado, realizou-se a classificação manual de um conjunto de 196 tweets selecionados aleatoriamente do conjunto total de dados.

A classificação manual foi feita em relação a apenas um dos candidatos contidos no tweet, atribuindo uma das seguintes classes: positivo, negativo ou neutro. A classe neutra foi destinada a tweets: que relatavam notícias; cujo posicionamento não era claro; e que apenas faziam referência a algum link externo ao Twitter.

A Análise de Sentimento foi então aplicada individualmente, aos mesmos tweets, com a atribuição da nova classe neutra para tweets cujo número de sentimentos positivos se igualou ao de sentimentos negativos. O resultado da análise da acurácia (bem como outras métricas) com relação à classificação manual é mostrado a seguir:

```
Confusion Matrix and Statistics

          Reference
Prediction Negativo Neutro Positivo  Total
Negativo       21     21       20     62
Neutro         23     34       36     93
Positivo        9     21       11     41
Total          53     76       67
```

```
Overall Statistics

Accuracy : 0.3367
95% CI : (0.271, 0.4075)
No Information Rate : 0.3878
P-Value [Acc > NIR] : 0.93919

Kappa : -0.0065

Mcnemar's Test P-Value : 0.04185


Statistics by Class:

                     Class: Negativo Class: Neutro Class: Positivo
Sensitivity                   0.3962        0.4474         0.16418
Specificity                   0.7133        0.5083         0.76744
Pos Pred Value                0.3387        0.3656         0.26829
Neg Pred Value                0.7612        0.5922         0.63871
Prevalence                    0.2704        0.3878         0.34184
Detection Rate                0.1071        0.1735         0.05612
Detection Prevalence          0.3163        0.4745         0.20918
Balanced Accuracy             0.5548        0.4779         0.46581
```

A acurácia do modelo aplicado foi tão baixa quanto o esperado em uma classificação aleatória. Todavia, o método utilizado nessa validação difere ligeiramente do método aplicado ao conjunto inteiro dos tweets, cujos resultados foram exibidos na seção anterior. Pode-se notar também que a classe de maior número é a neutra.

7 Discussão

O processo de coleta de tweets iniciou-se exploratoriamente, sendo realizados muitos testes antes que se pudesse encontrar um conjunto de parâmetros que retornasse uma quantidade e formato de dados esperado. Mesmo com um considerável tempo de ajuste do algoritmo de coleta, posteriormente foram descobertas melhorias a serem feitas. Uma delas seria gravar todos os tweets coletados em arquivos JSON, que apresentam menos erros ao serem abertos novamente.

O tempo demandado pelos algoritmos de coleta era grande, em torno de 4 horas por dia. Todavia, o custo computacional era baixo, uma vez que, na maior parte do tempo, o algoritmo apenas esperava os 15 minutos limitantes para coletar mais tweets. Ainda, através da habilitação de acesso remoto ao computador sendo usado na coleta de dados, a localização e horário não interferiram no processo de

coleta.

O modelo de Análise de Sentimento léxica construído utilizou a remoção de *stopwords* e aplicação de *stemming* sobre um léxico vasto, porém não apresentou boa acurácia quando aplicado a um pequeno conjunto de teste rotulado manualmente. Os resultados da Análise de Sentimento ao longo da campanha eleitoral não demonstraram forte correlação com as pesquisas eleitorais, e tampouco verificou-se uma clara preferência dos usuários do Twitter por um dos candidatos.

7.1 Considerações Finais

A Análise de Sentimento em tweets apresenta uma série de desafios que foram observadas no decorrer desse trabalho. O formato do Twitter resulta em atributos esparsos, e há casos em que o texto por si não é o suficiente para analisar o sentimento. A aparição de gírias, erros de digitação e *memes* também dificulta o trabalho de classificação.

Em um amplo estudo sobre fake news [28] constatou-se que o baixo custo e a facilidade de propagação de tweets constituem os mesmos fatores de estímulo de propagação de notícias de baixo custo intencionalmente falsificadas para disseminar confusão e conturbar a realidade factual. No Brasil, a distorção causada pelas fake news no cenário político é semelhante, ou talvez mais acentuada do que nos países analisados no estudo.

Por essa razão, tem sido trabalhado também com metadados (aumento ou diminuição de valores, frequência relativa associada a eventos do cenário político etc.) como forma de mitigar o efeito nefasto do conteúdo intencionalmente distorcido de factóides gerados por fake news.

7.2 Trabalhos Futuros

A Análise de Sentimento pode contar com os demais dados do tweet além do texto, podendo ser feita uma pré-seleção dos tweets neutros. Sua precisão pode aumentar com a utilização de um léxico mais específico como o *Unilex*, bem como com a utilização de técnicas de aprendizado de máquina.

Para isso faz-se necessário um conjunto de tweets rotulados maior do que o apresentado, e aplicações web de sistemas conhecidos como o *ifeel2.0* podem ser úteis.

Há ainda, outras técnicas não abordadas neste trabalho para a análise. A lematização é uma técnica concorrente ao *stemming*, mais complexa, que traz resultados potencialmente melhores [29]. O LDA (*Latent Dirichlet Allocation*) é uma técnica de clusterização que agrupa documentos por assunto [30].

Uma vez construído um modelo mais robusto, tanto para a coleta quanto para a análise do sen-

timento dos tweets, futuros eventos de escala nacional, de natureza política ou não, poderão ser analisados em tempo real.

Agradecimentos

Ao meu orientador José Artur por me auxiliar e motivar sempre, à minha família pelo apoio constante e à minha amiga Julia Yoko pelo apoio e auxílio na classificação manual dos tweets.

Referências

[1] LIU Bing. *Sentiment analysis: mining opinions, sentiments, and emotions*. Cambridge University Press, 2015.

[2] POZZI Federico Alberto et al. *Sentiment analysis in social networks*. Morgan Kaufmann, 2016.

[3] LOPES Flávia Valério. A reconfiguração dos veículos tradicionais de informação frente à popularização das mídias sociais. *XV Congresso de Ciências da Comunicação na Região Sudeste*, 2010. Acesso em 30 jun. 2018. Disponível em: http://www.intercom.org.br/papers/regionais/sudeste2010/resumos/R19-0905-1.pdf.

[4] BHAYANI Richa, HUANG Lei, and GO Alec. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 2009. Acesso em 5 fev. 2019. Disponível em: http://www.yuefly.com/Public/Files/2017-03-07/58beb0822faef.pdf.

[5] KEARNEY Michael. Package 'rtweet'. *CRAN Repository*, 2018. Acesso em 30 jun. 2018. Disponível em: https://cran.r-project.org/web/packages/rtweet/rtweet.pdf.

[6] GENTRY Jeff. Package 'twitter'. *CRAN Repository*, 2016. Acesso em 30 jun. 2018. Disponível em: https://cran.r-project.org/web/packages/twitteR/twitteR.pdf.

[7] TUMASJAN Andranik et al. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 2010. Acesso em 30 jun. 2018. Disponível em: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1441/1852.

[8] ANJARIA Malhar and GUDDETI Ram Mohana Reddy. Influence factor based opinion mining of twitter data using supervised learning. *IEEE*, 2014. Acesso em 30 jun. 2018. Disponível em: https://ieeexplore.ieee.org/document/6734907/.

[9] KAGAM Vadim et al. Using twitter sentiment to forecast the 2013 pakistani election and the 2014 indian election. *IEEE*, 2015. Acesso em 30 jun. 2018. Disponível em: https://ieeexplore.ieee.org/abstract/document/7030167/.

[10] GRAGNANI Juliana. Exclusivo: investigação revela exército de perfis falsos usados para influenciar eleições no brasil. *BBC Brasil em Londres*, 2017. Acesso em 26 abr. 2018. Disponível em: http://www.bbc.com/portuguese/brasil-42172146.

[11] RUEDIGER Marco Aurélio et al. Robôs, redes sociais e política no brasil: estudo sobre interferências ilegítimas no debate público na web, riscos à democracia e processo eleitoral de 2018. *FGV, DAPP*, 2017. Acesso em 30 jun. 2018. Disponível em: https://bibliotecadigital.fgv.br/dspace/handle/10438/18695.

[12] CONOVER Michael D et al. Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*, 2011. Acesso em 22 jul. 2019. Disponível em: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPaper/2847.

[13] Using twitter. 2019. Acesso em 22 jul. 2019. Disponível em: https://help.twitter.com/en/using-twitter.

[14] ZIMBRA et al. The state-of-the-art in twitter sentiment analysis: a review and benchmark evaluation. *ACM Transactions on Management Information Systems (TMIS)*, 9(2):5, 2018. Acesso em 22 jul. 2019. Disponível em: https://dl.acm.org/citation.cfm?id=3185045.

[15] DE SOUZA Karine França, PEREIRA Moisés Henrique Ramos, and DALIP Daniel Hasan. Unilex: Método léxico para análise de sentimentos textuais sobre conteúdo de tweets em português brasileiro. *Abakós*, 5(2):79–96, 2017. Acesso em 22 jul. 2019. Disponível em: https://dicionariounilex.wixsite.com/unilex/publicacoes.

[16] ARAUJO Matheus Lima et al. ifeel 2.0: A multilingual benchmarking system for sentence-level sentiment analysis. In *Tenth International AAAI Conference on Web and Social Media*, 2016. Acesso em 22 jul. 2019. Disponível em: https://homepages.dcc.ufmg.br/~matheus.araujo/ifeel2.pdf.

[17] Search tweets. 2019. Acesso em 15 jan. 2019. Disponível em: https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html.

[18] Pesquisa datafolha: Bolsonaro, 24%; ciro, 13%; marina, 11%; alckmin, 10%; haddad, 9%. 2018. Acesso em 15 jan. 2019. Disponível em: https://g1.globo.com/politica/eleicoes/2018/noticia/2018/09/10/pesquisa-datafolha-bolsonaro-24-ciro-13-marina-11-alckmin-10-haddad-9.ghtml.

[19] HENRIQUE Jefferson. Get old tweets. *Github*, 2016. Acesso em 30 jun. 2018. Disponível em: https://github.com/Jefferson-Henrique/GetOldTweets-python.

[20] Ibope inteligência eleições. 2018. Acesso em 15 jan. 2019. Disponível em: http://www.ibopeinteligencia.com/eleicoes/.

[21] Datafolha eleições presidente. 2018. Acesso em 15 jan. 2019. Disponível em: http://datafolha.folha.uol.com.br/eleicoes/2018/presidente/indice-1.shtml.

[22] Resultados das pesquisas eleitorais para presidente xp/ipespe. 2018. Acesso em 15 jan. 2019. Disponível em: https://www.infomoney.com.br/mercados/politica/noticia/7513479/resultados-das-pesquisas-eleitorais-para-presidente-xpipespe.

[23] PORTER Martin F. Snowball: A language for stemming algorithms. 2001. Acesso em 22 jul. 2019. Disponível em: https://snowballstem.org/.

[24] PEREIRA Giuliano Lemes. Projeto analise de sentimentos. 2017. Acesso em 22 jul. 2019. Disponível em: https://rpubs.com/giuice/analisesentimentos1.

[25] CHEN Yanqing and SKIENA Steven. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, 2014. Acesso em 22 jul. 2019. Disponível em: https://www.aclweb.org/anthology/P14-2063.

[26] Veja linha do tempo dos acontecimentos políticos até as eleições 2018. 2018. Acesso em 30 jun. 2018. Disponível em: https://www.correiobraziliense.com.br/app/noticia/politica/2018/10/28/interna_politica,715609/eleicoes-2018-tem-disputa-inedita-nas-ruas-e-nas-redes.shtml.

[27] Eleições 2018: calendário, debates e programa dos candidatos à presidência do brasil. 2018. Acesso em 30 jun. 2018. Disponível em: https://brasil.elpais.com/brasil/2018/08/14/politica/1534276152_537579.html.

[28] SHU Kai et al. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017. Acesso em 22 jul. 2019. Disponível em: https://arxiv.org/abs/1708.01967.

[29] DA SILVA João Ricardo Martins Ferreira. *Shallow processing of Portuguese: From sentence chunking to nominal lemmatization*. PhD thesis, Universidade de Lisboa, 2007. Acesso em 22 jul. 2019. Disponível em: http://xisque.di.fc.ul.pt/publicacoes/Silva2007.pdf.

[30] CAHYANINGTYAS Risma Mustika et al. Emotion detection of tweets in indonesian language using lda and expression symbol conversion. In *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*, pages 253–258. IEEE, 2017.