UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE AUTOMAÇÃO E SISTEMAS

Pedro Casali Guedes

Aplicação de técnicas de *data mining* para previsibilidade eleitoral

Florianópolis
2018

Pedro Casali Guedes

Aplicação de técnicas de *data mining* para previsibilidade eleitoral

Relatório submetido à Universidade Federal de Santa Catarina como requisito para a aprovação na disciplina **DAS 5511: Projeto de Fim de Curso** do curso de Graduação em Engenharia de Controle e Automação.
Orientador: Prof. Hector Bessa Silveira

Florianópolis
2018

Pedro Casali Guedes

Aplicação de técnicas de *data mining* para previsibilidade eleitoral

Esta monografia foi julgada no contexto da disciplina DAS5511: Projeto de Fim de Curso e aprovada na sua forma final pelo Curso de Engenharia de Controle e Automação.

Florianópolis, 13 de dezembro de 2018

Banca Examinadora:

Prof. João Artur de Souza
Orientador Local
IGTI – EGC - UFSC

Prof. Hector Bessa Silveira
Orientador no Curso
Universidade Federal de Santa Catarina

Prof. Marcelo Ricardo Stemmer
Avaliador
Universidade Federal de Santa Catarina

Bruno Roberto Carvalho
Debatedor
Universidade Federal de Santa Catarina

Guilherme Cornelli Souza
Debatedor
Universidade Federal de Santa Catarina

A meus pais, Luís e Cíntia, pelo apoio incondicional.

AGRADECIMENTOS

Agradeço, primeiramente, a meus pais, Cíntia e Luís. Todo o amor e suporte que vocês me deram desde meu nascimento foram imprescindíveis para que eu chegasse até aqui.

À minha irmã, um grande exemplo de pessoa e futura médica e a quem devo muitas de minhas conquistas.

Ao meu orientador, Prof. Hector Bessa Silveira, pela paciência e pelos conselhos ao longo deste projeto e da graduação como um todo.

Aos meus orientadores no IGTI, Prof. João e Alessandro, pelos ensinamentos, pela experiência e apoio.

A meus colegas de turma, pelo companheirismo e pela amizade.

A meus amigos de longa data, pelas risadas e pela inspiração contínua.

A todos os docentes que tive ao longo de minha carreira acadêmica, na UFSC e na França, pelo aprendizado.

A demais familiares, amigos, colegas e todos aqueles que contribuíram para esse trabalho.

Muito obrigado!

“A poderosa peça continua e você pode contribuir com um verso”.
(Walt Whitman)

RESUMO

Após casos de imprevisibilidades eleitorais nos Estados Unidos e no Reino Unido em 2016, havia uma preocupação para que o mesmo ocorresse nas eleições presidenciais brasileiras de 2018 desde o início do período eleitoral. A Ciência Política, por sua vez, apresenta diversas teorias que buscam avaliar essa questão. Entretanto, há limitações quanto aos dados utilizados nas pesquisas dessa área, além de não considerarem os efeitos das redes sociais em suas análises. Por fim, com a Reforma Político-Eleitoral de 2017, a duração das campanhas eleitorais foi reduzida, bem como foram impostas limitações financeiras a elas, demandando novas estratégias dos candidatos. Desse modo, este projeto propõe um novo método de prever resultados de eleições, baseando-se na teoria de estratégias emocionais da Neuropolítica. Aplicando a *Design Science Research Methodology* (DSRM), uma das metodologias mais utilizadas em projetos de pesquisa, o objetivo deste trabalho é categorizar as diferentes publicações (de 1° de janeiro até 6 de outubro de 2018) na rede social *Twitter* de oito dos principais candidatos à Presidência da República conforme os sentimentos transmitidos. Para isso, foram aplicadas e avaliadas diversas técnicas de mineração de opiniões (uma aplicação de *data mining*), como: aquisição de dados por meio de um *web scraper*, pré-processamento dos textos utilizando de diferentes ferramentas, categorização (com base no uso de léxicos e de seis algoritmos de aprendizado de máquina) e até a apresentação visual dos resultados. Para a construção do modelo de categorização, optou-se pelo uso do algoritmo de aprendizado de máquina de Naïve-Bayes, sendo este um dos trabalhos pioneiros no Brasil em uso de técnicas de *machine learning* na previsibilidade eleitoral. A partir de sua aplicação, os *tweets* publicados foram classificados em três emoções: positivas, negativas e neutras. Os resultados obtidos neste trabalho foram preliminares e não conclusivos em relação à previsão das eleições. Entretanto, foi permitido diferenciar as estratégias emocionais de cada candidato, as quais apresentaram relação com eventos externos que ocorreram ao longo do período eleitoral, comprovando que a teoria da Neuropolítica aplicada possui influência nas campanhas e que o modelo de classificação criado é válido para avaliá-la. Com isso, o trabalho apresenta perspectivas de maiores contribuições futuras à área de previsibilidade eleitoral com base no uso de técnicas de aprendizagem de máquina.

Palavras-chave: mineração de textos. Análise de sentimentos. Previsibilidade eleitoral.

ABSTRACT

After cases of electoral unpredictability in the United Stated and in the United Kingdom, in 2016, there were concerns that the same could happen in the 2018 Brazilian presidential elections since the beginning of the campaign period. Within Political Science, there are a few different theories which focus on studying such issues. However, there are some limitations with respect to the data used in those researches, as well as not including the effects of social media in their analysis. Finally, with the Political-Electoral Reform of 2017, campaigns' durations were reduced as well as their allowed budget, which demands new strategies by the candidates. With such scenario, this project proposes a new method to predict election results, based on Neuropolitics' emotional strategies theory. By applying the Design Science Research Methodology (DSRM), one of the most used methodologies in research projects, the goal is to categorize *tweets* shared by eight of the main candidates to Brazil's Presidency during the period of January 1st until October 6th of 2018, according to their emotions. In order to achieve this, many techniques of text mining (one of many areas of data mining) were applied and evaluated, such as: data acquisition through a web scraper, text pre-processing with the use of different tools, classification (based on the use of dictionaries as well as six different machine learning algorithms) and even visually presenting the results. For building the classification model, it was applied machine learning's Naïve-Bayes Algorithm, which makes this work one of the firsts in Brazil to use machine learning techniques in electoral predictability. With it, the available tweets were classified into three different emotions: positive, negative and neutral. The results obtained with this work were preliminary and non-conclusive about its capability of predicting elections. However, one was able to differentiate each candidate's emotional strategies, which showed relation to events (such as breaking news) that occurred during the electoral period, thus showing that the Neuropolitcs' theory applied does have influence in campaigns and that the classification model built is valid to evaluate this theory. Therefore, this work presents future perspectives for significant contributions in electoral predictability based on the use of machine learning techniques.

Key-words: text mining. Sentiment analysis. Electoral predictability.

LISTA DE ABREVIATURAS E SIGLAS

IBGE - Instituto Brasileiro de Geografia e Estatística

UFSC - Universidade Federal de Santa Catarina

TSE - Tribunal Superior Eleitoral

PT - Partido dos Trabalhadores

PSDB - Partido da Social Democracia Brasileira

PSL - Partido Social Liberal

EUA - Estados Unidos da América

IGTI - Núcleo de Estudos em Inteligência, Gestão e Tecnologias para Inovação

EGC - Departamento de Engenharia e Gestão do Conhecimento

PFC - Projeto de Fim de Curso

TF - *Term Frequency*

IDF - *Inverse Document Frequency*

CSV - *Comma Separated Values*

TXT - *Text File*

BoW - Bag of Words

NOVO - Partido Novo

PMB - Partido da Mulher Brasileira

SUMÁRIO

1 INTRODUÇÃO

Neste capítulo, primeiramente serão apresentados, brevemente, o contexto do projeto e o que motivou seu desenvolvimento. Em seguida, nas Seções 1.1, 1.2 e 1.3, serão detalhados, respectivamente, os objetivos geral e específicos, bem como qual a solução proposta para resolver o problema identificado. Na sequência, na Seção 1.4, será apresentada a instituição em que esse trabalho foi desenvolvido, o laboratório IGTI. Por fim, a última Seção deste capítulo apresentará a estrutura do documento.

No segundo semestre de 2018, ocorreram as eleições presidenciais brasileiras. Em 2016, com a ocorrência de casos cruciais ao redor do mundo em que as previsões eleitorais falharam (no Reino Unido e, após, nos EUA), havia um risco de que o mesmo fosse ocorrer no Brasil. Além disso, em 2017, foi publicada a Lei nº 13.488, popularmente conhecida como Reforma Político-Eleitoral, que alterou diversas regras para o pleito deste ano e, por consequência, influenciaria as estratégias das campanhas dos candidatos em 2018 e nas eleições seguintes, já que reduziu a duração das campanhas e limitou o acesso a recursos. Por fim, as diferentes teorias de decisão de voto da Ciência Política ainda dependem fortemente da realização de questionários (*surveys*) junto a uma amostra de eleitores para conseguirem validar suas hipóteses, sem aplicação de técnicas avançadas de ciência de dados e sem avaliarem a influência de redes sociais nos pleitos.

A partir desse conjunto de constatações, o presente trabalho trata do problema de desenvolver novos métodos de previsão de resultados eleitorais, sem depender apenas de pesquisas de opinião e, em paralelo, aplicar técnicas mais modernas de aquisição e análise de dados para contribuir às pesquisas da Ciência Política, tornando-se um dos trabalhos pioneiros no Brasil. A partir disso, pode-se definir os objetivos desse projeto, a serem listados a seguir.

1.1 Objetivo Geral

Dado o cenário introduzido acima, define-se como objetivo geral deste Projeto de Fim de Curso: “**aplicar técnicas de *data mining* para previsão de campanhas eleitorais**”.

1.2 Objetivos específicos

Com base no objetivo geral do projeto, determinam-se cinco objetivos específicos:

- Demonstrar diferentes modelos de decisão de voto, analisá-los criticamente quanto à aplicabilidade em previsão eleitoral e selecionar um para se usar no projeto;
- Levantar as possíveis fontes de dados a serem usadas em previsibilidade do voto e avaliar qual(is) utilizar;
- Apresentar e aplicar diferentes técnicas de *data mining* para texto (mineração de opiniões);

- Analisar criticamente os resultados obtidos com as técnicas de *data mining* para validar o método proposto;
- Apontar novas ferramentas, metodologias e técnicas a serem usadas no futuro para aprimorar as previsões eleitorais.

1.3 Solução proposta

A partir dos objetivos mencionados acima, a solução proposta consiste em desenvolver e avaliar uma nova maneira de se prever resultados de eleições. A metodologia utilizada neste trabalho foi a *Design Science Research Methodology* (DSRM), que é referência na literatura para projetos de pesquisa e será apresenta no Capítulo 5. O desenvolvimento desse novo método foi baseado na teoria de estratégias emocionais da Neuropolítica (ver Seção 3.5), uma abordagem mais moderna de previsibilidade eleitoral e ainda pouco estudada dentro da Ciência Política, mas que coaduna com o cenário eleitoral brasileiro de 2018. Ela consiste em avaliar os efeitos que as emoções transmitidas pelos candidatos possuem no direcionamento do voto do eleitorado. Neste projeto, optou-se pela análise dos sentimentos transmitidos pelas campanhas nas publicações em redes sociais, graças ao crescimento de seu uso pelos eleitores e também por conta das limitações impostas pela Reforma Político-Eleitoral de 2017 (ver Seção 2.1.3). Dentre as diversas mídias, a utilizada neste projeto foi o *Twitter*, devido a restrições de acesso às demais, como *Facebook* e *WhatsApp*.

Para essa análise, foram aplicadas diversas técnicas de mineração de opiniões, uma aplicação de *data mining* específica para conteúdos em texto. O método proposto, portanto, inclui um sistema de aquisição automática dos dados do *Twitter* (chamado de *web* scraper), pré-processamento dos textos, avaliação de diferentes abordagens de categorização e a apresentação visual dos resultados obtidos. Dessa forma, é um dos pioneiros no Brasil ao utilizar dessas técnicas em previsibilidade eleitoral.

Ao longo do trabalho, foram usadas algumas das ferramentas mais utilizadas em projetos de *data mining*, como as *open source R*, *Weka* e *Orange*. Através delas, pôde-se criar um modelo de classificação de *tweets* com uso do algoritmo de Naïve-Bayes. Com sua aplicação, notou-se diferentes estratégias emocionais utilizadas pelas campanhas de oito dos principais candidatos à Presidência da República em 2018, o que comprova a relevância da teoria escolhida neste trabalho, a Neuropolítica. Entretanto, os resultados obtidos foram preliminares e não conclusivos, o que impediu encontrar relação direta entre os sentimentos transmitidos e o resultado do pleito. Apesar disso, há perspectivas de que esse método possa dar origem a novas abordagens de previsibilidade eleitoral para a Ciência Política, já que permitiu diferenciar as estratégias utilizadas, bem como encontrar efeitos de eventos externos nas campanhas.

1.4 IGTI: o laboratório incubador

Com o objetivo de captar suporte humano e estrutural para realizar o Projeto de Fim de Curso, foram buscadas alternativas de laboratórios na UFSC que pudessem me ajudar. Ao pesquisar na Internet a respeito de Professores e/ou trabalhos, teses e

artigos da Universidade relacionados à política e tecnologia, o Núcleo de Estudos em Inteligência, Gestão e Tecnologias para Inovação (IGTI) apresentou-se como uma boa oportunidade.

Criado em 1997, o IGTI (Núcleo de Estudos em Inteligência, Gestão e Tecnologias para Inovação) está vinculado ao Departamento de Engenharia e Gestão do Conhecimento (EGC) da UFSC e o objetivo do grupo é desenvolver pesquisas aplicadas à gestão de negócios. Um dos pontos fortes do IGTI é a característica multidisciplinar de sua equipe, formada por Engenheiros, Estatísticos, Cientistas da Computação, Matemáticos, dentre outras formações e de todos os níveis acadêmicos - estudantes de graduação em estágio ou fazendo Trabalho de Conclusão de Curso, mestrandos e doutorandos.

Além disso, dentre suas linhas de pesquisa, identificam-se algumas que se encaixam diretamente com o presente trabalho:

- Inteligência para Inovação: estudo de métodos e ferramentas que viabilizam a coleta e análise de dados e informações dos ambientes externos e internos da organização, com técnicas relacionadas à *business intelligence* (BI), inteligência competitiva, dentre outras;
- Tecnologias da Informação aplicadas à Gestão de negócios: estudo de métodos, técnicas e ferramentas que visem tornar uma organização mais competitiva, com foco na busca por oportunidades nas áreas de tecnologia da informação e gestão da informação.

Por fim, vale comentar que duas dissertações de mestrado realizadas no IGTI trabalhavam com a área de mineração de opiniões, inclusive com uma delas servindo como referência para este trabalho.

1.5 Estrutura do documento

No Capítulo 2, serão apresentados em mais detalhes o contexto, a motivação e o problema abordado neste projeto, o que inclui: imprevisibilidades eleitorais ao redor do mundo, perspectivas para as eleições presidenciais de 2018 no Brasil, a Reforma Político-Eleitoral de 2017, histórico de projetos acadêmicos em previsibilidade eleitoral e outros projetos com temática similar.

Nos Capítulos 3 e 4, serão apresentados os diversos conceitos necessários para compreensão das técnicas utilizadas e da solução proposta. Para isso, no Capítulo 3, as diferentes teorias de decisão de voto mais aceitas pela literatura da Ciência Política serão explicadas. Em seguida, já no Capítulo 4, serão apresentados os aspectos conceituais relacionados à Engenharia de Controle e Automação que foram requisitados ao longo do projeto. Estes foram divididos em dois grandes temas: desenvolvimento de sistemas e reconhecimento de padrões, o que inclui os conceitos de *data mining* e os algoritmos de aprendizado de máquina utilizados.

Na sequência, no Capítulo 5, será apresentada a metodologia *Design Science Research Methodology*, referência na literatura para projetos de pesquisa, bem como as diferentes ferramentas usadas para sua implementação.

Já no Capítulo 6, serão detalhadas as diferentes etapas do desenvolvimento deste PFC, desde a etapa inicial de pesquisa (para detalhamento do escopo do projeto) até a parte de geração de informações por meio da aplicação das técnicas de *data mining*.

Os resultados obtidos para a previsibilidade eleitoral serão exibidos e comentados no Capitulo 7, incluindo um comparativo com eleições anteriores.

Por fim, no Capítulo 8 deste documento, serão apresentadas as conclusões do projeto desenvolvido e perspectivas de melhorias e/ou trabalhos futuros.

2 MOTIVAÇÃO E DESCRIÇÃO DO PROBLEMA

Neste capítulo, serão apresentadas as principais justificativas para o desenvolvimento do presente projeto. Primeiramente, na Seção 2.1, será feita uma descrição do cenário político brasileiro, o que inclui os principais casos recentes de imprevisibilidade eleitoral ao redor do mundo, as perspectivas para a eleições presidenciais brasileiras de 2018 e a Reforma Político-Eleitoral de 2017. Em seguida, na Seção 2.2, será feita uma análise do Estado-da-arte a respeito da temática do projeto, incluindo projetos acadêmicos e também do setor privado. Ao final do capítulo, na Seção 2.3, será feita uma recapitulação de todo esse contexto, resumindo o problema a ser resolvido por este trabalho.

2.1 Contexto

Nessa seção, serão apresentados os principais acontecimentos eleitorais que anteciparam o projeto, primeiramente em âmbito mundial e, em seguida, nacional.

2.1.1 Imprevisibilidade eleitoral ao redor do mundo

Recentemente, duas disputas eleitorais tiveram, além de forte impacto econômico e diplomático ao redor do mundo, resultados ditos “imprevisíveis”: o *Brexit*, no Reino Unido, e a disputa presidencial dos Estados Unidos, ambos em 2016.

Integrado à União Europeia desde 1973 (na época, a UE ainda era chamada de Comunidade Econômica Europeia - CEE), o Reino Unido nunca teve uma relação estável com o bloco econômico. Já em 1975, apenas dois anos após a adesão do país ao bloco, houve um referendo para decidir se ele permaneceria ou não na CEE. A população foi às urnas e optou pela conservação. Outra peculiaridade da relação dos britânicos era a questão monetária, já que o país optou por não ingressar na Zona do Euro, mantendo a libra esterlina como sua moeda oficial, o que também causava atritos com a UE (FERNANDES & SILVA, 2018).

Em 2015, durante as eleições gerais no Reino Unido, David Cameron, então Primeiro-Ministro britânico, propôs a realização de um referendo para avaliar a permanência do reinado na UE, caso seu partido vencesse o pleito, o que de fato ocorreu. Marcou-se então para o dia 23 de junho de 2016 a realização da votação, que ficou conhecida como *Brexit*, uma abreviação das palavras *britain* (Bretanha) e *exit* (saída). O resultado final foi surpreendente: com 51,9% dos votos válidos (totalizando mais de 17 milhões de votos), optou-se pela saída do país do bloco econômico (FERNANDES & SILVA, 2018).

O desfecho do plebiscito tornou-se um “*fracasso no mundo das previsões eleitorais*”, segundo Cohn (2016). Cinco horas antes do resultado ser divulgado, as casas de apostas britânicas apontavam 88% de chance de vitória para a permanência, métrica esta que até então era um previsor confiável (COHN, 2016). O mercado financeiro também não esperava essa conclusão, resultando em uma forte correção nos mercados de ações e câmbio. No geral, o *Brexit* liderou em apenas quatro das últimas onze pesquisas antes da votação, o que fez com que o “fico” tivesse 0,5% de

chance a mais ganhar, segundo o *HuffPuffPolster*, e 4% a mais, segundo a *YouGov*. Por fim, os próprios eleitores acreditavam em vitória da permanência, o que também, até então, funcionava razoavelmente bem para prever resultados (COHN, 2016).

Curiosamente, segundo Cohn (2016), o fracasso nas previsões do *Brexit* deveria ser aprendizado para outra votação que atrairia os holofotes mundiais em novembro do mesmo ano: as eleições presidenciais estadunidenses.

Com o final do segundo e último mandato do então Presidente norte americano Barack Obama se aproximando, os Estados Unidos deveriam escolher seu sucessor (ou sucessora) no dia 8 de novembro de 2016. A disputa, como de costume naquele país, foi muito polarizada, com o bipartidarismo cada vez mais evidente. De um lado, havia Hillary Clinton, do Partido Democrata, o mesmo de Obama. Ao longo de sua carreira na vida pública, ela já tinha sido Primeira-Dama, Senadora e Secretária de Estado, tornando-se a primeira mulher da história do país candidata à Presidência da República por um *major party* (um dos dois partidos mais relevantes dos EUA). Seu concorrente era Donald Trump, do Partido Republicano, um empresário bilionário e personalidade televisiva (STACK, 2016).

O resultado foi, novamente, inesperado. Donald Trump venceu, tornando-se o primeiro Presidente da história da nação sem experiência alguma em cargo público ou nas Forças Armadas. A contagem final dos votos do colégio eleitoral foi de 304 eleitores para Trump contra 227 para Clinton (SCHMIDT & ANDREWS, 2016). Assim como no caso do *Brexit*, o consenso de especialistas estava errado, já que previa vitória da candidata Democrata, apesar dos alertas de Cohn (2016).

Segundo a *Agence France-Presse* (AFP), terceira maior agência de notícias do mundo, das vinte maiores empresas de pesquisas que realizaram sondagens durante os dois últimos meses da campanha, apenas uma indicava vitória de Trump. O site *FiveThirtyEight*, um dos mais reconhecidos nos EUA por conta de seu trabalho preditivo de votações, projetou vitórias de Hillary nos quatro estados-chave das eleições (Pensilvânia, Flórida, Carolina do Norte e Wisconsin - todos vencidos por Trump) e 70% de chance de tornar-se Presidente. O *The New York Times*, por sua vez, estimou 83% de probabilidade de vitória da democrata (G1, 2016).

Não eram apenas as pesquisas que estavam equivocadas. O setor de apostas também errou em suas previsões. Nas casas de aposta do Reino Unido, no dia da votação, as chances de Clinton estavam em -300 (ou seja, uma pessoa precisaria apostar £300 para lucrar £100), enquanto as de Trump em +275 (para cada £100 apostados, o retorno seria de £275). Foram diversos os sites de aposta que também erraram em suas projeções, como *Sportsbook*, *Bookmaker* e *Ladbokes*. Muitas das agências de apostas, inclusive, tiveram problemas financeiros para pagar os apostadores vencedores após as eleições (G1, 2016).

Por fim, o mesmo ocorreu com o mercado financeiro. Especialistas do *Citibank*, por exemplo, colocavam as probabilidades de vitória da Hillary entre 70 e 80%, enquanto Trump não teria nem 33% de chance (SHEN, 2016).

Nota-se, portanto, um certo padrão em duas das principais disputas eleitorais no mundo dos últimos anos, algo que poderia se repetir no Brasil em 2018, conforme a ser apresentado na próxima seção.

2.1.2 Imprevisibilidade eleitoral no Brasil

No Brasil, por sua vez, as eleições presidenciais mostravam-se imprevisíveis desde seu início. O país vivenciava um período de alta instabilidade política desde as eleições presidenciais de 2014, graças à polarização que ela causou. Nesses 4 anos entre pleitos, passou por uma forte crise econômica, foram descobertos inúmeros casos de corrupção graças à Operação Lava-Jato e houve até um processo de *impeachment* da então Presidente da República Dilma Rousseff. Destaca-se ainda que não cabe a este projeto avaliar a efetividade nem a validade desses acontecimentos, mas todos, independentemente dessa avaliação, tiveram forte impacto nas eleições de 2018.

Neste ano de 2018, chegou-se ao período eleitoral com um governo federal com a maior taxa de reprovação desde que tal indicador passou a ser medido (AMORIM, 2018) e com brasileiros com um forte anseio por renovação. Segundo pesquisas, até 96% dos eleitores disseram não se sentirem representados pelos atuais políticos e até 93% disseram acreditar na necessidade da formação de novas lideranças políticas (G1, 2018). Entretanto, as novas “regras do jogo” provenientes da Reforma Político-Eleitoral apontavam para a menor renovação da história do Congresso Nacional, de acordo com alguns especialistas, e o mesmo poderia se refletir nos cargos em disputa do Poder Executivo (VENTURINI, 2018).

No início do período da campanha, estimava-se uma disputa pulverizada, com muitos candidatos recebendo porcentagens entre 5 e 15% dos votos, mas sem polarização entre dois, muito menos sem um líder claro com capacidade de vencer no primeiro turno. Além disso, tivemos a indefinição a respeito da elegibilidade de um candidato até poucas semanas antes das eleições. Em um certo momento, o site *Predict It*, um dos mais importantes do mercado preditivo, chegou até a apontar a candidata Marina Silva (REDE) como a favorita, mas ela terminou com apenas 1% dos votos válidos (AMENDOLA, 2018).

Aliado a tudo isso, ainda havia um firme crescimento dos impactos que a Internet teria em processos eleitorais. Com o acesso à rede progressivamente mais democrático, com mais pessoas usando redes sociais e compartilhando conteúdo por meio delas, a estratégia de uso da Internet passa a ser fundamental na campanha de um candidato. Enquanto o candidato Jair Bolsonaro (PSL), que estava bem posicionado nas pesquisas de intenção, baseava sua campanha quase que unicamente no uso das redes sociais, outros, principalmente o candidato Geraldo Alckmin (PSDB), apostavam no uso da televisão. Por fim, a criação e disseminação de notícias falsas (*fake news*) por meio da Internet (principalmente através das redes sociais) também poderia ter seu impacto no processo eleitoral (FLORES, 2018).

Portanto, observa-se um cenário nacional dito imprevisível no início das campanhas, com os métodos atuais de previsão mostrando-se falhos, assim como notado no exterior. Aliado a isso, as eleições de 2018 mostravam-se como um divisor

de águas na política nacional, graças à implementação da Reforma Político-Eleitoral, a ser descrita a seguir.

2.1.3 Reforma Político-Eleitoral de 2017: as novas regras do jogo

O Projeto de Lei n° 8.612 de 2017 (PL8612/2017), que foi aprovado em outubro do mesmo ano pela Câmara dos Deputados e pelo Senado, tornou-se na Lei Ordinária nº 13.488 de 2017 ao ser sancionada pelo então Presidente da República Michel Temer. Essa Lei é vital para se compreender o contexto das eleições federais de 2018, bem como das que ocorrerão nos próximos anos, pois trata-se da chamada “Reforma Político-Eleitoral”.

Seguindo anseios populares de uma reestruturação na estrutura político-eleitoral brasileira, o Congresso Nacional iniciou discussões e negociações para aprovar uma reforma. Para que as mudanças desejadas passassem a valer já nas eleições de 2018, era necessária a aprovação das novas regras até no máximo um ano antes de sua ocorrência. Ou seja, a Lei supracitada deveria ser publicada até o dia 6 de outubro de 2017 (CALEGARI, 2017).

Assim, depois de meses de debates e de acordos feitos nos bastidores (que incluíram até o pré-estabelecimento de vetos a serem dados pelo Presidente Temer), o PL8612/2017 foi sancionado com os devidos vetos, os quais foram aprovados pelo Senado, e publicado como a Lei 13.488/2017 justamente no último dia possível (CALEGARI, 2017). Não coube ao projeto, tampouco a este relatório, avaliar a efetividade nem a qualidade da reforma. Entretanto, compreender as alterações por ela impostas é fundamental, as quais são descritas na sequência.

2.1.3.1 Cláusula de barreira

Em 1965, foi criado pelo então Presidente da República Humberto Castello Branco, através do Artigo 60 da Lei Orgânica dos Partidos Políticos (LOPP), o Fundo Especial de Assistência Financeira aos Partidos Políticos, mais conhecido como **Fundo Partidário**. Ele é mantido com recursos de diferentes fontes: dotações orçamentárias da União, multas, outras verbas atribuídas conforme a aprovação de novas leis e até por doações de pessoas físicas ou jurídicas.

Antes da Reforma Político-Eleitoral de 2017, 5% dos recursos totais do fundo eram distribuídos igualmente a todos os partidos, independentemente se obtiveram ou não uma quantidade mínima de votos e/ou de eleitos, enquanto os demais 95%, por sua vez, eram divididos de forma proporcional à quantidade de votos recebidos para os candidatos à Câmara dos Deputados Federais. Em 2017, por exemplo, o fundo totalizou pouco mais R$750 milhões, segundo estimativas do Tribunal Superior Eleitoral (TSE). Conforme a legislação vigente na época, a divisão dos 5% do montante total (aproximadamente R$37,5 milhões) fez com que a cota mínima a receber era de pouco mais de R$1 milhão (que foi o caso de dois partidos: o NOVO e o PMB). Enquanto isso, o PT - partido que recebeu mais recursos do Fundo - totalizou quase R$100 milhões de receita em 2017 através dele. Já em 2018, o montante do

fundo deverá ser de mais de R$800 milhões, com a divisão sendo feita da mesma maneira.

A Reforma Político-Eleitoral de 2017, no entanto, alterou os critérios para distribuição dos recursos do Fundo Partidário para a partir do ano de 2019. Para ter acesso à qualquer cota do montante, um partido precisa atingir uma quantidade mínima de votos para a Câmara Federal de Deputados, que ficou conhecida como a cláusula de barreira ou cláusula de desempenho. Para as eleições de 2018, ficou estabelecida como a "barreira" um piso de 1,5% do total dos votos válidos para a Câmara, com pelo menos 1% dos votos válidos de 9 (ou um terço das) unidades da federação, **ou** eleger pelo menos 9 Deputados Federais em 9 estados diferentes. Nas eleições federais seguintes, as exigências aumentam gradativamente até atingir, em 2030, 3% dos votos válidos para a Câmara.

Além de não terem acesso algum ao Fundo Partidário, os partidos que não cumprirem com a cláusula de desempenho terão outras duas restrições a suas atuações: não poderão possuir estrutura própria e funcional nas casas do Congresso (terão espaço físico menor) e também não terão direito ao horário eleitoral de televisão e rádio, que até então era dividido de forma similar ao Fundo Partidário, com 10% do tempo dividido igualmente entre todos os partidos e o restante era partilhado conforme a quantidade de Deputados eleitos.

Com essa alteração, a perspectiva é de que algumas legendas chamadas de "nanicas" deixem de existir sem o fundo partidário: sem exposição na TV e rádio, dificulta a eleição de seus candidatos; sem eleger seus candidatos, fica sem acesso a esses recursos (Fundo, TV), entrando num ciclo vicioso. Nas eleições deste ano, 14 dos 35 partidos não cumpriram com a cláusula (G1, 2018). A medida, segundo alguns especialistas, visa acabar com essas siglas, muitas das quais existem apenas para formar coligações com partidos maiores, oferecendo a eles mais tempo de TV, sem ter representatividade entre o eleitorado brasileiro. Por fim, a cláusula de barreira também dificulta o surgimento de novos partidos, já que estes não terão financiamento público nem acesso a propagandas de rádio ou televisão até que atinjam uma popularidade considerável.

2.1.3.2 Coligações partidárias

Para que candidatos a vereador e deputado de um partido tivessem mais chances de serem eleitos, era comum a prática da formação de alianças entre partidos, chamadas de coligações, muitas vezes independente do alinhamento ideológico dos coligados.

A Reforma Político-Eleitoral, portanto, propôs a proibição de coligações para eleições aos cargos cuja escolha ocorre por meio do voto proporcional. Entretanto, devido à grande resistência dos próprios partidos e políticos, o Presidente Michel Temer optou por colocar em prática essa nova regra apenas nas eleições de 2020. Vale ressaltar que, para as eleições com voto majoritário (para prefeitos, governadores, presidente e senadores), as coligações partidárias permanecem autorizadas.

2.1.3.3 Fundo Eleitoral

Um dos pontos mais polêmicos da Reforma Político-Eleitoral foi a criação do Fundo Especial de Financiamento de Campanha (FEFC), mais conhecido como Fundo (ou Fundão) Eleitoral. Financiado apenas por meio de dotações orçamentárias da União (ou seja, composto apenas por dinheiro público), em 2018, no primeiro ano de sua existência, ele totalizou um montante de R$1,71 bilhão. Entretanto, durante as discussões no Congresso a respeito da Reforma, chegou-se a cogitar um valor de até R$3,6 bilhões.

Ao contrário do Fundo Partidário, que existe para custear a manutenção dos partidos em todos os anos, o FEFC foi criado para apoiar as campanhas eleitorais e existe apenas em anos de eleição, começando pela de 2018. A divisão do Fundão também é diferente:

- 2% é igualmente dividido entre todos os 35 partidos registrados;
- 35% entre os partidos que tenham pelo menos um Deputado na Câmara Federal, divididos conforme a proporção dos votos conquistados nas últimas eleições;
- 48% divididos entre os partidos de forma proporcional ao número de deputados na Câmara, considerando os titulares do mandato;
- 15% divididos entre os partidos de forma proporcional ao número de senadores, considerando os titulares do mandato.

Ao contrário do que ocorre com o Fundo Partidário, o Eleitoral é distribuído aos partidos conforme as regras acima, independente do cumprimento ou não da cláusula de barreira. Existe também uma exigência quanto à distribuição dos recursos do FEFC de um partido a seus candidatos: 30% deve ser transferido a candidatas mulheres. Ademais, não há outras regras: cada partido pode dividir como quiser, seja por probabilidade de eleição ou igualmente entre os candidatos.

Em 2018, os partidos que mais receberam verbas pelo FEFC foram, em ordem: MDB (pouco mais de R$234 milhões), PT (pouco mais de R$212 milhões) e PSDB (quase R$186 milhões). Seis partidos receberam a mesma cota mínima, a partir da divisão de 2% do Fundão, o que totalizou em pouco mais de R$978 mil para cada.

2.1.3.4 Limite de gastos

Até então, não havia limite de gasto imposto aos candidatos. A Reforma Político-Eleitoral instituiu tetos, que variam conforme o cargo pleiteado e/ou a Unidade de Federação do candidato. Dessa forma, torna-se necessário que os candidatos busquem maneiras mais eficientes, financeiramente, para realizarem suas campanhas (incluindo usar mais da Internet e das redes sociais, por exemplo). O detalhamento das restrições é feito a seguir:

- Para as campanhas ao cargo de Presidente da República, colocou-se como limite para os candidatos um valor de R$70 milhões para o

primeiro turno. Em caso de segundo turno, o teto para o período da nova campanha é de R$35 milhões. Estes são os maiores tetos estipulados pela Reforma. Sendo assim, o máximo que um candidato pode gastar em sua campanha é R$105 milhões - aproximadamente um terço do que a ex-Presidente Dilma Rousseff, do PT, declarou gastar nas eleições de 2014;

- Para os candidatos ao Governo de suas Unidades da Federação, o teto para todo o período de campanha (incluindo segundo turno) varia entre R$2,8 milhões e R$21 milhões, dependendo da quantidade de eleitores da UF;
- Aos postulantes ao Senado, o limite máximo varia entre R$2,5 milhões e R$5,6 milhões, novamente conforme a quantidade de eleitores da UF;
- Para os que almejam uma vaga na Câmara Federal de Deputados, o teto imposto é de R$2,5 milhões - valor que independe da quantidade de eleitores do estado;
- Por fim, para os que desejam ingressar na Assembleia Legislativa de sua UF como Deputado Estadual/Distrital, o limite é de R$1 milhão, valor novamente independente da quantidade de eleitores.

2.1.3.5 Doações, *crowdfunding* e auto-financiamento

Em 2015, o Supremo Tribunal Federal (STF) proibiu que pessoas jurídicas pudessem investir em campanhas políticas. Com isso, boa parte das mudanças impostas pela Reforma Político-Eleitoral ocorreram nos critérios de financiamento, como a criação do Fundo Eleitoral, já apresentado anteriormente.

Uma das mais inovadoras foi a autorização da realização de *crowdfunding,* uma "vaquinha" *online*. Essa arrecadação pode ser feita já a partir do dia 15 de maio do ano da eleição, ou seja, mesmo antes de ter a candidatura registrada. Os *sites* realizadores dessas campanhas deverão divulgar a identidade dos doadores (por meio do número do Cadastro de Pessoa Física, o CPF) e o valor da doação. Além disso, passam a ser permitidas a promoção de eventos de arrecadação pelos partidos/candidatos e as vendas de bens ou serviços. Entretanto, nem tudo é novidade.

As doações de pessoas físicas se mantiveram inalteradas, com teto para o doador de 10% das receitas brutas obtidas no ano anterior. Chegou a ser discutida uma redução desse limite para 10 salários mínimos, porém ela não foi aprovada. O autofinanciamento dos candidatos também permanece inalterado, graças ao Presidente Michel Temer, que vetou a proposta vinda do Senado de limitar a até 10 salários mínimos. Desse modo, a totalidade dos gastos da campanha ainda pode vir de autofinanciamento (conforme os limites apresentados no tópico anterior).

2.1.3.6 Outras mudanças

A duração do horário eleitoral em rádio e televisão foi reduzida a 35 dias. Fora do período eleitoral, não serão permitidas propagandas de partidos ou candidatos. Também foi alterada a data de início das campanhas do segundo turno. Passa a ser

permitida a impulsão (paga) de publicações em redes sociais ou em ferramentas de buscas. Por fim, as emissoras de rádio e televisão que organizarem debates deverão convidar todos os candidatos de partidos com pelo menos 5 cadeiras na Câmara Federal (antes, o número mínimo era de 9). Novamente, essas mudanças impostas pela Reforma apontam para um uso mais intenso, pelos candidatos, da Internet.

De acordo com o exposto acima, conclui-se que o cenário político brasileiro para as eleições de 2018 apresentava-se com incertezas, tanto por conta de falhas recentes em previsões eleitorais ao redor do mundo, quanto devido à pulverização na disputa presidencial. Além disso, graças às novas regras impostas pela Reforma Político-Eleitoral, a perspectiva era de que os partidos e candidatos passassem a utilizar mais intensamente das redes sociais para as suas campanhas, o que vai de acordo com o novo perfil do eleitorado brasileiro, que busca, cada vez mais, a Internet para adquirir informações (DATAFOLHA, 2018).

A próxima seção apresenta uma revisão do estado-da-arte em previsibilidade eleitoral na Ciência Política.

2.2 Estado-da-arte

Desde que a Ciência Política passou a estudar a previsibilidade do voto, muitas teorias foram criadas e, para cada uma, há diversas metodologias de pesquisa utilizadas. Não cabe a esta seção detalhar a respeito das teorias em si, entretanto, serão apresentadas as diferentes maneiras que seus pesquisadores coletaram dados até então. O objetivo é demonstrar que há oportunidades para aplicar técnicas da Engenharia de Controle e Automação nas pesquisas de previsibilidade eleitoral, apresentando-a como uma ferramenta de grande valia para esse processo.

Para todas as três teorias clássicas da Ciência Política (Sociológica, Psicológica e Racional), as primeiras pesquisas eram feitas somente com base em *surveys* realizados em uma amostra especifica do eleitorado (FIGUEIREDO, 2018). Outra prática comum era utilizar de dados censitários fornecidos pelos governos. Entretanto, nesse caso, os pesquisadores não conseguem concluir com exatidão a respeito dos fatores influentes do voto, segundo Carreirão (2000): não se pode inferir a respeito de uma decisão individual apenas com base em dados do coletivo. Essas duas práticas (*surveys* e censos) dominaram os estudos de previsibilidade eleitoral por alguns anos, até que surgiram as pesquisas de opinião pré-voto (CARREIRÃO, 2000).

Com elas, segundo Carreirão (2000), os trabalhos de previsibilidade eleitoral se revolucionaram, já que essas pesquisas passaram a coletar informações vitais dos indivíduos, em massa e com maior periodicidade. Delas, pode-se encontrar, com maior precisão, as diferentes correlações entre variáveis “internas” de um eleitor e sua decisão do voto. Entretanto, elas vêm se mostrado muito imprecisas quanto à capacidade de previsão, conforme apresentado anteriormente.

Por fim, ainda segundo Carreirão (2000), a Ciência Política ainda não realizou estudos que buscassem analisar os efeitos das redes sociais no direcionamento do voto. Como, até então, elas não vinham sendo muito usadas pelo eleitorado brasileiro,

isso pode ser justificado. Entretanto, segundo o Datafolha (2018), para as eleições de 2018, uma grande maioria dos eleitores não apenas utilizam as redes sociais, como também tem, nelas, as principais fontes de informação.

Percebe-se, portanto, que até então, os diferentes estudos feitos na Ciência Política a respeito da previsibilidade eleitoral apresentam alguns *gaps*: ou não conseguem tirar conclusões precisas a respeito do indivíduo ou têm se mostrado imprecisos quanto às previsões, além de relevarem o impacto de redes sociais. Com isso, apresenta-se uma oportunidade de melhoria através da aplicação de técnicas da Engenharia, como *data mining*. De um ponto de vista pessoal, essa temática foi de acordo com o desejado pelo autor, já que englobava assuntos de interesse pessoal (política e eleições) com uma área de conhecimento do curso não antes aprendida (*data mining* e análise de sentimentos).

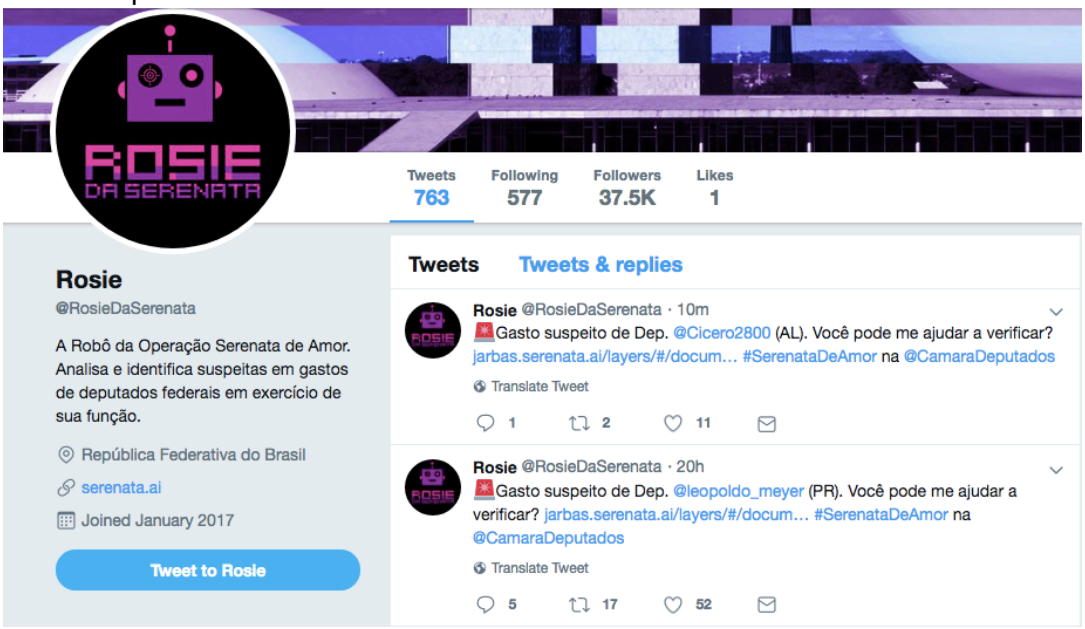
Porém, existem alguns exemplos de projetos que aplicam técnicas de Informática ou de Engenharia na área política, como na fiscalização de gastos de Parlamentares, na criação de indicadores para campanhas eleitorais ou até na simulação de comportamentos de eleitores. Para o presente trabalho, esses *cases* serviram de exemplo ou inspiração. O primeiro exemplo encontrado foi a *Cappra Data Science*, empresa de Porto Alegre que trabalha com projetos, pesquisa e consultoria em Ciência de Dados. O que chamou a atenção nela foi seu fundador, Ricardo Cappra, que trabalhou para o Governo de Barack Obama, nos EUA, bem como para as campanhas do Democrata em 2008 e 2012. Em entrevista à revista *Veja* (2012), ele explicou como aplicava a ciência de dados em seu trabalho no período eleitoral:

> Entendíamos o que estava acontecendo em uma região e então sugeríamos uma ação para os estrategistas da campanha. Tudo isso, vale ressaltar, em tempo real. Cerca de 15 dias antes do primeiro debate do Obama, já sabíamos que seu desempenho não seria satisfatório. Dessa forma, planejamos uma ação on-line, que foi ao ar uma hora depois do programa. (CAPPRA, 2012)

Segundo ele, foram usados apenas dados públicos e ferramentas *open source*. Com essa pesquisa, conseguiram prever quais eleitores decidiram as eleições e, assim, direcionavam as propagandas a eles. Ao final, previram o resultado das eleições com 30 dias de antecedência e 96% de precisão, segundo Cappra (2012), que também afirmou: "*a estratégia digital política no Brasil é ridícula*".

Outro projeto inspirador, também liderado por brasileiros, é a chamada *Operação Serenata de Amor*, que é *"um projeto aberto que usa ciência de dados para fiscalizar gastos públicos e compartilhar as informações de forma acessível a qualquer pessoa*", segundo o próprio *site* da organização. Para isso, criaram duas iniciativas: a Rosie e o Jarbas. A Rosie é um robô que analisa os reembolsos feitos a Deputados e Senadores durante exercício de seus mandatos, identifica gastos suspeitos e publica em seu *Twitter* quando isso ocorre. A *timeline* da Rosie pode ser vista na Figura 2.1. Nela, observa-se diferentes observações de suspeitas publicadas pela ferramenta.

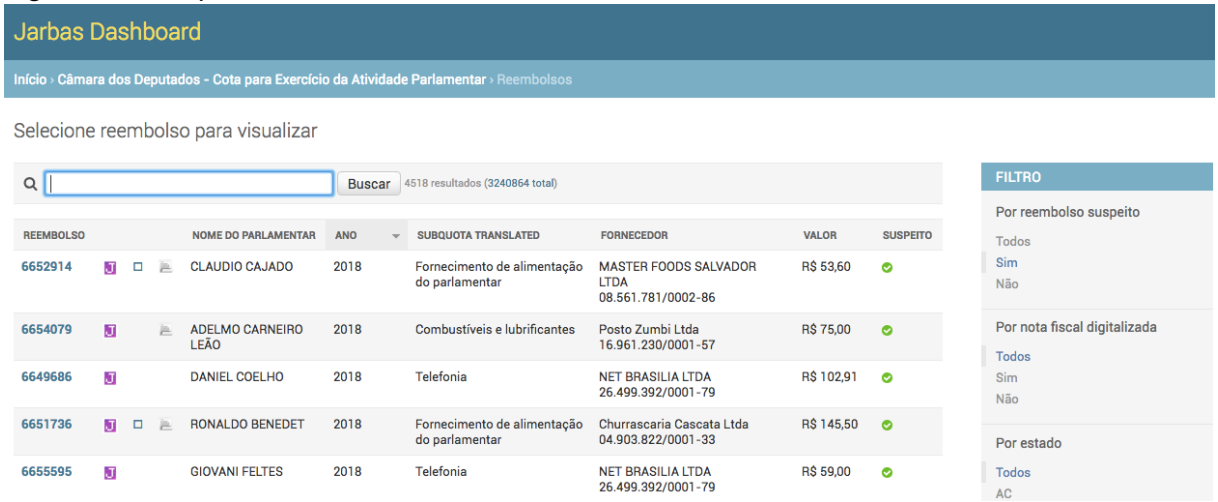
Figura 2.1 - Captura de tela da *timeline* do robô Rosie



Fonte: https://twitter.com/rosiedaserenata

O Jarbas, por sua vez, é um *site* que apresenta em detalhes os motivos da suspeita identificada pela Rosie, de forma que qualquer um consiga iniciar uma investigação mais detalhada. Na Figura 2.2, está apresentada a tela principal do *site*, onde, novamente, observa-se as diferentes anomalias que o usuário pode selecionar para buscar mais informações.

Figura 2.2 - Captura de tela do *dashboard* do robô Jarbas

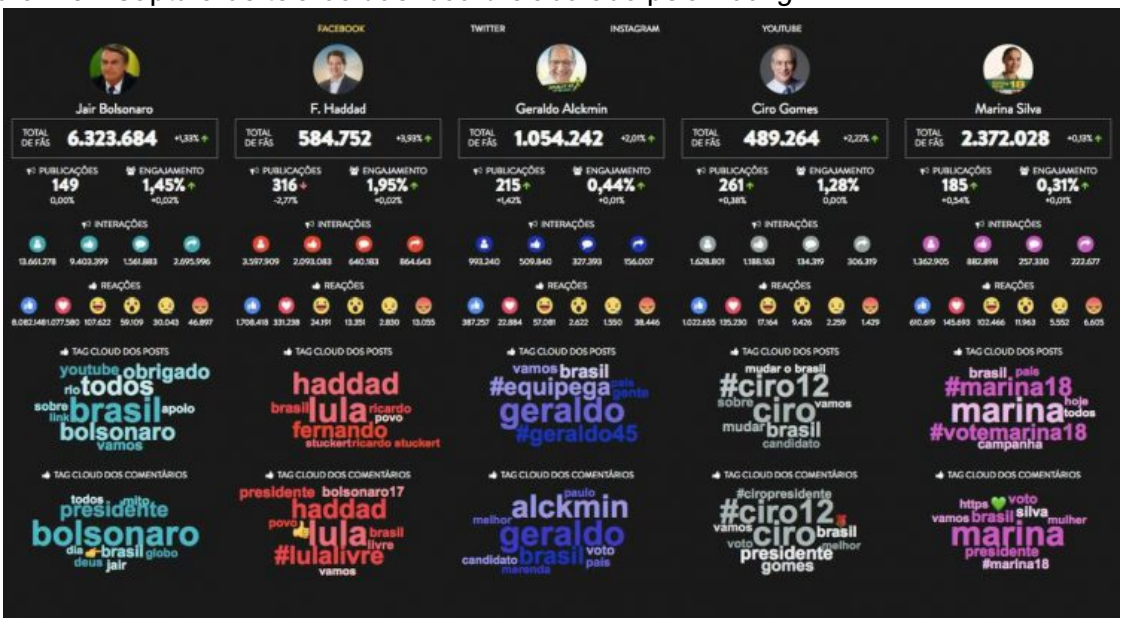


Fonte: https://jarbas.serenata.ai/

Um terceiro *case* foi elaborado pela *Zeeng* (2018), empresa brasileira que trabalha com análise de dados para comunicação e marketing. Para as eleições presidenciais desse ano, ela elaborou uma *dashboard* (ferramenta de visualização de

informações, mostrada na Figura 2.3) *online* com alguns indicadores das redes sociais dos candidatos. Nela, observa-se, em cada coluna (no total, são cinco) indicadores das paginas no *Facebook* de um candidato, como quantidade de curtidas e palavras mais ditas em suas publicações.

Figura 2.3 - Captura de tela do *dashboard* elaborado pela *Zeeng*



Fonte: Zeeng (2018)

Entretanto, ela coleta apenas indicadores simples, como quantidade total de curtidas ou seguidores, que são de fácil coleta. Isso impede (ou, no mínimo, dificulta) a realização de análises mais completas a respeito do uso das redes pelos postulantes.

Dentre as iniciativas acadêmicas, um projeto que aplicou técnicas de Informática, um dos pilares da Engenharia de Automação, em política foi um artigo de Eisenberg e Vale (2009). Ele desenvolveu junto ao Centro de Políticas Públicas e Avaliação da Educação da Universidade um simulador de comportamento eleitoral. Esse projeto tomou como base para simulação as interações sociais dos indivíduos com a mídia, colegas de trabalho e/ou estudo, amigos e familiares.

O simulador foi feito partindo de uma quantidade inicial de eleitores entrevistados na cidade do Rio de Janeiro, que informaram seus candidatos de preferência à Presidência. O simulador, então, procurou prever como as suas opiniões mudariam conforme fossem interagindo com outros eleitores e com a mídia ao longo de sua rotina. Nesse caso, cada interação era considerada como um evento discreto e que necessariamente haveria discussão a respeito de política, com um lado podendo afetar a visão do outro a respeito de seu candidato. Essas interações ocorriam de forma sistemática ao longo do dia-a-dia: de manhã cedo, ao ler o jornal e/ou assistir televisão; no trabalho/escola, ao conversar com colegas; em atividades de lazer, ao conversar com amigos, *etc*.

Ao comparar o resultado previsto pelo simulador com a realidade, não se obteve uma precisão muito boa. Para as eleições de 2002, no Rio, o candidato Lula (PT) foi o mais votado com 47% dos votos. O simulador, por sua vez, só previu sua vitória em 26,25% dos cenários: nesses casos, ele obteve uma média de 43% dos votos. Entretanto, o próprio autor aponta melhorias. A principal delas, segundo Eisenberg e Vale (2009), é que o projeto não conseguiu modelar o acontecimento de "escândalos" ao longo da campanha, como descobertas de casos de corrupção. Além disso, o simulador não leva em conta as interações feitas em redes sociais nem em aplicativos de mensagem, algo de extrema relevância nos dias atuais.

2.3 Descrição do problema

Levando em conta o contexto atual descrito acima, o presente trabalho considera o seguinte cenário:

- Imprevisões eleitorais em eleições de forte impacto global em 2016, com os casos do Reino Unido e dos Estados Unidos;
- Imprevisão quanto às eleições presidenciais no Brasil, que indicavam, inicialmente, uma disputa pulverizada e até com incerteza a respeito da elegibilidade de um candidato;
- Alterações nas estratégias de campanha, graças às mudanças causadas pela Reforma Político-Eleitoral, que reduziram o período das campanhas e limitaram seu financiamento;
- Pesquisas de previsibilidade eleitoral, feitas na Ciência Política, com falhas reconhecidas em suas metodologias de coletas de dados;
- Avanço de uso das redes sociais, tanto pelo eleitorado, que a utiliza como fonte de notícias, até pelas campanhas eleitorais, graças às restrições impostas pela Reforma Político-Eleitoral de 2017. O impacto das redes, entretanto, até então ainda não tinha sido mensurado na literatura da Ciência Política - o que constata o pioneirismo deste trabalho.

O problema tratado no presente trabalho é propor novos métodos para o aprimoramento da previsibilidade eleitoral através de dados coletados na rede social *Twitter*. Com base na teoria de estratégias emocionais da Neuropolítica, o objetivo é utilizar técnicas de *data mining* para propor novos métodos de previsão eleitoral. O próximo capítulo apresenta os principais conceitos de Ciência Política utilizados ao longo do trabalho.

3 FUNDAMENTOS DA CIÊNCIA POLÍTICA

Dentro da Ciência Política, foi necessário compreender as diferentes abordagens teóricas a respeito da decisão de voto existentes na literatura. Existem três que são tradicionais e são as mais aceitas e citadas e todas serão aqui apresentadas. Entretanto, antes disso, será feita uma introdução à Ciência Política e seus princípios na Seção 3.1. Já na Seção 3.2, será explicada a abordagem sociológica, organizada por pesquisadores da Universidade de Columbia. Em seguida, na Seção 3.3, a abordagem psicológica ou psicossociológica, proveniente da Universidade de Michigan será apresentada. Na sequência, a teoria da escolha racional (também chamada de Modelo Downsiano) será exibida na Seção 3.4. Além delas, uma quarta, que tem ganhado respaldo mais recente na área, será explicada na Seção 3.5: a neuropolítica.

3.1 Ciência Política: abordagens teóricas sobre o comportamento eleitoral

Para modelar um sistema que proponha resolver um certo problema em qualquer domínio de conhecimento, exige-se, por parque de quem conceberá o projeto, um aprofundamento no tema para, assim, poder modelar as soluções. Portanto, apesar do presente trabalho estar inserido no curso de Engenharia de Controle e Automação, foi necessário adquirir considerável conhecimento a respeito de alguns temas da Ciência Política, os quais serão detalhados a seguir.

Os fenômenos sociais são os objetos de estudo da Ciência Social, ou Sociologia. Segundo Tckeskiss (1934), "*cada fenômeno social tem de estar ligado aos esforços que forjam a sociedade*", neste caso, somos nós, seres humanos que os fazem. Portanto, a Sociologia procura estudar as diferentes ações dos homens, enquanto conjunto (em grupos, organizações, sociedades), e como elas estão relacionadas com sua consciência.

Um exemplo simples e que está em nosso dia-a-dia é a moda, que muda/define uma ação coletiva humana (a maneira como nos vestimos) mas que também influencia (ou é influenciada pela) nossa consciência (define valores vigentes da sociedade, é uma maneira de exprimir nossa personalidade, nossos ideais). Sendo assim, um fenômeno social é, portanto, psicológico. Entretanto, um fenômeno também se relaciona com o organismo humano como um todo. Afinal, fatores fisiológicos e químicos também afetam não apenas o processo de decisão humano, mas também a maneira com que nos desenvolvemos. Portanto, fenômenos sociais são de grande complexidade, que se relacionam com uma imensidão de variáveis e que, ao contrário da engenharia, possuem altos graus de inexatidão (TCKESKISS, 1934).

Dentro da Ciência Política, que é uma das áreas da Sociologia, há um fenômeno social fascinante e que também faz parte de nosso dia-a-dia: o voto. Para uma sociedade se engajar em um processo eleitoral, são necessários milhões de indivíduos (no caso do Brasil) tomarem a decisão de participar do evento (através do ato de votar) e, não apenas isso, mas também direcionar o voto a alguém. Entretanto, para os cientistas sociais, não basta analisar um voto apenas *a posteriori*: boa parte

dos estudos e bibliografias na área estão voltadas à previsibilidade eleitoral (PIMENTEL, 2006).

Essa área da Ciência Política ainda é relativamente recente, segundo Figueiredo (2008). Ela teve início há aproximadamente sete décadas, com o advento das pesquisas de opinião e dos sistemas de processamento de dados. Desde então, os cientistas sociais estudam este fenômeno social nos mais diversos contextos sociais e políticos, analisando diferentes fontes e aplicando diferentes metodologias. Normalmente, os diferentes modelos explicativos feitos nas pesquisas científicas procuram responder à seguinte pergunta genérica: *por que as pessoas vão votar e por que dão o seu voto para este ou aquele candidato ou partido?*

Dentre os diversos modelos concebidos ao longo desses anos, são três os que mais se destacam na literatura mundial, bem como no Brasil:

- **Teoria sociológica**, periodicamente difundida na literatura como "modelo de Columbia", em alusão aos pesquisadores dessa Universidade, nos Estados Unidos, que desenvolveram esse modelo em *The People's Choice*, de 1944 e *Voting*, de 1954;
- **Teoria psicológica ou psicossociológica** do comportamento eleitoral, também chamada na literatura como "modelo Michigan", fazendo referência aos pesquisadores da Universidade de Michigan, também nos EUA, que o elaboraram com o livro *The American Voter*, publicado em 1960;
- **Teoria da escolha racional**, que é comumente chamada de "modelo Downsiano", em homenagem ao seu criador, Anthony Downs, que o fez em *An Economic Theory of Democracy*, em 1957.

Para responder à pergunta supracitada, é preciso entender que o processo eleitoral vai além do ato de votar e da contabilização dos resultados. Antes de tudo isso, há um longo processo social por que uma população passa para escolher os projetos sociais a serem desenvolvidos nos próximos anos. Os diferentes modelos concebidos se diferem justamente ao tentar reconstruir o processo que levou ao resultado final, ou seja, determinar a fonte originária que leva à decisão. Entretanto, todos se assemelham ao reconhecerem as particularidades de cada eleitor: cada indivíduo, ao longo de sua vida, adquire seu próprio sistema de crenças, desejos, valores, ódios (o que, eventualmente, transforma-se em sua ideologia). Ao expressar-se em uma democracia através de seu voto, toda a sua história social contribui para sua decisão (FIGUEIREDO, 2008):

> Explicar o voto (...) é o mesmo que revelar como variadas motivações e percepções se agregam na opção por um candidato. (...) Neste sentido, adquire particular relevo o estudo das semelhanças e diferenças no comportamento de distintas camadas sociais, pois é na acentuação ou na neutralização das propensões destas que se encontra quase sempre o cerne de uma estratégia eleitoral (*apud* Lamounier, 1978).

As diferentes teorias serão apresentas a seguir, conforme ordem cronológica de sua criação, iniciando pela Teoria Sociológica.

3.2 Teoria Sociológica ou Modelo de Columbia

Os primeiros trabalhos na área de previsibilidade eleitoral ocorreram nos Estados Unidos, na década de 40, por conta de pesquisadores da Universidade de Columbia, com liderança de Paul Lazarsfeld (1940). A proposta inicial dos colaboradores era de mensurar o impacto que as chamadas “forças de curto prazo” teriam nas decisões eleitorais (LAZARSFELD; BERELSON; GAUDET, 1940). Na época, essas forças seriam causadas pelas mídias de massa vigentes: a televisão começava a surgir na Europa e nos EUA, mas ainda era um eletrodoméstico muito elitista; o rádio, por sua vez, já havia se popularizado nesses países, graças aos anseios da população em acompanhar eventos da Segunda Guerra Mundial com menor atraso; e a mídia impressa também circulava com alta presença no dia-a-dia de sua população (FIGUEIREDO, 2008).

Sendo assim, os pesquisadores procuraram entender como as propagandas políticas veiculadas nesses meios em época de campanha influenciavam a direção do voto das pessoas. Optaram por realizar um teste em 1940 no condado de Erie, no estado de Ohio, nos Estados Unidos. Foram selecionados de forma aleatória quatro grupos de 600 eleitores cada do condado - três desses conjuntos foram entrevistados duas vezes (no começo e ao final da pesquisa) e o quarto, mensalmente, entre maio e novembro daquele ano. As entrevistas questionavam não apenas quanto ao voto do sujeito nas eleições presidenciais daquele ano (entre o Democrata Franklin Roosevelt, que buscava reeleição, e o Republicano Wendell Willkie), mas também os seus posicionamentos em questões polêmicas, como a venda de munição pelos EUA aos Aliados, confiabilidade na mídia, dentre outros (LAZARSFELD; BERELSON; GAUDET, 1940).

Os resultados das entrevistas, entretanto, apresentaram algo inesperado pelos pesquisadores: poucos eleitores alteraram seus votos ou suas visões políticas em geral ao longo do período de campanhas. Portanto, puderam concluir que as chamadas “forças de curto prazo” exercidas pelas mídias de massa vigentes tinham pouco impacto na estruturação do voto. Com base nesses resultados, chegaram à conclusão de que as campanhas eleitorais não persuadiam a população e que o eleitorado já teria uma “predisposição”, segundo os próprios colaboradores de Columbia, a respeito das propostas a serem apresentadas (LAVAREDA, 2011). Mas se as campanhas não têm forte influência, o que então direciona o eleitor?

Apesar de praticamente não cambiarem os votos da amostra, as campanhas tiveram um papel fundamental no comportamento eleitoral. Segundo Lazarsfeld, Berelson e Gaudet (1940), elas “ativaram” ou reforçaram as predisposições mencionadas acima: para os eleitores indecisos, a tendência foi que, com o decorrer das campanhas, eles orientassem seus votos às suas predisposições. Além disso, para aqueles que já tinham um posicionamento estabelecido antes do período de estudo, as forças de curto prazo tendiam a reforçar suas convicções - mas não modificar.

Portanto, a partir dos achados da pesquisa, a chamada teoria sociológica foi estabelecida. Seu princípio era que a previsibilidade eleitoral pode ser feita através de

uma análise macro, dividindo o eleitorado em diversos grupos conforme seu contexto social. Isso significa que os comportamentos e as escolhas são determinados pelas condições socioeconômicas, culturais, educacionais, religiosas e urbanas do indivíduo (LAVAREDA, 2011). Determinar quais condições possuem mais influência requer uma análise específica da população, já que podem variar conforme o local e o período de estudo:

> A clivagem social, a diferença entre ricos e pobres, o *status*, a renda, o pertencimento aos sindicatos de trabalhadores, a religião, tudo gerava impacto sobre a decisão sobre a posição política a adotar (MARTINS JUNIOR, 2009).

No caso do estudo de Lazarsfeld, Berelson e Gaudet (1940), no condado de Erie, as três características mais relevantes para formação dos grupos foram a classe socioeconômica, a religião, o gênero e o tipo urbano de sua residência (se era urbano ou rural). Ele e seus colegas puderam comprovar que um eleitor homem, católico, de classe econômica mais baixa e que residia em um ambiente urbano tendia a estar alinhado com os ideais do Partido Democrata e, por consequência, votar no Roosevelt nas eleições presidenciais. Dessa análise, os pesquisadores de Columbia puderam chegar à outra conclusão importante: a identidade política (ou seja, o conjunto de princípios, valores e ideologias) de um grupo social tende a convergir em votos ao partido que defenda essas crenças do grupo.

Dentre esses requisitados, o principal é a existência de partidos que de fato possuam uma clara identidade política, conseguindo assim se identificar ideologicamente com algum grupo social. Além disso, é preciso que os eleitores tenham um certo nível de consciência para reconhecer o grupo em que estão, chamada de “identidade interna”. Por fim, é necessário que os grupos sociais consigam identificar os partidos que realmente se identificam com sua identidade - algo que depende dos partidos (comunicarem-se de forma efetiva e honesta com a população) e também do eleitorado (buscar informações a respeito dos partidos) (MARTINS JUNIOR, 2009).

Desde o estudo feito por Lazarsfeld e demais colaboradores da Universidade de Columbia, outros cientistas sociais e/ou políticos replicaram a mesma análise em suas regiões e em diferentes décadas. Na Grã-Bretanha, conforme aponta Martins Junior (2009), já na década de 60, a mesma conclusão foi obtida com duas pesquisas diferentes: uma realizada com eleitores da cidade de Bristol e outra, com os de Greenwich. Lá, as características prevalentes foram, novamente, gênero, religião e classe social. A principal diferença entre essas pesquisas britânicas em comparação com a de Columbia é que, enquanto nos EUA as eleições analisadas foram as presidenciais, no Reino Unido, o pleito estudado foi para o cargo de deputado.

Já no Brasil, esse tipo de estudo foi feito pelas primeiras vezes entre as décadas de 50 e 70, contendo dados eleitorais e censitários agregados. As variáveis que mais pesaram na decisão do eleitorado, nessa época, foram índices de urbanização e industrialização. Depois disso, houve um foco de análise na disputa polarizada entre dois partidos, ocorrida, pelo menos, nas últimas 5 eleições presidenciais brasileiras (MARTINS JUNIOR, 2009). Tanto Terron & Soares (2010)

quanto Singer (2012) chegaram à mesma conclusão: a tendência do eleitorado de classe social mais baixa era de votar em um partido, enquanto aqueles de classe mais elevada, tendiam a optar pelo outro.

Há, contudo, uma grande ressalva na teoria sociológica e que faz com que ela não seja inteiramente aceita entre pesquisadores das ciências sociais. A racionalidade (termo usado para definir a capacidade de reconhecer suas próprias ideologias, bem como identificar as dos partidos/candidatos) necessária dos cidadãos recebeu um forte "golpe" em uma nova leva de pesquisas de Lazarsfeld: grande parte do eleitorado não possuía informações sobre os detalhes de propostas ou sobre os posicionamentos dos candidatos. Assim, estava formado um paradoxo: embora boa parte da população não possuía racionalidade, votavam como se tivessem, e foi nesse contexto que a teoria descrita abaixo surgiu (ACHEN; BARTELS, 2016).

3.3 Teoria Psicológica ou Psicossociológica

A teoria psicológica de decisão eleitoral surgiu na Universidade de Michigan, nos EUA, na década de 60, como um complemento à escola de Columbia. O objetivo principal de Campbell e Converse, seus idealizadores, era de "*preencher o vazio decorrente da constatação de um público de massas desinformado sobre as questões em tela nas disputas eleitorais*" (LAVAREDA, 2011). Segundo eles, partindo do paradoxo do modelo de Columbia, a maioria do eleitorado é errática na conceituação do mundo político, formando sistemas idiossincráticos, com atitudes com sinais trocados. Afirmam também que apenas uma porção altamente politizada da sociedade (15% da população em países desenvolvidos, estimam eles) é coerente (ACHEN; BARTELS, 2016).

De início, os pesquisadores de Michigan reconheceram a importância do contexto social no direcionamento do voto. Entretanto, buscaram demonstrar, segundo Pimentel (2007) que "*existem fatores intervenientes de natureza individual que explicam melhor a decisão do voto do que as predisposições sociais*". Sendo assim, o modelo de Michigan tem como prioridades o indivíduo, suas crenças, suas atitudes e suas motivações psicológicas como unidade de análise - o que tem relativa independência do contexto social do sujeito.

O trabalho foi feito por meio de diversas pesquisas empíricas e questionários (*surveys*), através dos quais centenas de dados foram gerados. Com isso, os cientistas introduziram dois conceitos fundamentais para compreensão do modelo: sistema de crenças e identificação partidária (FIGUEIREDO, 2008).

Philip Converse definiu um sistema de crenças, em seu livro "*The nature of belief systems in mass publics"*, como um conjunto de atitudes que o indivíduo tem junto ao sistema político (como por exemplo, defender um candidato em seu círculo social). Quanto a esse sistema, segundo Sabin (2018), o que ele notou foi que a vasta maioria do eleitorado tem pouco entendimento de suas crenças, não tem base para justifica-las e possuem, muitas vezes, crenças conflitantes (como por exemplo, segundo o pesquisador, apoiar redução da carga tributária e expansão de programas sociais). A exceção está presente apenas em uma minoria intelectual ou politizada,

conforme já mencionado anteriormente. Os números exatos de sua pesquisa dizem que:

> Apenas 3,5% possuíam um sistema de crenças razoavelmente abstrato e ideológico; 12% faziam uso desses conceitos, mas não entendiam exatamente seu significado e 84,5% eram totalmente estranhos a esses termos ideológicos. Assim, concluiu Converse (1964), o sistema de crenças do país era instável e desprovido de coerência (LAVAREDA, 2011).

Essa alta porcentagem de eleitores que eram "*totalmente estranhos a termos ideológicos*" motivou a definição de outro termo fundamental para o modelo de Michigan: a alienação política. Criado por Robert Lane em 1962, o termo significa uma rejeição consciente de todo o sistema político por meio de apatia (SABIN, 2018) e é composta por três atitudes do indivíduo - algumas delas, certamente, já foram ditas por quase todo o eleitorado:

- Não sou sujeito, mas sim objeto da política (não tenho influência);
- O governo não administra no meu interesse; e
- Não aprovo o processo de tomada de decisões, as regras são injustas e a Constituição pode até ser fraudulenta.

O termo "identificação partidária", por sua vez, foi criado para explicar o direcionamento de voto do eleitorado sem um sistema de crenças racional. Segundo Figueiredo *apud* Campbell (2008), ele representa os laços afetivos dos eleitores em relação aos partidos. Assim, as legendas tornam-se as referências para o publico. O que os pesquisadores observaram foi que havia uma estabilidade partidária entre eleições nos EUA, mesmo se um partido apresentasse outro candidato ou até se alterasse pontos de suas propostas.

Partindo das conclusões que o eleitorado possui alta identificação partidária, o que resulta em um sistema de crenças pouco racional (em média) ou até em uma alta alienação política por parte dos eleitores, resta entender o que constrói esse apego aos partidos nos indivíduos.

Segundo o modelo de Michigan, a identificação é criada ao longo do período de aprendizado do eleitor, com maior influência do ambiente familiar, cultural e histórico em que está inserido. Ou seja, é durante a socialização com outros indivíduos a sua volta, ainda na infância e principalmente no âmbito familiar, que a perspectiva política é formada - e é ao continuar socializando com aqueles em seu círculo social que o eleitor pode ser influenciado ou influenciar outros a modificarem sua identificação partidária (ou então mudar suas atitudes em relação à ela) (FIGUEIREDO, 2018).

Dito tudo isso, de forma resumida, o modelo de Michigan pode ser resumido ao afirmar que o comportamento político de um sujeito é resultado de interações sociais ao longo de sua vida com outros sujeitos (e suas atitudes) de seu círculo social, principalmente familiares. A teoria, entretanto, tem suas críticas.

Já a partir da década de 70, a proposta iniciada por Campbell e Converse começou a ser questionada. Vários estudos da época, como os de Nie, Verba e Petrocik, em 1976, comprovaram um aumento do número de eleitores mais informados e uma queda do percentual daqueles que se identificavam com um partido específico. Segundo os cientistas, essa mudança foi causada pelo aumento do impacto de temas políticos na vida pessoal dos eleitores e, por isso, mais pessoas passaram a prestar mais atenção nos debates ideológicos entre as elites (PIMENTEL, 2007).

Porém, essa própria afirmação de que o eleitorado vem se tornado mais sofisticado politicamente também não é unânime no campo da Ciência Política. Entretanto, o que é concordância de todos é a redução do apego aos partidos - inclusive no Brasil (MARTINS JUNIOR, 2009). O declínio do partidarismo gerou o principal paradigma do Modelo de Michigan e que introduziu o conceito de racionalidade econômica do voto (LAVAREDA, 2011), a ser vista na próxima seção do documento.

3.4 Teoria Racional

Idealizada por Anthony Downs na segunda metade da década de 50, a teoria da escolha racional (também conhecida como modelo downsiano) afirma que o comportamento eleitoral ocorre conforme um auto-interesse, em que o indivíduo é um "ator racional maximizador utilitarista", ou seja, ele direciona seu voto para maximizar seus próprios ganhos e reduzir seus custos (LAVAREDA, 2011).

Segundo Downs, a lógica do voto do eleitor é baseada na premissa de que, diante de um conjunto de escolhas possíveis (os candidatos, seus partidos e seus planos de governo), ele torna-se um ator racional e escolhe pela opção que lhe trará mais benefícios, por meio das atividades governamentais propostas. De certa forma, esse comportamento é similar ao de um consumidor no âmbito do mercado - referenciado como *homo economicus* por economistas, a teoria da escolha racional irá adaptar este conceito para criar o *homo politicus* (MARTINS JUNIOR, 2009).

O "homem político" por meio de um diferencial de utilidade esperada dos partidos. Ou seja, ele compara o desempenho do partido que está atualmente no governo com os benefícios imaginados caso a oposição estivesse em seu lugar. Para fazer essa análise, Morris Fiorina, já na década de 80, propôs que a avaliação retrospectiva é mais econômica do que a prospectiva (LAVAREDA, 2011), ou seja, é mais "barato" (mais fácil para o eleitor, portanto menos custoso para ele) analisar o que já foi feito do que propostas e ideias. Ao avaliar o passado, o indivíduo consegue projetar o futuro: é o voto prospectivo com base na avaliação retrospectiva. Com isso, segundo Fiorina *apud* Figueiredo (2008), "*a lealdade e as identificações políticas/partidárias não resistem ao teste dos fatos*".

De forma resumida e simplificada, o eleitorado se posiciona como um juiz do governo: se a economia vai bem, os governantes ganham mais votos; se não, a oposição se favorece. As ideologias, a identificação partidária e os valores dão espaço a um sistema de interesses.

Ainda segundo a escola racional, entretanto, o eleitorado pode ter dois sistemas de interesses diferentes, dividindo os indivíduos em egoístas e sociotrópicos. Enquanto o primeiro grupo tem como interesse seus interesses pessoais, sua vida doméstica, seu próprio bolso (baseia-se na pergunta: *what have you done for me lately?*), o segundo possui como escala de comparação o estado econômico da sociedade como um todo, votando de acordo com o "bolso do pais" (baseia-se na pergunta: *what have you done for the country lately?*). Para desempenhar essas avaliações de desempenho do governo, os pesquisadores desta teoria afirmam que os principais indicadores analisados são: inflação, taxa de desemprego, crescimento real da renda e a seguridade social (FIGUEIREDO, 2008).

Este modelo, contudo, é válido apenas com restrições, além de ser alvo de diversas críticas. A primeira reprimenda feita por cientistas políticos é quanto à capacidade do eleitor ser de fato racional. Argumenta-se que a grande maioria não tem acesso às devidas informações para uma decisão racional, conforme comprovado pelos pesquisadores da escola de Michigan. Os racionalistas, entretanto, argumentam que o acesso ao conhecimento não é necessário para uma decisão racional e trouxeram o conceito de "níveis de racionalidade". Segundo eles, mesmo sem serem muito informados nem terem coerência ideológica, os eleitores não votam de moto irracional (ou seja, não realizam um voto emocional ou algo similar) - apenas possuem algumas limitações (LAVAREDA, 2011).

Outra forte crítica à teoria é quanto à ambiguidade dos partidos. O que pesquisadores afirmam é que, na tentativa de angariar mais votos, os partidos "jogam o jogo da ambiguidade", ou seja, tornam-se flexíveis em relação às suas ideologias e posicionam-se de forma que eleitores de diferentes perspectivas possam se identificar com suas propostas. Nas sociedades atuais, em que temos diversas questões políticas e sociais entrelaçadas, o relaxamento da rigidez ideológica é fundamental para conquistar sucesso eleitoral. Com isso, deparamo-nos com uma grande imprevisibilidade quanto aos partidos e, consequentemente, o eleitorado se volatiza: os candidatos convergem para um centro ambíguo e, assim, crescem as chances de decisões aleatórias e de abstenções entre os indiferentes (FIGUEIREDO, 2008).

Quanto às restrições do modelo, destaca-se, inicialmente que ele só é valido (com sua aplicação através do calculo retrospectivo) para eleições redutíveis a no máximo dois candidatos ou dois "blocos de candidatos" diferentes: situacionistas e oposicionistas. Com três ou mais opções, a teoria apresenta-se limitada, já que boa parte das pesquisas feitas a respeito dela foram nos EUA e na Inglaterra, onde predominam sistemas bipartidários (FIGUEIREDO, 2008). Aqui no Brasil, portanto, ela tem sido mais aplicada em situações de segundo turno (CARREIRÃO, 2000).

Além disso, uma dificuldade imposta é que a relação de satisfação com o voto não é linear, ou seja, segundo Figueiredo (2008), "*os partidos do governo são muito menos recompensados pelos bons tempos do que castigados pelos maus tempos*". Por fim, algo apontado por Fiorina, é que nada impede um eleitor de avaliar positivamente os atuais governantes, mas preferir optar, racionalmente, pela oposição - bem como o contrário.

Portanto, mesmo em um processo de decisão eleitoral que se enquadre no modelo racional (bipartidarismo, com eleitores conscientes e bem informados e partidos não-ambíguos), questiona-se até que ponto essa escolha não possui aspectos emocionais envolvidos. Foi a partir desse questionamento, aliado ao surgimento de novas técnicas de análise de atividades cerebrais e de estudos feitos a respeito da “cognição” (aspectos da mente, como atenção, percepção e memória) que surgiu a próxima teoria a ser apresentada: a neuropolítica (LAVAREDA, 2011).

3.5 Neuropolítica

A neuropolítica é a investigação da relação entre o cérebro e a política. É uma área multidisciplinar, sendo a intersecção da neurociência com ciência política, psicologia, genética comportamental, dentre outras. De uma maneira geral, a neuropolítica aplica métodos e técnicas da neurociência cognitiva para solucionar a problemas relevantes clássicos da ciência política, inclusive decifrar como os indivíduos direcionam seus votos, como formam suas ideologias, dentre outros (LAVAREDA, 2011).

Um conceito que é bom definir desde o início e que é premissa fundamental da neuropolítica é o de cognição. Enquanto as áreas mais tradicionais da psicologia e da ciência política definem-na como o processo consciente para a realização de atividades como o exercício da memória, o raciocínio, a atenção, dentre outras, as abordagens mais recentes já abrem mão do requisito da consciência e passam a focar naquilo que não percebemos nem temos controle (LAVAREDA, 2011).

Apesar de terem seu valor para registrar o direcionamento de um voto, segundo Law e Redlawsk *apud* Lavareda (2011), os *surveys* não são o suficiente para indicar como o individuo chegou à decisão. Afinal, segundo Pradeep *apud* Lavareda (2011), nosso cérebro consciente só consegue processar no máximo 40 bits de informação por segundo, enquanto nossos sentidos captam 11 milhões de bits por segundo. Ou seja, aproximadamente 99% do conteúdo que “recebemos” são processados de forma inconsciente, por isso a nova definição de cognição.

A neuropolítica procura focar, portanto, justamente naquilo que nós, enquanto eleitores, não temos percepção: o inconsciente e seus efeitos nos comportamentos políticos.

O reconhecimento do papel das emoções e da dimensão do inconsciente no papel das decisões passou a ser estudado no âmbito da ciência política apenas na década de 90. Logo de início, seu objetivo era constatar os limites dos métodos chamados “declarativos” de compreensão do direcionamento político (LAVAREDA, 2011). Desde então, dezenas de teorias foram elaboradas dentro da grande área da neuropolítica. Será apresentada, a seguir, uma das com maior relevância na literatura (LAVAREDA, 2011) e que analisa os discursos dos candidatos durante suas campanhas:

3.5.1 Estratégias emocionais

Essa abordagem, como dito acima, procura avaliar os discursos das diferentes campanhas eleitorais e os efeitos causados pelos sentimentos por elas transmitidos. Um dos primeiros cientistas a fazê-lo foi Ted Brader, em 2006, sendo o primeiro a trazer compreensão teórica e testes em laboratório a respeito dos apelos emocionais. Nesse caso, Brader procurou avaliar as diferentes propagandas (também chamadas de peças ou *spots*) exibidas pelos candidatos na televisão, quais sentimentos elas transmitiam e quais eram seus impactos na corrida eleitoral (LAVAREDA, 2011).

Brader chegou a uma conclusão principal que confirmava a chamada teoria da inteligência afetiva. Segundo ele, enquanto peças que transmitiam entusiasmo apenas reforçaram a lealdade dos eleitores àquele partido/candidato, reforçando um sentimento de polarização, aquelas que transmitiam medo incentivaram o indivíduo a buscar novas informações e a reconsiderar suas escolhas. Com isso, Brader pôde concluir que o medo acabou sendo mais persuasivo do que o entusiasmo ao tentar causar mudanças na direção do voto do telespectador (LAVAREDA, 2011).

Entretanto, umas das conclusões mais interessantes de seu estudo contrariou uma suposição usual. Ao contrário do que era estimado por especialistas (inclusive os pesquisadores da Universidade de Michigan), os eleitores que mais se mostraram suscetíveis à chamada “manipulação emocional” eram os mais bem informados e com maior escolaridade. Na verdade, isso ocorre, pois, esses eleitores tendem a ter maior interesse em saber mais e, com isso, pesquisam e vão atrás de novas informações (LAVAREDA, 2011).

No Brasil, Lavareda (2011) repetiu a mesma análise de Brader, mas em 2009 e usando como foco as eleições presidenciais de 1998, 2002 e 2006. De imediato, ele notou que a tendência era de que candidatos brasileiros utilizassem com menos frequência as emoções negativas (medo, tristeza, raiva) do que os estadunidenses em suas propagandas. Na verdade, de todas as peças das três eleições brasileiras, 67% delas eram com emoções positivas; nos EUA, esse número é de 58%; no Reino Unido, 69%.

Lavareda (2011) destacou também as diferentes abordagens que brasileiros e norte-americanos costumam ter ao expressar emoções positivas. Enquanto no Brasil, nessas eleições, reinaram os *jingles*, nos EUA, as peças positivas costumavam apresentar mais conteúdo cômico. Além disso, enquanto os brasileiros preferem expor mais a compaixão como destaque positivo (algo que, segundo Lavareda, é causado pelo excesso de desigualdade aqui), o estadunidense privilegia o orgulho.

De maneira geral, enxergando apenas os *spots* brasileiros, os candidatos vencedores apelaram mais para um discurso com apelos ao entusiasmo, orgulho e compaixão. Enquanto isso, os perdedores apresentaram mais conteúdo com raiva. Vale ressaltar que cada país costuma ter sua própria “receita” vencedora e que, às vezes, varia até de eleição para eleição.

4 FUNDAMENTOS DE DESENVOLVIMENTO DE SISTEMAS E DE RECONHECIMENTO DE PADRÕES

No decorrer deste projeto, foi necessário aplicar diversos conhecimentos de Engenharia para realizar as análises desejadas. Pode-se dividir esses conceitos em duas grandes áreas: desenvolvimento de *software* e reconhecimento de padrões. A primeira será exposta na Seção 4.1, incluindo as metodologias de desenvolvimento de sistemas estudadas para o trabalho. Já na Seção 4.2, serão apresentados os tópicos de reconhecimento de padrões, como as diferentes etapas em um projeto deste tema e até os algoritmos de aprendizado de máquina utilizados neste trabalho.

4.1 Desenvolvimento de *softwares*

Área de estudo reconhecida também como desenvolvimento de sistemas, desenvolvimento de *software* significa a criação (incluindo a ideação e a implementação) de um sistema computacional, ou seja, transformar as necessidades de algum cliente (que também podem ser chamadas de requisitos do usuário) em um programa. Para fazê-lo, é necessário que o desenvolvedor (ou grupo de desenvolvedores) execute um processo de desenvolvimento de *software*, algo estudado dentro da Engenharia de Software. Um processo nada mais é do que um conjunto de atividades a serem feitas, existindo diversos métodos (ou modelos) genéricos a serem seguidos (LEITE, 2008).

Dentre as diferentes abordagens existentes para a elaboração de um sistema, apenas duas foram estudadas para este trabalho: os chamados Modelo em Cascata e Modelo em V, ambos a serem apresentados a seguir.

4.1.1 Modelo em V

O Modelo em V, criado na década de 80 e apresentado na Figura 4.1, é um dos métodos mais usados e mais importantes da Engenharia de Sistemas. Feito, inicialmente, apenas para aplicações em Sistemas de Informação, teve seu uso expandido para sistemas complexos em geral na década de 90. É uma versão modificada do Método em Cascata (que será apresentado posteriormente), mas que propõe, como o nome diz, uma estrutura com formato em V (RAMOS, 2012).

Figura 4.1 - Diagrama representativo do Modelo em V

Fonte: adaptado de Ramos (2012)

Neste formato, a parte decrescente é para a chamada **verificação** e representa o período de caracterização e definição do projeto. Em seguida, a ponta do V é a etapa de desenvolvimento (programação) do sistema em si. A parte crescente, por sua vez, é de **validação** e contém diversas rodadas de testes. Com isso, os períodos de teste passaram a ter uma importância integral no desenvolvimento do projeto (tanto quanto a definição do escopo) (RAMOS, 2012).

Conforme demonstrado na Figura 4.1, temos as etapas, em ordem:

- **Análise de requisitos;**
- **Especificações do sistema;**
- **Projeto de alto nível:** também chamada de “Concepção geral”;
- **Projeto detalhado:** também chamada de “Concepção detalhada”;
- **Implementação:** também chamada de “Programação” ou “Desenvolvimento”;
- **Testes unitários:** verificam o cumprimento dos projetos detalhados;
- **Testes de integração:** verificam o cumprimento dos projetos de alto nível;
- **Testes de sistema:** também chamada de “Testes de validação”, verificam as especificações da segunda etapa;
- **Testes de aceitação:** também chamada de “Receita”, verificam os requisitos da primeira etapa;

Um detalhe importante é que o modelo conta com verificações constantes: é sempre necessário verificar as etapas anteriores antes de se passar para a próxima - por isso, o modelo também é chamado de **Modelo de Validação e Verificação**. Quanto aos testes, o modelo ensina como fazer uso efetivo deles nos estágios iniciais do desenvolvimento - o que funciona melhor se os requisitos estão claramente definidos e se o *software* em si é relativamente simples. Para isso, “Testador” e Programador/Desenvolvedor precisam ser pessoas distintas, trabalhando de forma paralela (RAMOS, 2012).

Podemos resumir o Modelo em V às seguintes vantagens e desvantagens, segundo Ramos (2012):

- **Vantagens:**

- Progresso avança de maneira sistemática;
- Mais apropriado para projetos de pequeno e médio porte;
- Testes começam na fase de requisitos;
- Fácil de manter controle do progresso.

- **Desvantagens:**

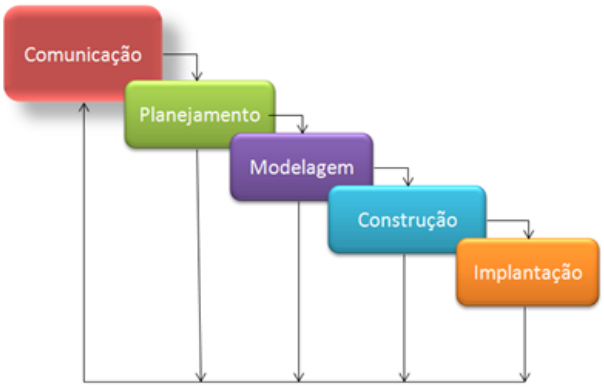
- Não apropriado para projetos grandes e complexos;
- Não é uma boa opção se requisitos mudam constantemente;
- O cliente vê apenas o projeto final e não os módulos intermediários.

4.1.2 Modelo em Cascata

Apresentado na Figura 4.2, é considerado como o modelo mais clássico e tradicional para desenvolvimento de *software*. O Modelo em Cascata propõe uma

abordagem linear, sistemática, com uma sequência bem definida de passos e atividades a serem feitos.

Figura 4.2 - Diagrama do modelo em cascata



Fonte: adaptado de Couto (2011)

Chamado, em inglês, de *Waterfall model*, ele foi concebido na década de 70 por W. W. Royce (COUTO, 2011). Em sua proposta, o modelo possui sete etapas, as quais foram reduzidas a cinco:

- **Comunicação** ou requerimentos: levantamento de requisitos ou necessidades junto ao cliente (serviços a serem fornecidos, limitações, objetivos). Inclui também analisar a viabilidade do requisitado;
- **Planejamento** ou projeto: definição do cronograma do projeto (etapas, prazos, responsáveis) e da maneira de acompanhar seu cumprimento;
- **Modelagem** ou implementação: definição da estrutura de dados, da arquitetura, das interfaces... é um “desenho” do que será feito na próxima etapa;
- **Construção** ou verificação: codificar, mas também testar o que foi construído. Nesta etapa é onde se incluem os testes unitários e os testadores externos;
- **Implantação** ou manutenção: também chamada de emprego. É quando é realizada a entrega (instalação completa, incluindo de banco de dados, se necessário) e também o suporte e a manutenção do *software*.

Existem outras variações deste modelo e que elencam outras etapas, mas as cinco aqui citadas englobam o conteúdo de outras que porventura possam existir.

- **Vantagens:**

- Torna processo de desenvolvimento sistemático, ordenado, estruturado e linear;
- De certa forma, esta abordagem é uma norma usada até nos modelos mais modernos;
- Simples e de fácil entendimento.

- **Desvantagens:**

- Pouco foco é dado aos testes, aumentando o risco de serem encontradas anomalias no futuro (ou até do *software* estar incompatível com o desejado pelo cliente);
- Pouco foco dado à manutenção;
- Chega a ser metódico demais;

- Não permite entregas parciais, aumentando o risco de inconformidade com o solicitado pelo cliente.

4.2 Reconhecimento de padrões

O reconhecimento de padrões é uma área da ciência que possui como objetivo classificar (ou categorizar) objetos dentro de um certo conjunto de classes (ou categorias) conforme similaridades de características (ou padrões) dos membros de cada classe. Ela faz parte de um aglomerado de técnicas de *data mining*, ou mineração de dados (RIBEIRO, 2018).

Um *objeto*, por sua vez, pode ser de diferentes tipos: imagens, sinais (luz, radio…) e até mesmo seres humanos - que podem ser categorizados conforme sua personalidade, características físicas, etc. Outro termo de importante definição é *padrão*. Pode-se definir um padrão como um conjunto das características que permitem classificar objetos similares dentro de uma mesma classe. Para isso, a partir de um conjunto de informações de entrada, deve-se extrair quais informações são as mais relevantes dos objetos. Por fim, deve-se definir também que uma *classe* nada mais é do que um conjunto de características relevantes em comum (ou seja, um padrão) entre um conjunto de objetos (KAUER, 2016).

Em nosso dia-a-dia, o reconhecimento de padrões é feito inúmeras vezes. Um exemplo de fácil compreensão é a separação do lixo produzido em casa. Neste caso, podem existir duas classes (lixo reciclável e lixo não-reciclável) e os objetos a serem classificados são os resíduos gerados. As propriedades relevantes podem ser o tipo de material (papel, alumínio, orgânico) e seu estado (limpo, sujo, engordurado). Ao observar os resíduos que já foram corretamente descartados, o padrão da classe “lixo reciclável” pode ser definido como “material do tipo papel ou alumínio e estado do tipo limpo”. Enquanto isso, o padrão da classe “lixo não-reciclável” seria “material orgânico (independente do estado) ou materiais de outros tipos (papel, alumínio) porém com estado sujo ou engordurado”.

Nesse exemplo, as opções de classes eram apenas duas (lixo reciclável ou não) e pré-determinadas. Entretanto, pode-se ter reconhecimento de padrões com mais categorias ou até sem categorias pré-determinadas. Quando o conjunto de classes é desconhecido, cabe ao “classificador” definir quais são as classes e suas características (clusterização), o que se chama de um processo não-supervisionado de aprendizagem. No caso de as categorias serem pré-definidas, independente de quantas ou quais forem, o classificador deve reconhecer os padrões das diferentes classes com base nos objetos que já foram categorizados. Chama-se isso de processo supervisionado de aprendizagem (RIBEIRO, 2018). Neste caso, deve-se atentar para que a escolha das amostras de objetos de cada *cluster* seja de fato representativa, ou seja, para que a partir delas, consigamos definir as propriedades mais relevantes de cada padrão. Um projeto genérico de reconhecimento de padrões possui as seguintes etapas (RIBEIRO, 2018):

- Extração dos objetos a classificar ou descrever, bem como suas características;
- Seleção das características mais relevantes;

- Construção do classificador.

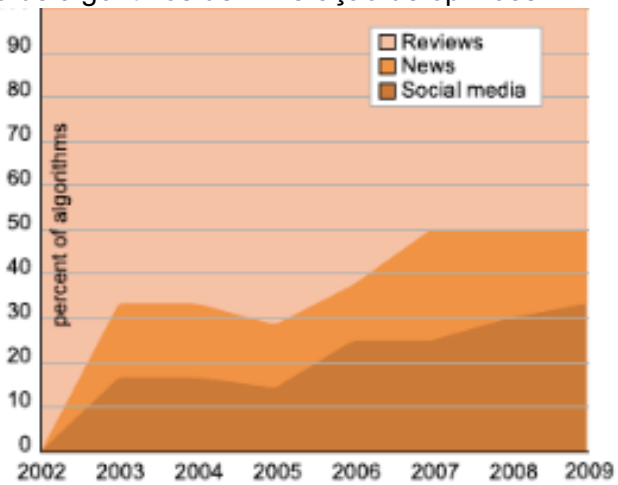
Na área da computação, os primeiros sistemas computacionais capazes de aprender foram desenvolvidos na década de 50, a partir, principalmente, do modelo de neurônio feito em 1943 por McCulloch e Pitts e também da regra de aprendizado neuronal, de elaboração de Hebb em 1949. Entretanto, foi apenas a partir do final da década de 80 que as pesquisas dessa área se voltaram para áreas mais comuns de nosso dia-a-dia, como esportes, exames de diagnóstico por imagem, dentre outros. Um exemplo bem simples de reconhecimento de padrões feito por sistema computacional e que temos contato diário é usado nos filtros de *spams* (BECKER e TUMITAN, 2013). Outra área de aplicação do reconhecimento de padrões é na mineração de opiniões - que foi o caso desse projeto e será detalhada no capitulo a seguir.

4.2.1 Mineração de opiniões: conceitos introdutórios

A mineração de opiniões é definida, segundo Becker e Tumitan (2013), como "*qualquer estudo feito computacionalmente envolvendo opiniões, sentimentos, avaliações, atitudes, afeições, visões, emoções e/ou subjetividade de forma textual*". Também chamada de análise de sentimentos ou análise de subjetividade, ela pode ser tratada como um problema de reconhecimento de padrões: os objetos são os textos (por isso, também pode ser chamada de *text mining*), as classes são os sentimentos (ou opiniões) e os padrões são características desses textos (palavras, autores, dentre outros).

A análise de sentimentos tem crescido nos últimos anos, principalmente com a popularização das mídias sociais - que incluem redes sociais, fóruns de discussão e sites de vendas. Hoje, é comum o uso dessa análise no meio corporativo para melhoria de seus serviços aos clientes: ao saber o que seu público-alvo pensa/sente a respeito de certos assuntos (como um produto/serviço), é possível adequar seu *portfolio* conforme os anseios. Inclusive, Tsytsaru (2009) identificou o aumento de algoritmos feitos na área de mineração de opiniões em redes sociais. A Figura 3.3 apresenta esse crescimento, principalmente a partir de 2008/2009.

Figura 4.3 - Distribuição de algoritmos de mineração de opiniões



Fonte: Tsytsaru (2009)

Um projeto genérico de mineração de opiniões possui as seguintes etapas (BECKER; TUMITAN, 2013):

- Identificar as opiniões expressas em um conjunto de documentos/textos, o que pode ser dito, em linguagem genérica de reconhecimento de padrões, elencar as diferentes classes presentes nos objetos;
- Classificar os diferentes textos, conforme seus sentimentos;
- Apresentar os resultados de forma agregada e sumarizada.

Um objeto, na mineração de opiniões, pode ser qualquer tipo de texto, independente de seu tamanho ou formato. Alguns exemplos de fontes tradicionais de objetos são páginas *web*, notícias, *posts* em redes sociais, comentários em *sites* de venda de produtos, *tweets*, dentre diversos outros. No geral, uma opinião é formada de duas características principais. A primeira delas é um alvo, que pode ser definido como um tópico ou uma entidade (um produto, uma pessoa, uma organização, a economia…) - entretanto, existe a possibilidade do alvo ser *null*. A segunda é uma opinião, que pode ser através de um sentimento ou uma emoção - que são conceitos diferentes (KAUER, 2016).

Sentimentos são categorizados, muitas vezes, de forma binária (positivo ou negativo) a respeito do alvo. Existe também a possibilidade de serem classificados de forma discreta (*e.g.* com os sentimentos positivo, negativo e neutro) ou até contínua, através de um intervalo que representa sua intensidade (geralmente de -1 a 1). Emoção, por sua vez, é usada para definir pensamentos mais subjetivos do autor, como raiva, felicidade, tristeza, dentre outros e não estão necessariamente relacionados com um alvo (BECKER; TUMITAN, 2013).

4.2.2 Níveis de análise textual

A análise de sentimentos pode ocorrer em diferentes granularidades. A escolha da mais adequada depende do contexto do projeto e cabe ao projetista fazê-la devidamente. As diferentes possibilidades são, segundo (BECKER; TUMIAN, 2013):

Análise em nível de documento: o documento (texto, noticia, post…) é tratado como um todo, resultando em apenas um sentimento ou emoção. Ela é adequada quando o documento tem apenas um alvo;

Análise em nível de sentença: como o próprio nome diz, refere-se determinar o sentimento de cada sentença ou frase. Seu uso é recomendado quando o documento possui opiniões a respeito de diferentes alvos - por exemplo, um post no *Facebook* de um eleitor que compara diferentes candidatos;

Análise em nível de entidade e aspecto: neste caso, é determinado o sentimento específico para cada entidade/aspecto dentro do documento (independente se a análise for por sentença/oração ou do documento como um todo). É o nível mais complexo de análise, pois requer identificar quais entidades/aspectos foram avaliados no decorrer do documento e tem sido mais usado na revisão de produtos e serviços.

Em todos os níveis de análise, uma vez definidos os objetos (sejam eles documentos, sentenças ou entidade/aspecto), esses podem ser divididos em objetivos - quando apresentam apenas um fato e que, portanto, não são usados na análise - ou subjetivos - quando apresentam opinião ou sentimento.

4.2.3 Análise linguística

Opiniões são um conceito subjetivo e que são expressadas em linguagem natural. É importante entender como opiniões são expressas para melhor reconhecê-las em um texto. Dentro dessa expressão, há diferentes particularidades que dificultam o processo de mineração. A seguir, serão listadas três dessas situações.

A primeira delas é a chamada correferência. Este termo significa referenciar a mesma coisa utilizando expressões diferentes, seja com sinônimos ou até com pronomes (BECKER; TUMITAN, 2013). Ou seja, na Internet, por exemplo, encontraremos diferentes referências ao candidato Jair Bolsonaro (PSL) em um mesmo documento: “Bolsonaro”, “Bolso”, “candidato do PSL”, “Deputado”, dentre outras. Além disso, a referência pode ser feita através de pronomes. No trecho: “o professor é muito bom, ele é atencioso e sua didática é ótima” temos três termos diferentes para referenciar o professor: “professor”, “ele” e “sua”. Nesses casos, as diferentes referências devem ser unificadas.

A segunda particularidade está em analisar contextos. Alguns adjetivos podem parecer intuitivos para identificar um sentimento ou emoções: “bom”, “ruim”, “ótimo” e “péssimo” são alguns exemplos. Entretanto, analisar apenas essas palavras não é o suficiente para reconhecer o real sentimento do objeto. Às vezes, o contexto modifica seu significado: nas frases “muito obrigado pela ajuda” e “a contragosto, fui obrigado a fazer isso”, a palavra “obrigado” possui sentimentos diferentes em ambas - na primeira, sendo positivo; na segunda, negativo. Em outros casos, podem não existir adjetivos que deixam os sentimentos explícitos, como na frase “a sua comida está com cheiro de chulé”. Por fim, há que levar em consideração o uso de advérbios de negação, que podem alterar inteiramente uma opinião (BECKER; TUMITAN, 2013).

Por fim, a terceira dificuldade está no tratamento de ironias e sarcasmos, algo muito corriqueiro em documentos - principalmente naqueles que se referem a política ou esportes (BECKER; TUMITAN, 2013).

4.2.4 Redes sociais

A mineração de opiniões em redes sociais pode ter os problemas citados acima agravados, principalmente o terceiro (a respeito de ironia). Em um local cujo linguajar é mais informal, é mais comum encontrar situações desse tipo. Além disso, é frequente também se deparar com gírias e expressões populares. Por fim, outra dificuldade está na constante presença de abreviações e erros de digitação, de ortografia e/ou gramaticais. Em compensação, uma forte vantagem de trabalhar com mídias sociais está na grande quantidade de dados que podem ser gerados a respeito

de um alvo ou tópico, o que pode fazer com que esses pontos se tornem irrelevantes ou *outliers* (SOUZA; PEREIRA; DALIP, 2017).

4.2.5 Mineração de opiniões: etapas

Um projeto de análise de sentimentos é dividido em três etapas, conforme apresentado na Seção 4.2.1. Cada uma delas será apresentada em mais detalhes a seguir.

4.2.5.1 Identificação

A partir do conjunto de textos extraídos da fonte desejada (seja um portal de notícias ou alguma rede social), deve-se identificar os diferentes alvos bem como definir o tipo de sentimentos ou emoções desses alvos - etapa que depende da granularidade (ou nível de análise) que será utilizada. A dificuldade da identificação é algo que pode variar conforme a fonte dos textos e a estruturação deles. A aplicação mais frequente, até então, da análise de sentimentos é na avaliação de produtos e serviços. Nesses casos, pode ser mais fácil identificar os alvos: um *smartphone*, por exemplo, pode ter como alvos a bateria, a câmera, a memória, dentre outros; um hotel, por sua vez, teria, dentre outros, limpeza, vista, localização e refeições. Quando tratamos dos níveis de análise por documento e sentença, deve-se considerar que o documento ou a sentença, respectivamente, em sua integridade, trata apenas de um alvo. Determinar o alvo de cada documento/sentença, entretanto, não é trivial (BECKER; TUMITAN, 2013).

A solução mais simples é pré-determinar uma entidade para todos, o que não é aplicável em todas as minerações, além de agravar a questão das correferências. há também a possibilidade de aplicar técnicas de identificação de alvos a partir de uma lista de possibilidades, como: extração baseada em substantivos frequentes, extração através da relação entre o alvo e o sentimento ou extração usando modelos de tópicos (KAUER, 2016).

Por fim, nessa etapa, é possível que seja feita a diferenciação apresentada na Seção 4.2.1, entre objetos (sejam eles documentos ou sentenças) objetivos e subjetivos.

4.2.5.2 Classificação

Frequentemente, segundo Becker e Tumitan (2013), a classificação de sentimentos (também chamada de classificação da polaridade) é binaria, ou seja, com um conjunto de classes contendo apenas “positivo” e “negativo”. Entretanto, há outras maneiras de se classificar textos e que podem ser usadas conforme o interesse da análise. Alguns exemplos, conforme introduzidos na Seção 4.2.1:

- Classificações com diferentes graus de intensidade, como “Muito positivo”, “Moderadamente positivo”, etc, sendo uma categorização discreta e que só pode ser feita caso haja maneira de distinguir os níveis (como quantidade de adjetivos, uso de advérbios…);

- Classificação continua através de escala numérica, que varia entre -1 (mais negativo) e 1 (mais positivo);
- Inclusão da classe “neutra”, que pode incluir textos sem sentimento claro ou simplesmente sem sentimentos (BECKER; TUMITAN, 2013) - neste caso, a sua identificação pode ser feita tanto aqui quanto na etapa de identificação;
- Classificação discreta através de emoções (raiva, felicidade, medo…), que é difícil de ser feita, já que algumas palavras podem apresentar ambiguidade em relação a emoções transmitida.

Independente da abordagem escolhida para categorização, a solução dificilmente será trivial, já que algumas dificuldades são certas de serem encontradas. Conforme apresentado nas seções anteriores, há diversas dificuldades relacionadas à análise de linguagem natural, como erros de ortografia, gírias, sarcasmo, correferência e dependência de contexto. Além desses, uma opinião pode se diferenciar conforme o observador. há exemplos simples, como por exemplo na frase “O candidato Joao foi eleito de forma avassaladora” - para um apoiador/eleitor do Joao, a polaridade da frase seria positiva; entretanto, para quem fazia oposição a ele, a polaridade é negativa. Contudo, há situações mais “cinzas”, principalmente na classificação discreta através de emoções, em que a polaridade de conteúdo subjetivo nem sempre é objeto de consenso - estima-se que, para humanos, dificilmente o consenso chega a 75% (BECKER; TUMITAN, 2013).

4.2.5.3 Apresentação

Os resultados obtidos com uma mineração de opiniões nada mais são do que indicadores a respeito do alvo. Sendo assim, a forma mais adequada de exibir o que foi encontrado em um certo projeto depende de seu objetivo. Ao tratarmos de avaliações de um produto, por exemplo, o objetivo é, geralmente, mostrar ao potencial consumidor os pontos fortes e fracos de um produto. Para isso, costuma-se exibir um sumario que leva em consideração a opinião de todos os consumidores anteriores - como em um *site* de aquisição de eletrônicos (KAUER, 2016). A Figura 3.4 é um exemplo de método de visualização. Nela, observa-se o resumo de avaliações feitas a respeito de um restaurante da Grande Florianópolis, o qual apresenta avaliação geral, do serviço, do ambiente e de pratos específicos.

Figura 4.4 - Apresentação de avaliações feitas sobre um restaurante de Florianópolis

Fonte: https://www.tripadvisor.com/

Outros exemplos de exibição de resultados são através de associações a aspectos temporais (variação de um sentimento ao longo do tempo) ou geográficos (variação de um sentimento conforme a localização). Essas abordagens podem ser usadas, dentre inúmeras possibilidades, para avaliar a valorização de uma ação na bolsa de valores ou identificar em que bairros um time tem mais/menos torcedores que seu rival local.

4.2.6 Mineração de opiniões: abordagens de classificação

Segundo Tsytsaru (2009), as diferentes abordagens existentes na literatura para classificação em mineração de opiniões podem ser divididas em quatro grandes grupos: léxicas, com aprendizado de máquinas, estatísticas e semânticas. Ainda de acordo com esse autor, as abordagens dos dois primeiros grupos são as mais presentes, mas nenhuma das quatro apresenta desempenho consideravelmente superior às outras.

Dessas, apenas as abordagens léxicas e com aprendizado de máquinas foram estudadas e implementadas ao longo do projeto, justamente devido à abundância de material a respeito delas na literatura e também pela praticidade de implementação pratica. Entretanto, as demais também serão brevemente descritas aqui.

4.2.6.1 Pré-processamento

A etapa de classificação possui como primeiro passo fazer um tratamento (ou indexação) dos textos "adquiridos", feita entre a aquisição dos objetos e a classificação em si. Segundo Feldman e Sager (2006), essa é a parte mais crucial em mineração de textos: "*pode-se afirmar que a mineração de textos é definida por essa etapa*". É aqui em que os documentos e seus termos (palavras) são convertidos em valores numéricos - permitindo a sua análise estatística ou computacional. O produto final do pré-processamento é uma matriz numérica, esparsa, cujas linhas representam os diferentes documentos (D_i) e as colunas, os termos que podem aparecer (T_j). No campo (D_i, T_j), encontra-se frequência do termo T_j no documento D_i. Essa matriz é conhecida como *Bag of Words* (ou BoW).

Uma indexação bem feita ocorre quando a dimensão da matriz BoW é reduzida para otimizar o tempo de processamento na classificação e, para isso, apenas as informações mais relevantes são mantidas. Dessa maneira, a precisão do método de categorização tende a ser mais alta (FELDMAN; SAGER, 2006).

A primeira ação no pré-processamento, uma vez tendo selecionado os documentos para análise, é chamada de **tokenização**. Ela consiste em decompor os objetos de análise conforme seus termos (palavras). Para a decomposição, existem diferentes delimitadores que costumam ser usados: espaço em branco, tabulação, quebras de linhas e caracteres especiais são alguns exemplos. Existem também diferentes maneiras de ordenar os termos tokenizados, através de uma técnica chamada ***n*-gramas**. Com ela, podemos determinar a quantidade de termos a serem combinados por *token*, representada pelo *n*. Ou seja, com unigramas, cada *token* tem apenas uma palavra; com bigramas, duas palavras; e assim por diante (FELDMAN;

SAGER, 2006). Um exemplo: na frase “dez dicas rápidas”, ao tokenizar por unigramas, temos os *tokens* “dez”, “dicas” e “rápidas”; já por bigramas, os *tokens* seriam “dez dicas”, “dicas rápidas” e também “dez rápidas”.

Uma vez que os *tokens* de todos os documentos estejam levantados, as colunas da *BoW* são montadas a partir deles. Cada coluna representa um *token* T_j, sendo que não devem haver colunas distintas com o mesmo *token*.

Em seguida, deve ser feita uma **limpeza** do conteúdo restante. Novamente, existem diversas atividades possíveis de serem feitas para “limpar” um documento - a seleção das mais adequadas depende da fonte dos objetos e também dos objetivos da análise. Alguns exemplos de ações são: passar todos os caracteres para o minúsculo; remover caracteres especiais ou acentuados, numerais ou pontuação; remover URL’s e links; remover *emojis* (principalmente quando trabalhando com redes sociais); remover marcações de usuários e/ou páginas (principalmente quando trabalhando com redes sociais); dentre outros (SOUZA; PEREIRA; DALIP, 2017).

Outra operação necessária a ser realizada na limpeza é a chamada **remoções de *stopwords***. Uma *stopword* (ou palavra de parada) é uma palavra funcional, mas que pode ser considerada irrelevante para o conteúdo e, geralmente, pertence a uma das classes gramaticais: preposições (de, a, para, por…), artigos (o, um), pronomes (aquele, ali…) ou advérbios. Geralmente, a remoções de *stopwords* é feita a partir de uma lista de palavras “removíveis” (RIBEIRO, 2018).

Após a limpeza, ainda no pré-processamento, é feita uma **normalização linguística**, sendo a do tipo morfológica a mais comum e simples de ser feita. Ela pode ser realizada através de duas maneiras: *stemming* e redução canônica. A primeira reduz todas as palavras a seu radical (também chamado de *stem*), quando necessário, eliminando afixos e/ou sufixos. Por exemplo: supondo que em um documento, estejam presentes as palavras “estudante”, “estuda” e “estudando” uma vez cada. Após o processo de *stemming*, tornam-se todas as mesmas, “estuda”, indicando que ela apareceu três vezes. Já a redução canônica, também chamada de *lemmatization*, propõe reduzir os verbos ao infinitivo (que terminam em *-ar, -er, -ir* ou *-or*) e, para substantivos e adjetivos, é proposto unificá-los no masculino e singular (RIBEIRO, 2018).

A última etapa do pré-processamento é atribuir valores numéricos aos documentos e seus *tokens* e, novamente, existem diversas maneiras de fazê-lo. A primeira é uma atribuição binária (ou modelo booleano), ou seja, avaliar se no documento D_i aquele *token* (T_j) está presente (valor 1) ou não (valor 0) (GONÇALVES, 2018). A Figura 4.5 apresenta um modelo de matriz *BoW* booleana.

Figura 4.5 - Modelo de matriz *Bag of Words* booleana

Documentos	Termo 1	...	...	Termo k	...	...	Termo n
d_1	1	...	...	1	...	...	0
...	...	...	...	...	...	...	...
d_j	0	...	...	1	...	...	0
...	...	...	...	...	...	...	...
d_m	0	...	...	1	...	...	1

Fonte: Gonçalves (2018)

Outra possibilidade é de contar a quantidade de vezes que T_j aparece em D_i, já que ele pode estar presente mais de uma vez, o que pode influenciar seu sentimento (GONÇALVES, 2018). Além dessas, existe uma técnica chamada *TF-IDF*, que atribui valores numéricos da seguinte forma:

- TF (*term frequency*) é um índice que procura valorizar *tokens* que aparecem em abundância em **um** documento. Tal índice pode ser determinado através de contagem (número natural) ou relativamente:

$$TF_{ij} = \frac{f_{ij}}{\sum_j f_i}, \quad (1)$$

onde f_{ij} é a quantidade de vezes que o termo T_j está no documento D_i e $\sum_j f_i$ é a quantidade de termos no documento D_i.

Outras possibilidades são que TF seja calculada de forma normalizada (quantidade do *token* dividida pela maior quantidade de um *token* daquele documento) ou sublinear (logaritmo da TF apresentada na Equação 1).

No entanto, quando um *token* aparece muito no *corpus* (conjunto total de documentos), não é possível "diferenciar" os documentos entre si. Para tratar isso, é usado o IDF:

- DF (*document frequency*) é a quantidade de documentos em que um termo/token T_j aparece. Nesse caso, deve ser um valor natural (quantidade absoluta de documentos);

- IDF (*inverse document frequency*) é usado para valorizar os termos que aparecem raramente no conjunto de documentos e que, portanto, são bem relevantes na classificação. É calculado através da fórmula $IDF = \log\left(\frac{N}{DF}\right)$, onde N é o número total de documentos.

Para encontrar o peso TF-IDF de um *token* T_j em um documento D_i, basta multiplicar $TF - IDF = TF * IDF$. Com esse peso atribuído, é possível fazer duas observações: quanto maior a frequência de uma palavra em um documento, maior é sua relevância para ele; e quanto maior a frequência de uma palavra em **todos** os

documentos, **menor** é seu fator discriminante entre os documentos (GONÇALVES, 2018).

Ao final do pré-processamento, independente de quais técnicas foram usadas, o *corpus* é transformado: cada documento torna-se um vetor com n colunas numéricas, onde n é a quantidade total de termos presentes no *corpus*, formando assim a matriz *BoW*.

Outras etapas podem ser adicionadas ao pré-processamento, bem como as já descritas podem ser feitas com outros métodos e técnicas. Existem inúmeras formas de fazê-los e a escolha pela "ideal", que resulta em maior precisão da classificação (cujas abordagens serão descritas a seguir), depende do contexto do *corpus* e do objetivo da análise de sentimentos.

4.2.6.2 Abordagem baseada em dicionário

A maneira mais simples de se categorizar textos em sentimentos/opiniões, segundo Freitas (2013), é a chamada de abordagem léxica, linguística ou baseada em dicionário. A fundação de seu uso é através de dicionários (ou léxicos) de sentimentos, ou seja, listas extensas de palavras/termos, cada um com sua devida avaliação sentimental ou emocional. A partir dessa enumeração, é feita uma procura, em cada documento, dos termos avaliados no léxico. Geralmente, um termo do dicionário recebe uma nota, *e.g.* -1 para polaridade negativa e +1 para positiva, ou uma categoria. É importante que o léxico possua não apenas as palavras de sentimento, mas suas possíveis flexões - principalmente caso as técnicas de *stemmização* e/ou *lemmatização* não funcionem para a integridade dos *tokens*.

A maneira mais simples é de somar a polaridade total dos termos do documento para definir seu sentimento: quanto maior é sua nota, mais positivo ele será. Sendo assim, basta que, em um texto, exista apenas uma palavra com sentimento definido no léxico para conseguir avaliar suas emoções/opinião. Esse método é, portanto, mais usado no caso de a análise dos documentos serem de granularidade pequena, ou seja, em análise de nível de sentença ou em análise de nível de documento - mas quando o documento possui quantidade limitada de caracteres ou palavras, como um *tweet* ou um SMS (SOUZA; PEREIRA; DALIP, 2017).

Outra possibilidade é usar um método de co-ocorrência entre alvo e sentimento (BECKER; TUMITAN, 2013). Nesse caso, é criado um outro léxico com palavras-chave que definam os alvos possíveis dos documentos e, em seguida, é feita uma análise similar à descrita no parágrafo anterior. As condições para uso são as mesmas: análise em nível de sentença ou em nível de documento, quando este é curto.

Existe a possibilidade de encontrar léxicos-modelo na Internet, em diversos idiomas - mas a grande maioria é na língua inglesa. Para o português, dois exemplos de dicionários genéricos são o *OpLexicon* e o *SentiLex*-PT (BECKER; TUMITAN, 2013). Por serem genéricos, eles são importantes para auxiliar na mineração de qualquer tipo de texto. Entretanto, para obter resultados mais condizentes com a

avaliação humana, é necessário criar dicionários adaptados para o contexto - seja algum disponibilizado na Internet ou até um criado pelo projetista.

Esta abordagem é simples e, muitas vezes, pode ser suficientemente boa para categorizar textos. Entretanto, ela não leva em consideração ironias/sarcasmos, mudança de polaridade conforme o contexto e, para o caso de redes sociais, a abundância de gírias e erros de escrita pode prejudicar seu desempenho. Pensando nisso, uma abordagem mais inteligente (porém mais custosa) e que se aproxima mais de uma classificação feita por humano foi elaborada: a por aprendizado de máquina.

4.2.6.3 Abordagem baseada em aprendizado de máquina

As técnicas de aprendizado de máquina (também chamado de *machine learning*) consistem em delegar ao computador o descobrimento automático das regras de classificação dos documentos através de algum algoritmo escolhido e configurado pelo projetista. Esses algoritmos devem ser capazes de identificar as informações mais relevantes de cada classe de maneira que um ser humano, “na mão”, não conseguiria ou demoraria muito mais tempo para fazê-lo (BECKER; TUMITAN, 2013).

Existem dois tipos de *machine learning*: aprendizado supervisionado e não supervisionado, sendo o primeiro o mais utilizado em mineração de opiniões e o que será apresentado neste tópico.

Os métodos de aprendizado de máquina supervisionado partem de um conjunto de dados que devem ser rotulados, antes do treinamento, por um especialista no assunto (RIBEIRO, 2018). Esse conjunto é, em seguida, estudado pela máquina, que deve determinar um modelo de classificação com base nesse corpus e nas classes nele existentes (por exemplo, “positivo” e “negativo”). Apenas com esse modelo elaborado que é possível prever o sentimento de novos documentos.

Existem diversos algoritmos de classificação bem com diversas métricas de avaliação da qualidade do modelo - tanto os algoritmos quanto as métricas serão apresentados posteriormente.

Para obter melhores resultados, deve-se atentar para os dados de treino a serem fornecidos. Cada termo T_j da matriz *BoW* fornecida para treino corresponde a um atributo discriminante (ou *feature*). As emoções possíveis de serem classificadas são denominadas atributos alvo (ou *targets*). Geralmente, para uma análise mais otimizada, extraem-se apenas as *features* mais relevantes e que mais contribuam para a categorização, e não todas as existentes. Além disso, os documentos a serem previamente classificados e que serviram como “base” devem ser escolhidos cuidadosamente, de forma que melhor representem o conjunto total de objetos com a menor quantidade possível (GONÇALVES, 2018).

Esse método possui duas limitações principais (BECKER; TUMITAN, 2013). A primeira está justamente na definição das emoções dos dados para treino. Por ter de fazê-la manualmente, isso pode limitar a quantidade e a qualidade dos dados disponíveis para o aprendizado da máquina. A segunda é que, uma vez “aprendendo”

como categorizar os documentos, essa lógica de classificação servirá apenas para os objetos de mesmo domínio. Ou seja, a forma de classificar comentários a respeito de celulares certamente não será a mesma do que a usada para classificar *feedbacks* sobre um hotel.

4.2.6.4 Métricas de avaliação

Uma vez optando por um algoritmo específico de classificação com aprendizado de máquina supervisionado, deve-se verificar sua capacidade de categorizar corretamente um novo documento (RIBEIRO, 2018). Isso é feito após o treinamento, com base nos documentos usados para o próprio aprendizado e utilizando de quatro métricas principais:

- **Acurácia (Acc):** capacidade do modelo de prever corretamente, calculada através da fórmula:

$$Acc = \frac{Acp}{N}, \tag{2}$$

onde Acp representa a quantidade de amostras corretamente previstas e *N* a quantidade total de amostras.

- **Precisão (Pr):** mede a porção de previsões corretas de uma classe específica A:

$$\Pr(A) = \frac{Acp(A)}{Acp}, \tag{3}$$

onde $Acp(A)$ significa a quantidade de amostras corretamente previstas de uma classe A e Acp, a quantidade total de amostras corretamente previstas.

- **Revocação (também chamada de *Recall* ou Rec):** número de instâncias de uma dada classe A prevista na classe correta:

$$Rec(A) = \frac{Acp(A)}{N_A}, \tag{4}$$

onde $Acp(A)$ significa a quantidade de amostras corretamente previstas de uma classe A e N_A, a quantidade total de amostras da classe A.

- ***F-measure*:** média harmônica entre precisão e revocação, mais usada para avaliar casos de classificadores desbalanceados (com desempenho muito diferente entre classes):

$$F - measure = \frac{2}{\left(\frac{1}{\Pr}\right)+\left(\frac{1}{Rec}\right)} \tag{5}$$

4.2.6.5 Demais abordagens

Outros tipos de abordagens usados para categorização de textos são os chamados métodos estatísticos e os métodos semânticos (BECKER; TUMITAN, 2013). Para a primeira, a técnica mais usada é a chamada *Pointwise Mutual Information* (PMI), que avalia as co-ocorrências de dois termos, sendo um deles notoriamente positivo ou negativo (extraído de léxicos genéricos), julgando que o outro terá o mesmo sentimento caso ocorra repetidas vezes com ele. A segunda é até similar à abordagem estatística, entretanto, ela também leva em consideração a distância entre os termos dentro do documento e não apenas a co-ocorrência. Ambas costumam ser usadas como complemento aos métodos léxico e/ou com aprendizado de máquina supervisionado.

4.2.7 Algoritmos de aprendizado de máquina supervisionado

Conforme apresentado na Seção 4.2.6, existem diversos algoritmos de classificadores que fazem parte da abordagem por aprendizado de máquina. Esta seção irá descrever alguns dos mais utilizados na literatura.

4.2.7.1 Naïve-Bayes

O algoritmo de Naïve-Bayes é um algoritmo probabilístico relativamente simples feito com base no Teorema de Bayes. O termo "*naïve*", em inglês, significa ingênuo e é usado no nome do processo pois ele assume que há uma independência entre as *features* do modelo. Ou seja, para um minerador de opiniões, cujo objeto de análise é do tipo texto, isso significa que os *tokens* (palavras, bigramas…) não possuem relação uns com os outros (SANTANA, 2018).

O Teorema de Bayes trata da probabilidade condicional:

$$P(A|\mathrm{B}) = \frac{\mathrm{P}(\mathrm{B}|\mathrm{A})\ \mathrm{P}(\mathrm{A})}{\mathrm{P}(\mathrm{B})}, \tag{6}$$

onde P(AlB) é a probabilidade *a posteriori* de A condicional a B, P(BlA) é a probabilidade condicionada de B condicional a A e P(A) e P(B) são as probabilidades *a priori* de A e B, respectivamente.

Para o algoritmo, basta, portanto, aplicar essa fórmula na classificação e, assim, identificar para cada documento, com base nos termos pertencentes a ele, à qual classe ele tem maior probabilidade de pertencer.

Este classificador é um dos mais usados dentro da mineração de textos (*text mining*) com aprendizado de máquina. Caso o pré-processamento seja bem feito, seu desempenho é comparável a algoritmos mais modernos, custosos e/ou sofisticados (GONÇALVES, 2018).

4.2.7.2 Árvore de decisão C4.5

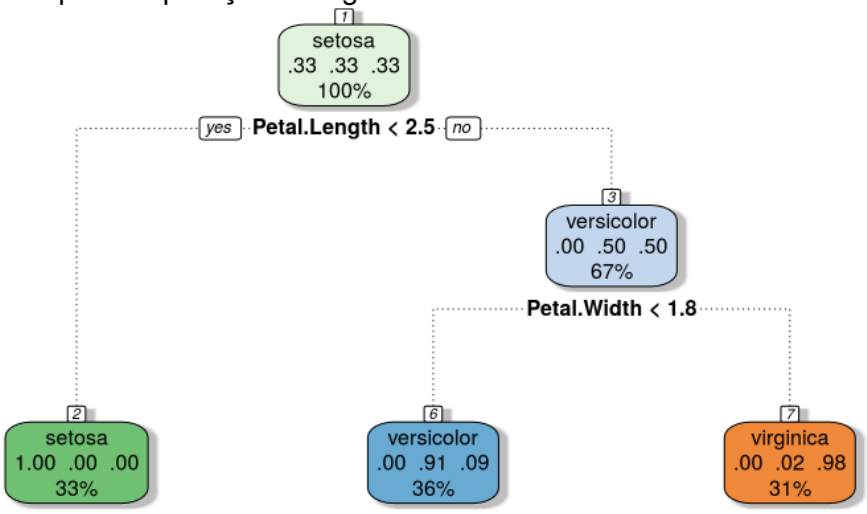
O algoritmo C4.5 faz parte da categoria de algoritmos de árvores de decisão, pois, ao final, constrói árvores de classificação a partir da amostra pré-classificada. Também chamado de J48 - esse é o nome da implementação do C4.5, feita na

ferramenta *Weka* - foi criado por Ross Quinlan em 1993 e é uma extensão de um outro algoritmo de árvore, de criação precedente, chamado ID3. O C4.5 é um dos mais usados na literatura de mineração de textos (GONÇALVES, 2018).

Para a construção da árvore, utiliza-se do conceito de entropia de informação (GONÇALVES, 2018). Ou seja, a cada nó da árvore, o algoritmo escolhe o *feature* que vai dividir as informações subsequentes de forma mais eficiente, sendo ele aquele contém mais "informações" que permitem categorizar (dado chamado de *information gain*). Isso significa que, quanto mais alto o nível do nó, mais relevante para a classificação ele é. As regras para avaliação de cada *feature* nas árvores são do tipo "se… então".

A Figura 4.6 apresenta um exemplo de árvore de classificação para bicicletas. Nesse caso, existem três categorias possíveis (*setosa*, *versicolor* e *virginica*) e, no início, a probabilidade é a mesma para as três (observa-se as três probabilidades iguais no meio do nó). Em seguida, é feita a primeira avaliação, em que se verifica o comprimento do pedal: se for menor que 2.5, é certo que a bicicleta é do tipo *setosa*. Se não, deve ser feita uma verificação da largura do pedal: se for menor que 1.8, então há 91% de probabilidade de a bicicleta ser do tipo *versicolor*; se não, 98% de ser *virginica*.

Figura 4.6 - Exemplo de aplicação do algoritmo C4.5



Fonte: *site EstatComp*

Outra característica fundamental do C4.5 é a possibilidade de fazer poda (*pruning*) da árvore criada (GONÇALVES, 2018). Isso significa substituir uma parte da árvore (sub-árvore) por uma "folha", com o objetivo de simplificar o algoritmo. Entretanto, uma poda só pode ocorrer quando o erro esperado da folha é menor do que o da sub-árvore retirada.

Para realização de *pruning*, é preciso basear-se na fórmula do erro esperado de um nó:

$$Erro(\text{nó}) = \frac{N-n+k-1}{N+k}, \tag{7}$$

onde N é o número de exemplos do nó, k é o número de classes e n é o número de exemplos de N pertencentes à classe C (classe com maior número de elementos).

Além disso, deve-se usar a fórmula do erro de uma sub-árvore:

$$Erro(subárvore) = \sum_i P_i * Erro(nó_i), \quad (8)$$

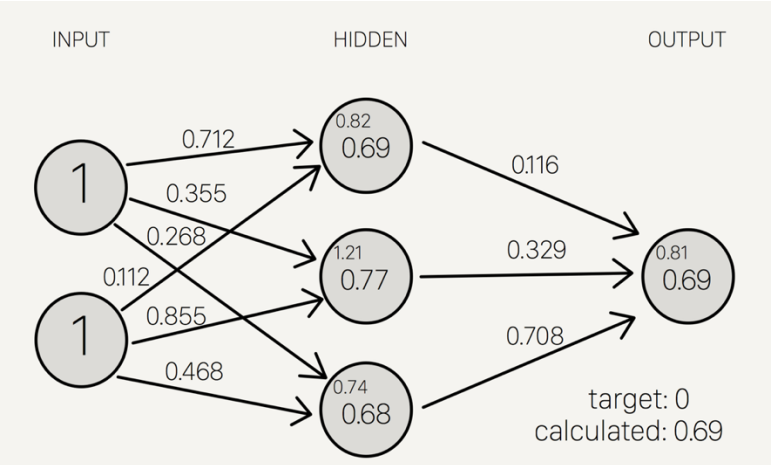
onde P_i representa a probabilidade da sub-árvore em relação ao nó pai ($Nó_i$).

4.2.7.3 Redes Neurais

Redes Neurais Artificiais (RNAs) são modelos computacionais que foram criados tomando como base os próprios neurônios humanos (ou seja, nosso sistema nervoso). Normalmente, esses modelos possuem neurônios interconectados em diferentes camadas os quais possuem capacidade de processar sinais de determinados tipos (imagem, som…) recebidos pelos neurônios de entrada (GURNEY, 1997).

Na Figura 4.7, podemos observar um exemplo simples de rede neural. Neste exemplo de rede, há duas entradas e uma saída, com uma camada intermediária "oculta" (*hidden*) para o usuário. Entre cada camada, observam-se os pesos modificam os valores numéricos, a serem explicados a seguir.

Figura 4.7 - Exemplo de aplicação de algoritmo de Rede Neural



Fonte: *https://stevenmiller888.github.io/*

Assim como para todos os outros algoritmos apresentados nessa seção, em mineração de opiniões, as Redes Neurais têm capacidade de aprender (e criar um modelo) a partir de uma amostra inicial já classificada (GURNEY, 1997). É nessa fase de aprendizado, também, que é determinada a quantidade de camadas de neurônios do modelo - ao contrário dos algoritmos iniciais, que possuíam esse valor pré-estabelecido.

Uma rede neural é chamada de aproximador de funções, ou seja, ela transforma um valor em outro. O principal tipo de uma RNA é chamado de

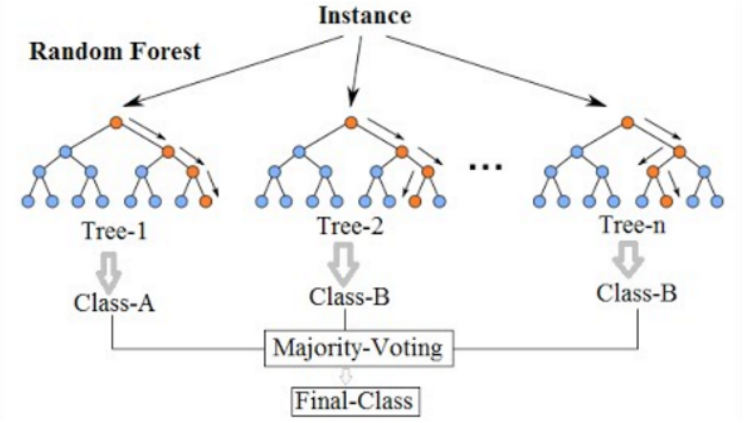
feedforward, ou seja, sem realimentação. Com ele, o sinal (valor) percorre o modelo em apenas uma direção: da entrada para a saída, através dos neurônios, sendo modificado ao longo do caminho. Cada conexão entre neurônios (ou no) recebe um valor numérico (peso), que multiplica o valor “recebido” antes de passar para o próximo neurônio. Vale comentar que uma conexão só pode ser feita entre dois neurônios de camadas distintas e consecutivas. Além disso, um neurônio de uma camada X pode receber valores de diferentes neurônios da camada X-1 - nesse caso, deve-se somar todos os valores “recebidos” para determinar seu valor (GURNEY, 1997).

A questão mais crítica de uma RNA é, durante a fase de aprendizado, determinar os pesos das conexões – para a qual existem diversas maneiras de fazê-lo, sendo a principal delas a propagação reversa. Nesse caso, são dados valores arbitrários para as conexões e simula-se, a partir de uma entrada, um valor para a saída. Esse valor calculado é então comparado com o valor correto, gerando um erro, a ser alimentado na rede para ajustar os pesos (GURNEY, 1997).

4.2.7.4 Floresta Randômica

O algoritmo da floresta randômica é simples de se entender: ele nada mais é do que um conjunto de árvores de decisão, que podem ser feitas com o algoritmo C4.5 apresentado anteriormente (CLÉSIO, 2018). A Figura 4.8 abaixo apresenta um exemplo genérico de floresta aleatória. Nela, observa-se que, a partir de uma instância, as avaliações são feitas em diversas árvores – cada uma, categorizando conforme seu algoritmo. Ao final, a categoria mais indicada pelas árvores é a determinada.

Figura 4.8 - Exemplo genérico de algoritmo de floresta aleatória



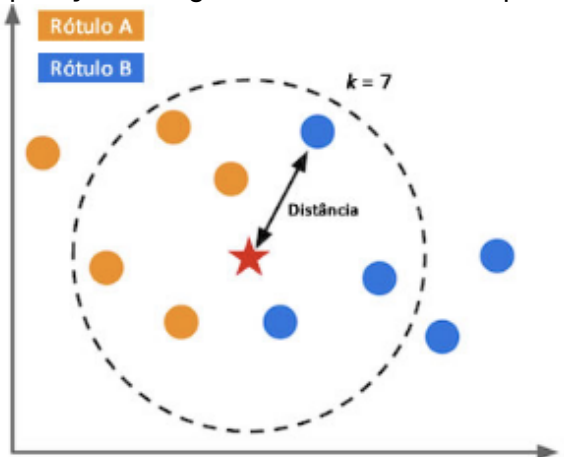
Fonte: *site Medium*

Comparado com o algoritmo de árvore de decisão, um algoritmo de floresta aleatória, como o nome diz, adiciona uma aleatoriedade extra ao modelo enquanto cria suas árvores. Essa aleatoriedade se da na hora de criar os nós: enquanto com o C4.5, seleciona-se o atributo (*feature*) com mais informação a ganhar, o algoritmo da floresta busca a melhor *feature*, mas apenas de um conjunto de atributos aleatoriamente selecionados (CLÉSIO, 2018).

4.2.7.5 K-Vizinhos mais próximos

Mais conhecido por seu nome em inglês (*K-nearest neighbours* ou k-NN), esse algoritmo é relativamente antigo (de 1967), porém é amplamente difundido. Ele é considerado simples de ser implementado, de fácil compreensão e com desempenho satisfatório, apesar da idade (GONÇALVES, 2018). O principal objetivo do algoritmo é determinar a classe de uma amostra com base nas classes dos seus "vizinhos" mais próximos já classificados. A Figura 4.9 apresenta um exemplo. Nela, nota-se que temos duas classes diferentes (A e B) e o valor de k é determinado como 7. Sendo assim, para classificar ao objeto desejado (estrela), deve-se encontrar os 7 vizinhos mais próximos (pré-classificados) e identificar suas categorias. Percebe-se que 4 são da categoria A (laranja) e 3 da B (azul) - como existem mais vizinhos de A, categoriza-se o objeto como sendo de A.

Figura 4.9 - Exemplo de aplicação do algoritmo k-vizinhos mais próximos



Fonte: *Site* Computação Inteligente

Existem, portanto, dois principais parâmetros do algoritmo: o valor de k e o método de calculo da distância. Determinar k depende do tamanho da amostra a ser classificada, bem como da já categorizada. Já para a distância, o método mais utilizado é o da distância Euclidiana (GONÇALVES, 2018):

$$d(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}, \quad (9)$$

onde $\boldsymbol{p} = (p_1, \ldots, p_n)$ e $\boldsymbol{q} = (q_1, \ldots, q_n)$ são dois pontos n-dimensionais. Um ponto representa o objeto a ser classificado, enquanto o outro, um vizinho; *n*, por sua vez, é a quantidade de atributos relevantes (*features*) - no caso desse exemplo, temos dois atributos: as coordenadas x e y.

Por mais que seja simples de se implementar, calcular a distância para todos os objetos pode ser bastante custoso para a máquina - o que pode consumir muito tempo de processamento. Por fim, o algoritmo é bastante sensível quanto à escolha de k, o que pode prejudicar seu desempenho de classificação.

5 METODOLOGIA UTILIZADA E SOLUÇÃO PROPOSTA

Neste capítulo, a metodologia utilizada no decorrer deste trabalho, *Design Science Research Methodology* (DSRM), será explicada na Seção 5.1. Em seguida, na Seção 5.2, a solução proposta para tratar do problema de previsibilidade eleitoral será descrita, com as justificativas das principais escolhas realizadas. Ao seu final, já na Seção 5.3, será apresentado o ferramental utilizado ao longo do projeto.

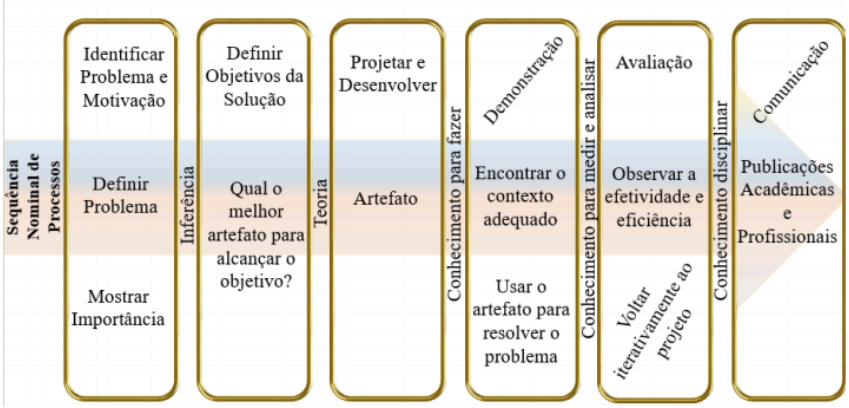
5.1 Metodologia utilizada

A solução desenvolvida foi tratada, desde o início, como um **projeto de pesquisa**. Segundo Ribeiro (2018), pesquisa é:

> O procedimento racional e sistemático que tem como objetivo proporcionar respostas aos problemas que são propostos. A pesquisa desenvolve-se por um processo constituído de várias fases, desde a formulação do problema até a apresentação e discussão dos resultados (*apud* Gil, 2007).

Elaborada por Peffers *et al.* em 2007, o nome da metodologia (*Design Science*) significa “ciência do projeto” e ela busca projetar e produzir sistemas inexistentes e, com isso, modificar situações existentes (RIBEIRO, 2018). É considerada, hoje, o estado-da-arte em metodologia de pesquisa, sendo citada em milhares de trabalhos desde sua criação. Segundo os autores da DSRM, os objetivos dessa ciência são projetar, criar e avaliar as técnicas de tecnologia da informação utilizadas para resolver os problemas identificados em algum contexto ou organização. É também parte fundamental da Design Science a comunicação dos resultados obtidos - que podem ocorrer por meio de inovações sociais, novas propriedades intelectuais, novos produtos, serviços ou processos - aos seus *stakeholders*. Conforme aponta Ribeiro (2018), outros autores do ramo reforçam o valor dessa ciência, como Aken e March & Smith. Além disso, apontam que o seu foco é em utilizar da tecnologia como solução para os problemas da sociedade. A metodologia foi resumida de forma gráfica por Lacerda *et al.,* em 2013, através do diagrama exposto na Figura 5.1. Nela, vemos as diferentes etapas relevantes do método em questão, as quais serão detalhadas a seguir.

Figura 5.1 - Representação gráfica da *DSRM*



Fonte: Ribeiro (2018)

Nota-se a presença de seis etapas na DSRM:

- **Identificar o problema e sua motivação:** define-se o problema a ser tratado pela pesquisa e justifica-se o desenvolvimento da solução;
- **Definir os objetivos para uma solução:** como seu título afirma, nessa etapa são listados os objetivos (geral e específicos) para a solução proposta;
- **Projetar e desenvolver:** etapa de criação, em si, da solução, a qual pode ocorrer de diversas maneiras, como um modelo, um método, um processo, um protótipo ou até mesmo um produto. É nesta etapa em que geralmente se faz uso de ferramentas computacionais, maquetes, protótipos, dentre outros;
- **Demonstração**: nessa etapa, demonstra-se a aplicação da solução proposta, seja através de experimentos, simulações ou estudos de caso, apresentando a resolução do problema (ou de parte dele);
- **Avaliação**: mensurar o impacto da solução desenvolvida e como ela atende as questões levantadas pelo problema. Através de métricas bem definidas, deve-se analisar se os objetivos propostos foram ou não cumpridos;
- **Comunicação**: quando apropriado, fazer a divulgação para outros pesquisadores, estudantes ou demais *stakeholders* o problema, sua relevância, a solução desenvolvida e os resultados obtidos.

Como objetivo do projeto é propor novos métodos ou novas ferramentas de previsibilidade eleitoral, essa metodologia apresentou-se como adequada para o uso ao longo do trabalho, como pode ser observado no Quadro 1. Nele, nota-se o que foi feito, nesta pesquisa, em cada etapa proposta pela metodologia:

Quadro 1 - aplicação da DSRM neste projeto

Etapa	**Aplicação no Projeto**
Identificar o problema e sua motivação	Imprevisibilidades eleitorais ao redor do mundo
	Imprevisibilidade eleitoral no Brasil
	Pesquisas acadêmicas de previsibilidade eleitoral sem aplicação de técnicas de *data mining*
Definir os objetivos para uma solução	Objetivo geral: ver Seção 1.1
	Objetivos específicos: ver Seção 1.2
Projetar e desenvolver	Método de previsão eleitoral com base em analisador de sentimentos de conteúdos publicados em redes sociais
Demonstrar	Demonstrar aplicabilidade e validade do novo método nas eleições presidenciais brasileiras de 2018
Avaliar	Analisador: através de indicadores como Acurácia e *F-measure*
	Método: através dos resultados da votação do primeiro turno das eleições
Comunicar	Apresentação dos resultados através deste documento.

Fonte: arquivo pessoal

Além da metodologia, é importante definir alguns conceitos a respeito da solução desenvolvida e detalhar um pouco a mais o que se pretendia construir no início do projeto. Isto será feito na seção a seguir.

5.2 Solução proposta

Neste trabalho, propõe-se a formulação de um novo método de se prever resultados de eleições a partir de aplicação de técnicas de *data mining*. Para guiar o desenvolvimento do estudo, foi aplicada a *Design Science Research Methodology* (DSRM), uma das metodologias mais referenciadas para projetos de pesquisa.

Dentre as diferentes teorias da Ciência Política apresentadas no Capítulo 3, a opção foi por basear-se na teoria de estratégias emocionais da Neuropolítica, já que estudar qualquer uma das três teorias clássicas não se mostrava lógico. Enquanto a teoria sociológica (da Escola de Columbia) já havia sido exaustivamente analisada no Brasil (Singer (2012), Terron & Soares (2010), Renno & Cabello (2014)), o Modelo de Michigan (ou teoria psicológica) e a Teoria Racional mostravam-se incondizentes com a realidade do cenário eleitoral brasileiro. Enquanto a falta de fidelidade partidária desqualifica o primeiro (LAVAREDA, 2011), a rejeição recorde do Governo (AMORIM, 2018) e o alto número de candidatos (G1, 2018) desqualifica a segunda. Além disso, através da Neuropolítica, a análise pode ser feita com base nas redes sociais, o que coaduna com o contexto que este trabalho se encontra e o posiciona como pioneiro no Brasil.

A respeito das redes sociais, o Datafolha (2018) indica que o *WhatsApp* e o *Facebook* são as mais populares entre o eleitorado brasileiro: aproximadamente 60% dos eleitores possuem conta nelas. O *Twitter*, por sua vez, era usado por apenas 10 a 15% dos votantes. Entretanto, o *WhatsApp* não fornece dados abertos a desenvolvedores de sistemas, o que impossibilita seu uso. Já o *Facebook* e o *Twitter* demandam uma licença autorizando a aquisição de seus dados. Para este trabalho, apenas o *Twitter* permitiu o acesso a suas informações e, por isso, foi a rede social considerada no estudo.

Para identificar as emoções transmitidas pelos candidatos no *Twitter*, optou-se pela aplicação de técnicas de mineração de opiniões (ou análise de sentimentos), uma área especifica de *data mining* cujo objetivo é justamente esse: determinar os sentimentos transmitido por textos. Essa análise será realizada nos conteúdos publicados pelos oito candidatos à Presidência da Republica de 2018 que lideraram a pesquisa de intenção de voto feita pelo Ibope em 24 de setembro de 2018 (G1, 2018). Os diferentes sistemas concebidos ao longo do trabalho, como um *web scraper* e um tradutor de *tweets*, foram desenvolvidos com uso do Modelo em Cascata apresentado no Capítulo 4, já que os aspectos envolvidos na solução proposta não demandam modelos mais complexos.

A essa descrição sucinta do projeto, adicionam-se os requisitos do sistema, que foram determinadas no início do desenvolvimento para guiar os trabalhos. Os requisitos para esse projeto são:

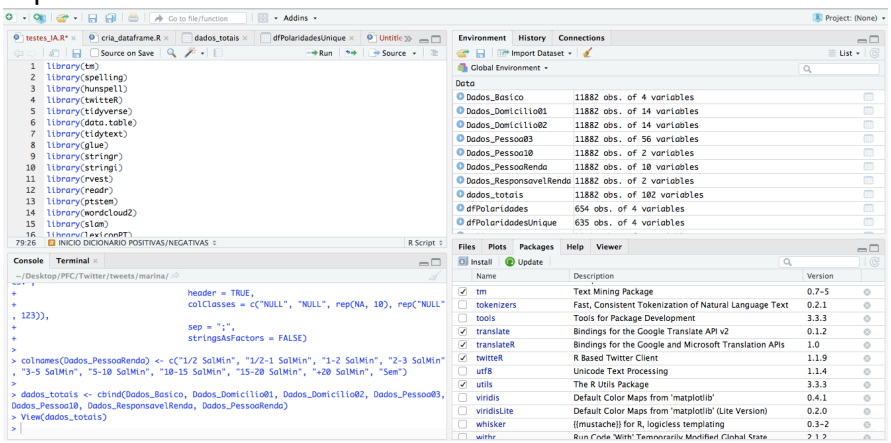
- O desenvolvimento não deve passar de 3 meses, com o prazo máximo de entrega sendo meados de novembro, conforme determinado pelo calendário acadêmico;
- Os *tweets* analisados dos candidatos devem ser a partir de 1° de janeiro de 2018, já que muitos deles manifestam-se ativamente em suas redes sociais desde antes do período eleitoral;
- Os *tweets* analisados dos candidatos devem ser até 6 de outubro de 2018 (incluso), sendo essa a data limite para manifestações dos candidatos em redes sociais imposta pelo Tribunal Superior Eleitoral:
- Não devem ser coletados *tweets* com conteúdo menor de 50 caracteres, já que publicações com tamanho menor a esse dificilmente apresentarão conteúdo suficiente para análise;
- Não devem ser coletados *tweets* que sejam respostas ou *retweets*, já que publicações desses tipos dificilmente apresentarão conteúdo suficiente para análise;
- A coleta será dos oito candidatos à Presidência que lideraram a pesquisa de intenção de voto realizada pelo Ibope em 24 de setembro;
- Categorizar, conforme o sentimento, os devidos *tweets* dos candidatos mencionados acima. Devem ser avaliadas diversas maneiras de categorização, mas apenas uma será escolhida para a construção do modelo final (com base nos desempenhos);
- Apresentar a classificação dos tuítes de diversas formas, de modo em que possam ser feitas as devidas análises a respeito da previsibilidade eleitoral;
- Avaliar a validade do novo método quanto à capacidade de prever o resultado de eleições.

5.3 Ferramentas utilizadas

5.3.1 *R* e *R-Studio*

Segundo o site de seu desenvolvedor, R é uma linguagem de programação muito utilizada para estatística e ciência de dados. É também um *software* que permite a aplicação dessa linguagem e, segundo a fabricante (a fundação *R Foundation for Statistical Computing*), sua popularidade tem crescido fortemente nos últimos anos. Todo o seu conteúdo é *open source* (*software*, bibliotecas, *packages*) e, por isso, optou-se pelo seu uso no projeto ao invés de seus concorrentes (como *SAS* e *Matlab*). Um dos *softwares* disponibilizados pela fabricante é o R-Studio, que permite a criação de sistemas/programas unicamente em R, mas com interface (ver Figura 5.2) mais para o usuário.

Figura 5.2 - Captura de tela da interface do *R-Studio*



Fonte: arquivo pessoal

Neste trabalho, o uso do *R* foi para a aquisição dos dados do *Twitter* (ou seja, para desenvolver o *web* scraper), já que possui bibliotecas específicas para isso. Foi usado, também, para aplicar algumas das técnicas do pré-processamento nos textos exportados da rede social. Por fim, seu uso ocorreu para fazer as classificações com base no uso de léxicos.

5.3.2 *Weka*

Segundo seu *site* institucional, o *Weka* (*Waikato Environment for Knowledge* Analysis) é um *software open source* desenvolvido em *Java* pela Universidade de Waikato, na Nova Zelândia, que engloba uma coleção de algoritmos de aprendizado de máquina para atividades de *data mining*. Para isso, contém também as ferramentas necessárias para fazer o pré-processamento dos dados, classificação, regressão e visualização. A janela inicial do programa é apresentada na Figura 5.3.

Figura 5.3 - Captura de tela da janela inicial do *software Weka*

Fonte: arquivo pessoal

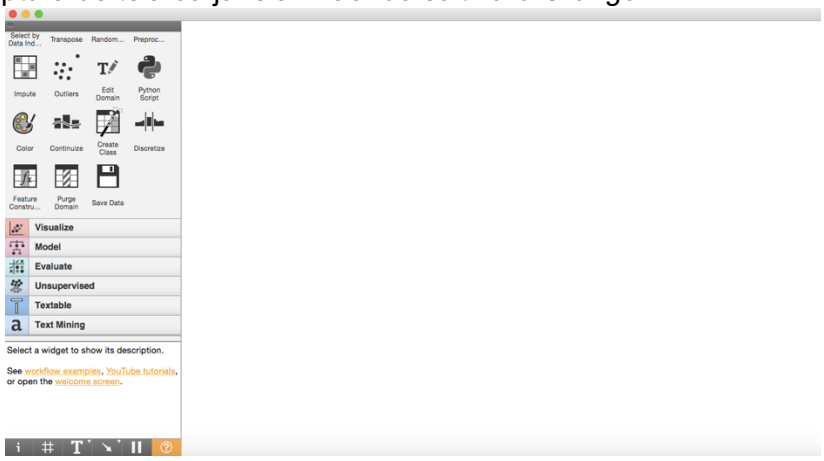
Ao longo do projeto, o *Weka* foi utilizado para avaliar, preliminarmente, as técnicas de pré-processamento e os diferentes algoritmos de classificação de aprendizado de máquina, já que seu desempenho era consideravelmente mais rápido do que o *Orange* (a ser apresentado na seção seguinte).

5.3.3 *Orange*

Segundo o site da ferramenta, o *Orange* foi desenvolvido pela Universidade de Liubliana (Eslovênia) e é uma ferramenta *open source* para aprendizado de máquina e visualização de dados criada em 1997. O ambiente de programação é visual, com uso de diagrama de blocos para a construção dos sistemas e onde cada função é chamada de *widget*. O programa também oferece a possibilidade de ser utilizado como uma biblioteca para a programação em linguagem *Python*.

A Figura 5.4 apresenta a janela inicial do *software*. À esquerda, estão exibidas algumas das funções (*widgets*) e também algumas das bibliotecas (*Text Mining, Textable*...).

Figura 5.4 - Captura de tela da janela inicial do *software Orange*



Fonte: arquivo pessoal

No projeto, o uso do Orange foi feito na construção do modelo final de categorização, com algoritmo de Naïve-Bayes, bem como em sua aplicação nos *tweets* dos candidatos. Além disso, foi usado para a visualização dos resultados, já que é um programa feito especificamente para isso.

6 MÉTODO DESENVOLVIDO

O presente capítulo descreve o método desenvolvido para o aprimoramento de previsões eleitorais com base em dados da rede social *Twitter* e na teoria de estratégias emocionais da Neuropolítica. O método proposto será aplicado no primeiro turno das eleições presidenciais brasileiras de 2018, avaliando o desempenho de oito dos principais candidatos ao pleito.

A primeira parte desse trabalho, aquisição das postagens da rede social, feita através de um *web scraper* desenvolvido na ferramenta R, será apresentada na Seção 6.1. Em seguida, na Seção 6.2, a etapa de identificação (conforme apresentada na Seção 4.2.5) será apresentada, detalhando as principais decisões tomadas quanto à categorização do conteúdo. Após, as diferentes técnicas utilizadas para o pré-processamento dos dados serão detalhadas na Seção 6.3, conforme explicadas na Seção 4.2.6. Junto disso, serão exibidas algumas análises relevantes para a posterior categorização das postagens. Em sequência, na Seção 6.4, serão apresentadas as diferentes abordagens experimentadas para fazer a categorização dos tuítes: por léxicos e por aprendizado de máquina, ambas explicadas na Seção 4.2.6. Para cada uma, diversos métodos foram aplicados e o desempenho de cada um será demonstrado, bem como a justificativa da escolha pelo algoritmo de Naïve-Bayes, que foi detalhado na Seção 4.2.7. O modelo por ele construído será apresentado na Seção 6.5, com o uso de imagens retiradas das ferramentas para sua exibição. Por fim, a Seção 6.6 deste capítulo contém a etapa de visualização (também chamada sumarização ou apresentação) dos dados minerados, introduzida na Seção 4.2.5. Diferentes abordagens serão apresentadas com amostras das respostas obtidas. Entretanto, as análises a respeito do conteúdo em si, ou seja, as diferentes conclusões que podem ser feitas a partir da visualização, só serão feitas no capitulo seguinte.

6.1 *Web scraper*

Para a coleta de dados da pesquisa, foi desenvolvido um sistema que a fizesse automaticamente com o uso da técnica de *web scraping*, implementada através da ferramenta *open-source* R.

Web scraping representa um processo de aquisição de dados/informações de alguma pagina *web*. De forma geral, é um pequeno sistema que percorre um ou mais *site*(s) em busca dos dados pré-determinados. Além disso, um *web scraper* deve, a partir da pagina especificada pelo usuário, carregar as informações desejadas e as exporta de tal maneira que continuem acessíveis para análise - para isso, são geralmente exportadas em diferentes formatos. A Figura 6.1 descreve essa etapa de maneira genérica.

Figura 6.1 - Apresentação das etapas de um *web scraper*

Fonte: Portal *ProWebScraping*

Para o desenvolvimento do sistema, foi utilizado como base o método em cascata, apresentado na Seção 3.2.1.2. Ele divide essa etapa de criação do *scraper* em cinco fases:

- Comunicação: levantamento de requisitos
  - Adquirir tuítes de usuários determinados pelo usuário do sistema;
  - Para cada *tweet*, informar as seguintes variáveis: usuário (candidato) que publicou, conteúdo (texto em si), data e horário de publicação, quantidade de caracteres, quantidade de curtidas, quantidade de "re-tuitadas", é um *retweet*?, é uma resposta a outro tuíte?;
  - Os tuítes precisam ser publicados de 01 de janeiro de 2018 até no máximo 23:59 do dia 06 de outubro do mesmo ano (já que no dia 07 ocorreram as eleições e, portanto, publicações eram proibidas);
  - É necessária conexão estável com a Internet;
  - É necessário fornecer os códigos de acesso à API do *Twitter*;
  - Salvar o conteúdo em formatos .CSV e .TXT;
- Planejamento:
  - Duração esperada de três dias de trabalho;
  - Etapas:
    - Estabelecer conexão com API do *Twitter*;
    - Aquisição inicial: fazer um teste com qualquer usuário;
    - Aplicar filtragens na aquisição, conforme requisitos de datas e variáveis;
    - Criar função para salvar dados em .CSV;
    - Criar função para salvar dados em .TXT;
    - Verificar resultados obtidos e corrigir, se necessário.
- Modelagem: pode ser vista na Figura 6.2, que apresenta um pequeno diagrama do que o sistema deveria fazer:

Figura 6.2 - Modelo do *web scraper*

Fonte: arquivo pessoal

- Construção: criação do código em si. Para isso, foi usada a ferramenta *R* e as suas bibliotecas: *rtweet*, *data.table*, *stringr*, *readr* e *stringi*;
- Implantação: no caso desse projeto, a implantação foi feita ao executar o sistema para aquisição dos tuítes, conforme seu objetivo. Ela foi realizada diversas vezes, sendo a final delas no dia 08 de outubro, após a eleição e, portanto, todas as publicações de interesse já tinham sido feitas;

Com a aplicação do *web scraper*, obteve-se 7778 tuítes de 8 dos principais candidatos à Presidência. A Figura 6.3 apresenta uma tabela criada na própria ferramenta R com algumas dessas publicações. Nela, observamos as variáveis desejadas nas colunas e cada linha correspondente a um *tweet*.

Figura 6.3 - Captura de tela da tabela criada, no R, com a aquisição dos *tweets*

created_at	screen_name	text	favorite_count	retweet_count	display_text_width	is_retweet	is_quote	reply_to_status_id
2018-10-07 02:39:18	jairbolsonaro	– UM FORTE ABRAÇO A TODOS OS BRASILEIROS! – DEUS NOS AB...	51491	11245	156	FALSE	FALSE	NA
2018-10-07 01:45:33	Haddad_Fernan...	Conheça nosso Plano de Governo. #HaddadPresidente https://t....	3200	1356	74	FALSE	FALSE	NA
2018-10-07 01:19:28	joaoamoedonovo	Cada um de nós é o salvador que a pátria precisa. Vote sem me...	2708	466	91	FALSE	FALSE	NA
2018-10-07 00:57:56	MarinaSilva	O Brasil clama para que a gente pare com essa polarização que ...	1009	166	179	FALSE	FALSE	NA
2018-10-07 00:42:10	cirogomes	O Brasil quer CIRO PRESIDENTE! #ViraViraCiro #CiroSim #Ciro12...	17631	4232	62	FALSE	FALSE	NA
2018-10-07 00:34:48	meirelles	Está chegando a hora da votação. Tenho viajado o País e conver...	165	16	278	FALSE	FALSE	NA
2018-10-07 00:23:29	GuilhermeBoulos	Vice de Bolsonaro falou hoje que seu neto é um "cara bonito" po...	40100	13792	255	FALSE	FALSE	NA
2018-10-07 00:21:55	joaoamoedonovo	Pessoal, essas últimas pesquisas comprovam: haverá 2º turno. ...	15682	2403	172	FALSE	FALSE	NA
2018-10-07 00:18:49	jairbolsonaro	Última live de Bolsonaro antes das eleições do dia 7 de outubro ...	11320	2040	113	FALSE	FALSE	NA
2018-10-07 00:09:36	GuilhermeBoulos	#AoVivo: Boulos comenta resultados do IBOPE e DATAFOLHA – ...	204	23	93	FALSE	FALSE	NA

Fonte: arquivo pessoal

Entretanto, no decorrer da construção, uma mudança considerável ao planejamento foi imposta. Como os tuítes deveriam passar por uma etapa de pré-processamento, julgou-se mais eficiente realizar essa etapa **antes** de gerar os arquivos .CSV e .TXT. Caso contrário, após criar os arquivos, estes precisariam ser acessados e modificados para finalmente estarem em sua versão final. Sendo assim, a criação dos arquivos só será feita após o pré-processamento, o que será apresentado na Seção 6.3.

6.2 Identificação

Conforme apresentado na Seção 4.2.5, em um projeto de análise de sentimentos, há uma etapa de identificação em que alguns parâmetros devem ser determinados para a continuidade das atividades. Essas escolhas devem ser feitas conforme o conjunto de documentos a serem classificados, bem como de acordo com os objetivos do projeto.

A primeira decisão relevante do projeto foi quanto ao nível de análise (ou granularidade). Como os documentos eram *tweets*, os quais possuem número reduzido de caracteres (no máximo 280), optou-se pela análise em nível de documento. Ou seja, cada tuíte é analisado como um todo, sem dividi-lo por sentença ou tema - conforme sugerido por Souza, Pereira & Dalip (2017).

Outro parâmetro importante para realizar mineração de opiniões é determinar os diferentes alvos possíveis. No caso desse projeto, como o conteúdo analisado é publicado por candidatos à Presidência, as possibilidades de alvo são inúmeras: economia, segurança pública, saúde, histórico pessoal próprio (ou de um rival), corrupção, meio ambiente… Com isso, surgem duas dificuldades: a primeira delas é listar todos os alvos com exatidão, o que pode ser até considerado inviável, já que trata-se de milhares de tuítes; a segunda é que mesmo que esse levantamento seja feito, alguns temas teriam poucos *tweets* a respeito, impedindo que sejam tiradas conclusões concretas no decorrer da análise. Portanto, foi decidido por não listar os possíveis alvos, considerando todos os *tweets* como se tivessem o mesmo "assunto": a campanha presidencial como um todo. Formalmente falando, para a categorização, isso significa que a lista de alvos contém apenas um elemento *null*.

Dado esse alvo, a próxima parte é definir as categorias de sentimentos. Conforme apresentado na Seção 4.2.5, em um projeto de *text mining*, pode-se encontrar categorias binárias, contínuas ou discretas. Para esse PFC em questão, optou-se por avaliar todas as possibilidades de classificação para, a partir do conjunto de resultados obtidos, optar por uma que apresentasse os melhores resultados.

A penúltima decisão dessa etapa foi determinar as *features* dos documentos. Uma *feature* representa um atributo do documento que possui relevância para a categorização. Para esse projeto, ficaram definidas duas *features*: o autor do documento, bem como o texto publicado em si. Posteriormente, o texto vai ser dividido em *tokens* e, então, cada *token* será uma *feature* - mas isso será melhor explicado na próxima seção.

Por fim, a última etapa foi selecionar uma quantidade de *tweets* para usar como amostra para o aprendizado da máquina, quando este for necessário. Para isso, com a ajuda da ferramenta R, foram escolhidos de forma aleatória aproximadamente 10% dos *tweets* de cada candidato para categorização manual. Esses documentos foram categorizados em três sentimentos (positivo, negativo e neutro) por um especialista na área, que no caso foi um Professor do Departamento de Ciência Política da Universidade.

6.3 Pré-processamento

A partir da base de tuítes gerada com o *web scraper*, são aplicadas as técnicas de pré-processamento apresentadas na Seção 4.2.6, que são de suma importância para assegurar a devida qualidade da categorização. Uma parte do pré-processamento foi feito com a ferramenta R, enquanto o resto, com a ferramenta utilizada para categorização (*Weka* ou *Orange*).

6.3.1 Pré-processamento no R

Para facilitar a implementação, optou-se por fazer a limpeza dos textos como primeira parte do pré-processamento, ao invés da *tokenização*. Para isso, foram feitas as seguintes operações, em ordem: passar todos os caracteres para minúsculo, remover as expressões “RT” ou “via” (comumente utilizadas na rede social), remover marcações de outros usuários (que são iniciadas sempre pelo caractere ‘@’), remover links e URL’s, remover *emojis*, remover caracteres de pontuação e, por fim, remover numerais. No Quadro 2, apresenta-se exemplos de documentos que foram alterados.

Quadro 2 - Pré-processamento dos *tweets* em R

***Tweets* de entrada:** "Hoje, às 20:30 (06/10/2018), realizaremos a última live no facebook antes das eleições! Nos vemos lá!"; "Com grande alegria participei da convenção do @pps23, celebrando a nossa aliança. Temos agora o desafio de fazer o Brasil crescer, gerar empregos e melhorar a vida da população. Essa não é uma missão para uma pessoa só, é uma tarefa coletiva. Vamos juntos para a vitória!"; "A prisão de um ex-presidente é um acontecimento triste em qualquer país. No entanto, numa democracia, as decisões da Justiça devem ser respeitadas por todos e aplicadas igualmente para todos."
***Tweets* de saída:** “hoje realizaremos última live facebook eleições vemos lá”; “grande alegria participei convenção celebrando aliança agora desafio fazer brasil crescer gerar empregos melhorar vida população missão pessoa tarefa coletiva vamos juntos vitória”; “prisão expresidente acontecimento triste país entanto democracia decisões justiça devem respeitadas aplicadas igualmente”

Fonte: arquivo pessoal

Após esse trabalho, foi feita a remoções das *stopwords*. Foram encontradas e juntadas três listas de *stopwords* em português, totalizando em um conjunto de 269 palavras.

A Figura 6.4 apresenta os primeiros 14 termos da lista *sw_merged*, em ordem alfabética. Com a criação feita, o próximo passo foi remover esses termos dos *tweets* dos candidatos.

Figura 6.4 - Termos presentes na lista de *stopwords*

1	a
2	ainda
3	alem
4	ambas
5	ambos
6	antes
7	ao
8	aonde
9	aos
10	apos
11	aquele
12	aqueles
13	as
14	assim

Fonte: arquivo pessoal

Ao final dessa etapa, os arquivos em formato .CSV e .TXT foram criados. A Figura 6.5 apresenta uma parte do arquivo .CSV gerado, com as primeiras amostras de *tweets* de um candidato e algumas das variáveis desejadas. As demais etapas do pré-processamento, como a *tokenização, stemmização* e atribuição numérica, foram feitas em outras ferramentas.

Figura 6.5 - Parte de arquivo em .CSV com *tweets* pré-processados

	A	B	C	D	E	F
1		created_at	screen_nam	text	favorite_count	retweet_count
2	1	10/7/18 02:39	jairbolsonarc	forte abravßo brasileiros! deus abenvßoe! jair bolsonaro link ca	51856	11374
3	2	10/7/18 00:18	jairbolsonarc	vʃltima live bolsonaro eleivßvµes dia outubro link youtube	11421	2057
4	3	10/6/18 23:07	jairbolsonarc	hoje realizaremos vʃltima live facebook eleivßvµes! vemos lv°!	24637	3709
5	4	10/6/18 22:15	jairbolsonarc	obrigado consideravßv£o guerreiro campev£o! forte abravßo	45602	9594
6	5	10/6/18 21:15	jairbolsonarc	obrigado apoio hashtag euvotobolsonaro vamos juntos mudar desti	83128	19004
7	6	10/6/18 20:47	jairbolsonarc	ratinho fala candidato senado paulo acesso candidatos brasil a	17942	4037
8	7	10/6/18 20:40	jairbolsonarc	hoje manhv£ carreata brasv≠lia forte abravßo distrito federal!	15590	2952
9	8	10/6/18 17:09	jairbolsonarc	fortes poucos recursos acordvµes tempo tv impossibilitado fazer c	58238	13737
10	9	10/6/18 17:08	jairbolsonarc	vencermos comevßamos diferentes livres escolher equipe critv©	41563	10271
11	10	10/6/18 17:08	jairbolsonarc	capazes reconhecer erros limitavßvµes enxergar potencial brasi	26559	6168
12	11	10/6/18 17:08	jairbolsonarc	tempo brasileiro escolher opvßvµes representava agora diferent	43591	10703
13	12	10/6/18 16:14	jairbolsonarc	jair bolsonaro soldado adv£o grupo artilharia campanha nioaque	28862	5411

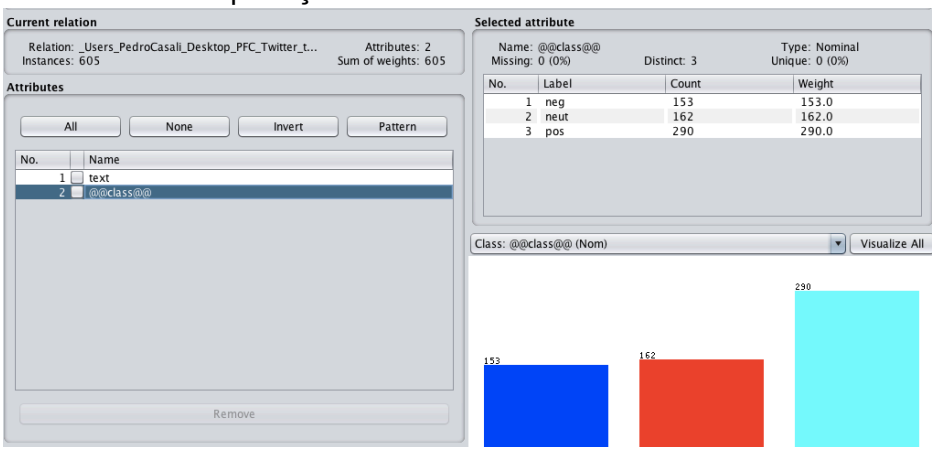
Fonte: arquivo pessoal

6.3.2 Pré-processamento no *Weka*

Um dos *softwares* usados para a categorização foi o *Weka*, conforme apresentado na Seção 5.3.2. Seu uso foi para avaliar os diferentes algoritmos de aprendizado de máquina. Portanto, para utilizá-lo, foi necessário importar em formato .TXT os arquivos dos *tweets* “limpos” no R. Contudo, apenas a amostra de 10% dos documentos foi importada, pois é a utilizada na construção dos modelos. Além disso, foram considerados os conteúdos de apenas seis dos oito candidatos (totalizando 605 documentos), para reduzir o tempo de processamento do *software*.

A Figura 6.6 apresenta o resultado da importação dos arquivos no *Weka*. Nota-se, na imagem, a informação de que 605 *tweets* foram importados, sendo que 153 com sentimento negativo, 162 neutro e 290 positivo.

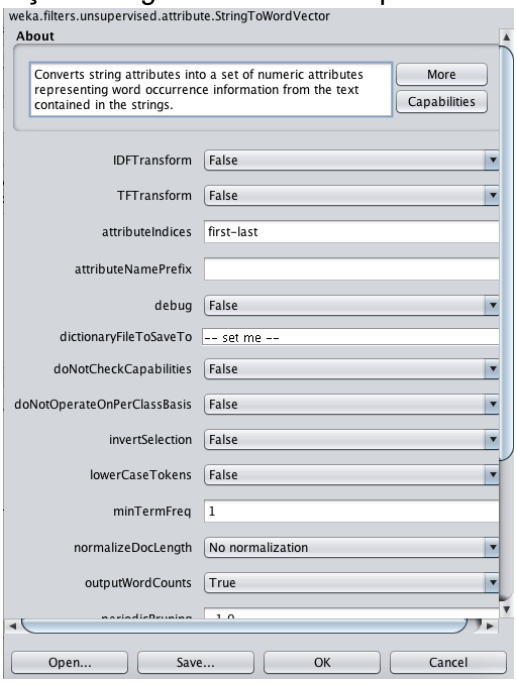
Figura 6.6 - Resultado da importação dos *tweets* no *Weka*



Fonte: arquivo pessoal

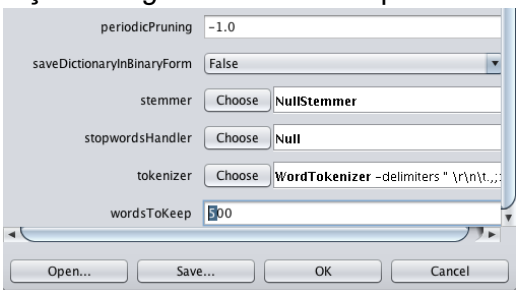
Em seguida, foi necessário construir a matriz *Bag of Words* (BoW). Para isso, o *Weka* possui uma função chamada *StringToWordVector*, cujo objetivo é dividir os diferentes documentos em vetores de palavras. Ao final, cada elemento desse vetor será considerado um *feature* (ou atributo). As Figuras 6.7 e 6.8 apresentam a configuração dessa função. Primeiramente, optou-se por fazer uma matriz que indicasse a quantidade de vezes que o termo T_j aparece no documento D_i, sem aplicar a TF-IDF. Para isso, os parâmetros *IDFTransform* e *TFTranform* foram definidos como *False*, enquanto o *outputWordCount*, como *True*. Além disso, a escolha foi de *tokenizar* os documentos em unigramas, ou seja, por palavras (parâmetro *Tokenizer* definido como *WordTokenizer*). Por fim, não foi possível adicionar *stemmização* ou *lemmatização*, já que o Weka só consegue fazê-lo para textos em inglês (parâmetro *stemmer* escolhido como *NullStemmer*).

Figura 6.7 - Configuração da função *StringToWordVector* – parte 1



Fonte: arquivo pessoal

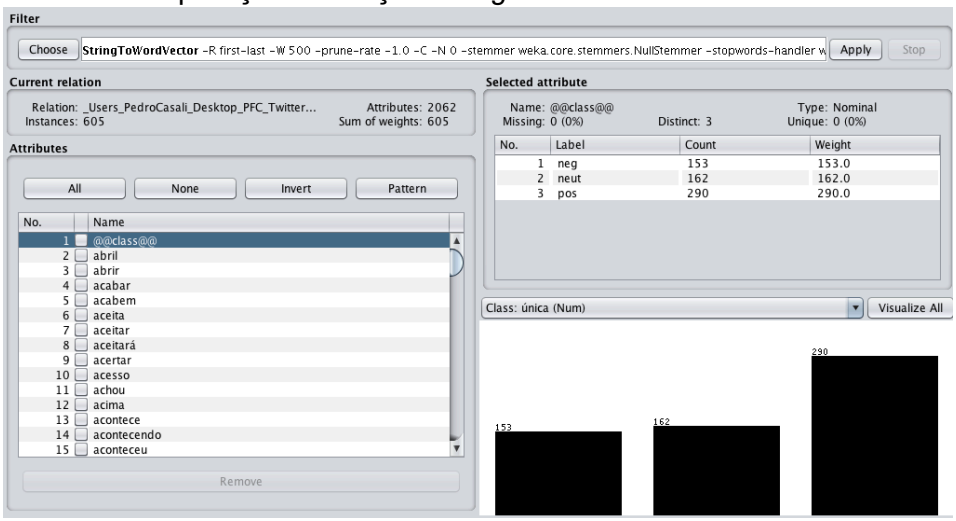
Figura 6.8 - Configuração da função *StringToWordVector* – parte 2



Fonte: arquivo pessoal

O resultado da aplicação dessa função pode ser observado nas imagens a seguir. Primeiramente, na Figura 6.9, observa-se que o processo resultou na identificação de 2061 palavras distintas (a categoria dos *tweets*, dada pelo atributo @@@*classs*@@@, é o 2062° atributo).

Figura 6.9 - Resultado da aplicação da função *StringToWordVector*



Fonte: arquivo pessoal

Entretanto, @@@*classs*@@@ é, na verdade, o conjunto de classes que os algoritmos deveriam utilizar como base, e não um atributo, conforme indicado na Figura acima. A Figura 6.10 apresenta a matriz *BoW* resultante após essa correção. Nela, observa-se a contagem dos primeiros *features* (em ordem alfabética) nos primeiros documentos. Nota-se que a palavra “acabar” está presente uma vez no Documento 22, enquanto “aceitar” está uma vez no Documento 14.

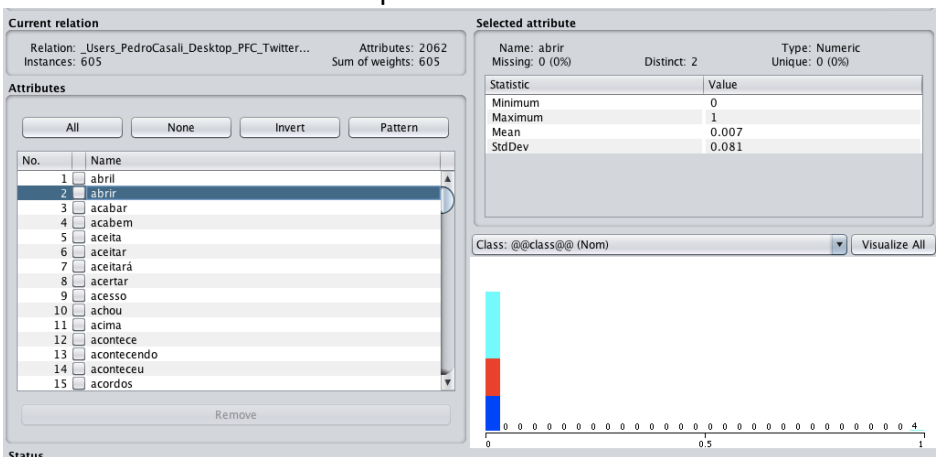
Figura 6.10 - Matriz *BoW* booleana gerada

No.	1: abril	2: abrir	3: acabar	4: acabem	5: aceita	6: aceitar	7:
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	
1	0.0	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	
5	0.0	0.0	0.0	0.0	0.0	0.0	
6	0.0	0.0	0.0	0.0	0.0	0.0	
7	0.0	0.0	0.0	0.0	0.0	0.0	
8	0.0	0.0	0.0	0.0	0.0	0.0	
9	0.0	0.0	0.0	0.0	0.0	0.0	
...	0.0	0.0	0.0	0.0	0.0	0.0	
...	0.0	0.0	0.0	0.0	0.0	0.0	
...	0.0	0.0	0.0	0.0	0.0	0.0	
...	0.0	0.0	0.0	0.0	0.0	0.0	
...	0.0	0.0	0.0	0.0	0.0	0.0	
...	0.0	0.0	0.0	0.0	0.0	1.0	
...	0.0	0.0	0.0	0.0	0.0	0.0	
...	0.0	0.0	0.0	0.0	0.0	0.0	
...	0.0	0.0	0.0	0.0	0.0	0.0	
...	0.0	0.0	0.0	0.0	0.0	0.0	
...	0.0	0.0	1.0	0.0	0.0	0.0	

Fonte: arquivo pessoal

Outra maneira de analisar cada *feature* é como demonstrado na Figura 6.11, que apresenta um pouco mais sobre a palavra "*abrir*": ela está presente no máximo uma vez (perceptível no canto superior direito da imagem, com o campo *Maximum*) e isso ocorre quatro vezes (conforme o gráfico de colunas) e, sempre que aparece, o sentimento do *tweet* é positivo (novamente indicado no gráfico, com a cor azul clara).

Figura 6.11 - Análise de *feature* no *Weka*: tipo 1

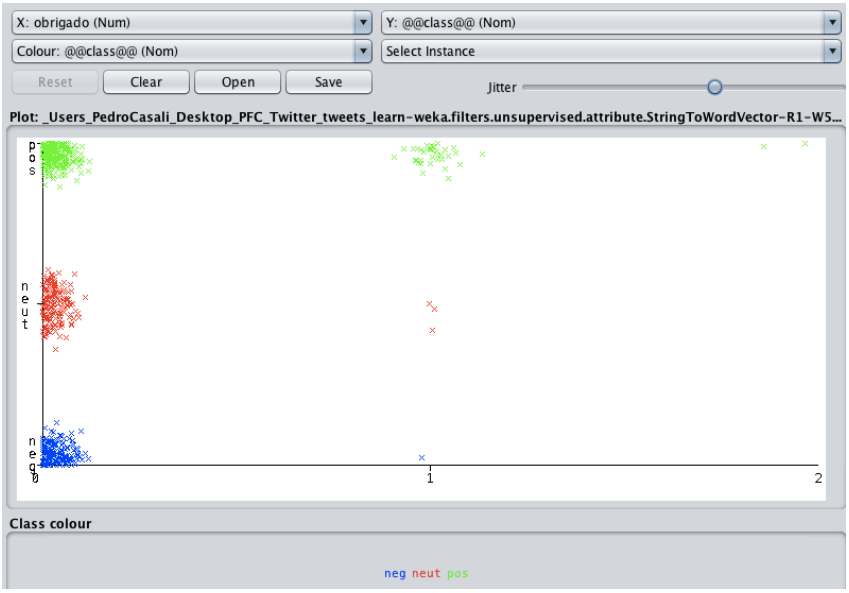


Fonte: arquivo pessoal

O *Weka* também permite visualizar essa disposição conforme exibido nas Figuras 6.12 e 6.13. Nelas, o eixo horizontal apresenta a quantidade de vezes que o termo aparece em um documento (*tweet*). O eixo vertical, por sua vez, apresenta a quantidade de *tweets* com cada sentimento: em verde, os positivos; em vermelho, os neutros: e em azul, os negativos. A primeira demonstra informações sobre a palavra "*obrigado*". Se ela não está presente em um documento, é difícil avaliar qual sentimento é mais provável que ele transmita. Entretanto, quando usada uma vez no *tweet*, houve um caso em que o sentimento foi negativo, três em que foi neutro e dezenas em que foi positivo. As informações da palavra "*milhões*" são apresentadas

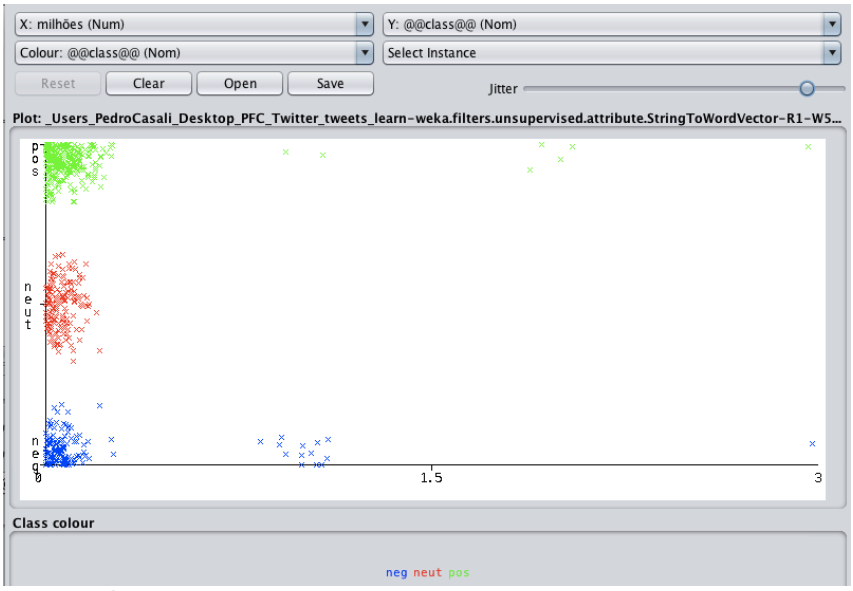
em seguida, em que outras análises podem ser feitas da mesma maneira: se ela aparece uma vez, há mais chance de o sentimento ser negativo.

Figura 6.12 - Análise do termo "obrigado" no *Weka*



Fonte: arquivo pessoal

Figura 6.13 - Análise do termo "milhões" no *Weka*



Fonte: arquivo pessoal

Após algumas simulações de classificação, notou-se que qualquer algoritmo que fosse utilizado obtinha melhor desempenho (maior acurácia) se a atribuição numérica aos elementos fosse feita utilizando TF-IDF. Isso faz sentido, já que essa técnica “premia” termos que aparecem mais em um documento e também termos que pouco aparecem no *corpus* como um todo. Para incluí-la, foi necessário alterar para *True* os campos *TFTransform* e *IDFTransform* apresentados na Figura 6.7.

A Figura 6.14 abaixo apresenta a matriz *BoW* já com a aplicação da técnica TF-IDF. Nota-se que os valores que antes eram iguais a 1 foram alterados para 3.478 e 4.439, o que demonstra que, para os documentos em que estão presentes, a palavra “aceitar” tem mais relevância do que “acabar”.

Figura 6.14 – Matriz *BoW* gerada no *Weka* com TF-IDF

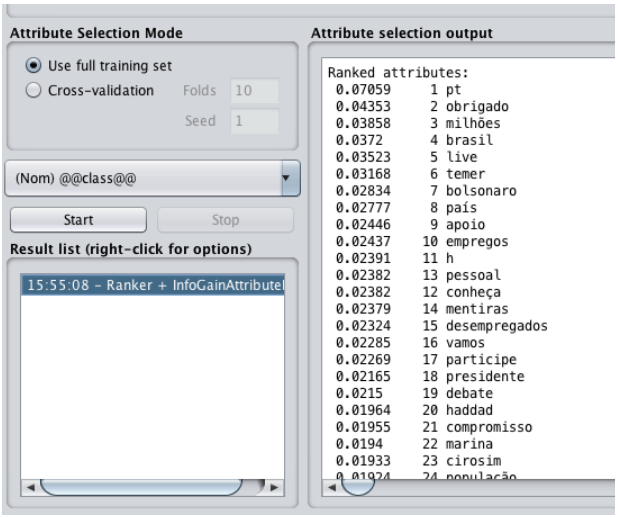
Relation: _Users_PedroCasali_Desktop_PFC_Twitter_tweets_learn-v

No.	1: abril Numeric	2: abrir Numeric	3: acabar Numeric	4: acabem Numeric	5: aceita Numeric	6: aceitar 7: Numeric
1	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0
...	0.0	0.0	0.0	0.0	0.0	0.0
...	0.0	0.0	0.0	0.0	0.0	0.0
...	0.0	0.0	0.0	0.0	0.0	0.0
...	0.0	0.0	0.0	0.0	0.0	0.0
...	0.0	0.0	0.0	0.0	0.0	0.0
...	0.0	0.0	0.0	0.0	0.0	0.0
...	0.0	0.0	0.0	0.0	0.0	4.439...
...	0.0	0.0	0.0	0.0	0.0	0.0
...	0.0	0.0	0.0	0.0	0.0	0.0
...	0.0	0.0	0.0	0.0	0.0	0.0
...	0.0	0.0	0.0	0.0	0.0	0.0
...	0.0	0.0	3.478...	0.0	0.0	0.0

Fonte: arquivo pessoal

Outra ação feita no pré-processamento no *Weka* foi através da função *AttributeSelect*, cujo objetivo é selecionar apenas algumas das *features* para serem usadas na categorização. Para a escolha dos termos, foi feito um ranqueamento (*Ranker*) de cada um com base na avaliação do ganho de informações deles (*InfoGainAttributeEval*). A Figura 6.15 apresenta parte do resultado desse ranqueamento. Nota-se que os termos “*pt*”, “*obrigado*” e “*milhões*” contribuem com mais conteúdo para a categorização. Foi estabelecido como 0.008 o valor mínimo a ser considerado pelo *Ranker*, o que resultou em utilizar apenas 123 *features*. Este valor é o mínimo permitido pelo *Weka*.

Figura 6.15 – Ranqueamento de termos no *Weka*

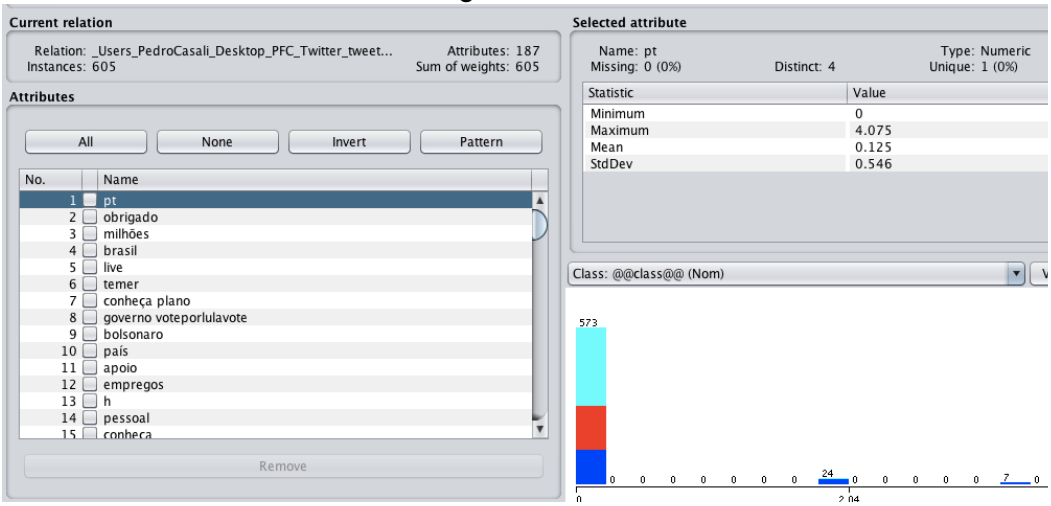


Fonte: arquivo pessoal

Entretanto, como o número de *features* restantes foi considerado pequeno, decidiu-se por também incluir os bigramas (conjunto de qualquer duas palavras) com maior relevância. Para isso, altera-se o campo *Tokenizer* (ver Figura 6.8) para usar a técnica de *N-Grams*. Determina-se que devem ser considerados os unigramas (*NGramMinSize = 1*) e os bigramas (*NGramMaxSize = 2*).

Ao incluir a mesma técnica de ranqueamento apresentada anteriormente, obtém-se 187 *tokens* relevantes, incluindo o bigrama "*conheça plano*", que é a sétima *feature* mais importante segundo o ranqueador, conforme apresentado na Figura 6.16.

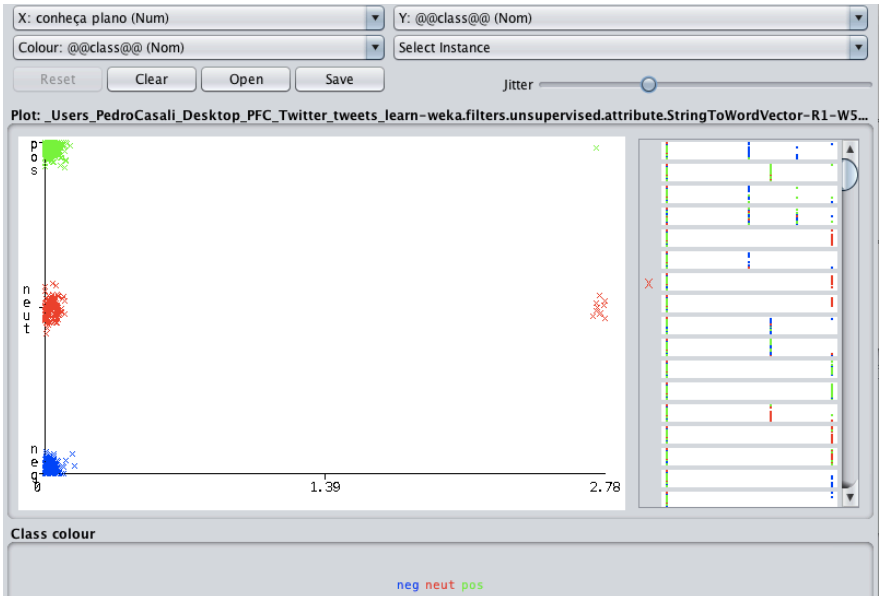
Figura 6.16 – Resultado da inclusão de bigramas



Fonte: arquivo pessoal

Para entender o porquê da relevância de "*conheça plano*", observa-se sua análise na Figura 6.17. Quando este *token* está presente em um documento, a probabilidade é alta de que este tenha sentimento neutro.

Figura 6.17 - Análise do bigrama "conheça plano" no *Weka*



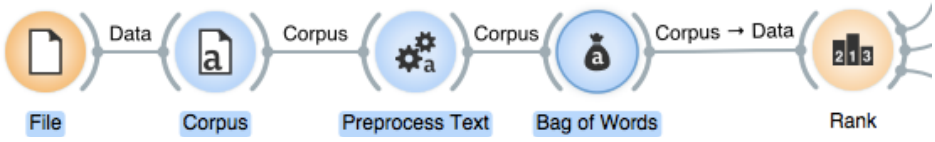
Fonte: arquivo pessoal

Portanto, no *Weka*, a categorização foi feita com as seguintes configurações no pré-processamento: **unigramas e bigramas, sendo considerados apenas aqueles com maior relevância; sem *stemmização* ou *lemmatização*; e atribuição numérica com a técnica TF-IDF**.

6.3.3 Pré-processamento no *Orange*

A ferramenta *Orange* foi apresentada na Seção 5.3.3. Neste trabalho, seu uso foi na construção do modelo de categorização e também na aplicação deste modelo aos 90% de *tweets* restantes para posterior visualização e análise. Portanto, o objetivo do pré-processamento no *Orange* era de replicar com a maior precisão possível o que foi feito no *Weka*. Para isso, foi necessário importar um arquivo em formato .CSV (criado pelo pré-processamento em R) com a amostra pré-avaliada de 10% dos conteúdos publicados pelos candidatos. Ao total, foram 800 publicações dos 8 candidatos escolhidos. A Figura 6.18 apresenta a parte do diagrama de blocos que contém o pré-processamento.

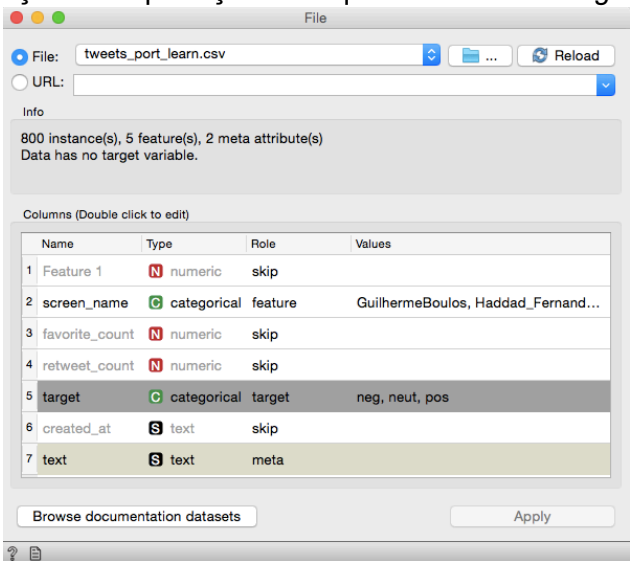
Figura 6.18 - Diagrama de blocos do pré-processamento no *Orange*



Fonte: arquivo pessoal

Logo na janela de configuração da importação, algumas decisões foram necessárias, conforme apresentado na Figura 6.19. Ao contrário do *Weka*, o *Orange* possui uma interface amigável e que permite escolher com facilidade as variáveis a serem consideradas como *features*. Além do conteúdo (texto) em si, foi incluída como atributo relevante a variável *screen_name*, que indica o usuário (candidato) que o escreveu. Posteriormente, essa escolha será justificada ao apresentar o ranqueamento das *features*.

Figura 6.19 – Configuração da importação do arquivo .CSV no *Orange*

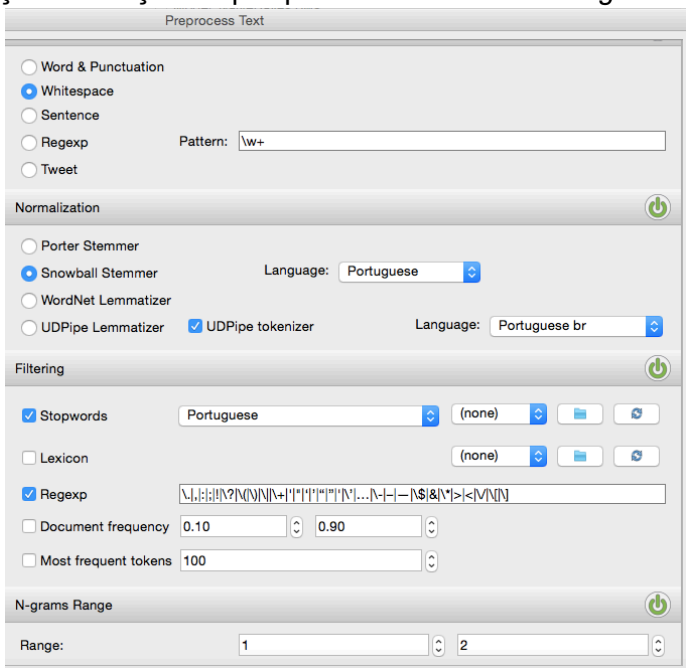


Fonte: arquivo pessoal

Uma vez a importação sendo feita, foi necessário indicar que os dados, na verdade, pertenciam a um *corpus*, através da função que recebe esse nome.

Em seguida, com o uso da função de pré-processamento de texto do próprio *Orange*, foram feitas algumas ações. A configuração dessa funcionalidade é apresentada na Figura 6.20. Primeiramente, foi aplicada uma técnica de *stemmização* com uso do método *Snowball* para a Língua Portuguesa, o que não era disponível no *Weka*. Optou-se por ela pois, com isso, obtiveram-se resultados ainda melhores nos testes de categorização. Além disso, mais uma etapa de limpeza foi realizada para garantir que esta esteja bem feita, com a remoções de *stopwords* (o *Orange* possui sua própria lista desses termos) e de pontuações. Por fim, na aba *N-Grams Range*, foi indicado para que fossem levados em conta os unigramas e também os bigramas, assim como feito no *Weka*.

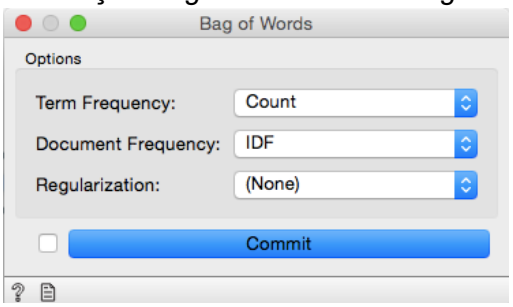
Figura 6.20 – Configuração da função de pré-processamento no *Orange*



Fonte: arquivo pessoal

Depois, foi feita a construção da matriz *Bag of Words*. A função para isso no *Orange* permite ao usuário determinar três configurações: frequência do termo, frequência em documentos e regularização, conforme pode ser visto na Figura 6.21, a qual já apresenta a configuração feita para o projeto.

Figura 6.21 – Configuração da função *Bag of Words* no *Orange*



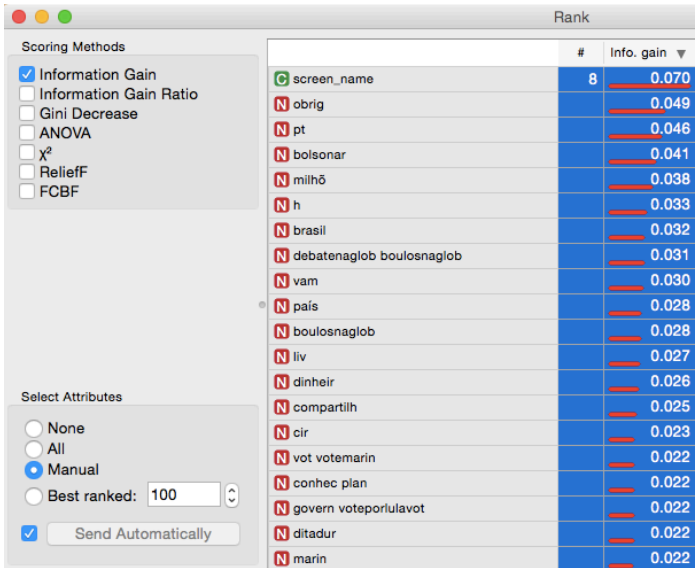
Fonte: arquivo pessoal

Para a frequência do termo (ou TF), três opções são possíveis (todas descritas anteriormente): binária, contagem ou sublinear, todas diferentes do utilizado no *Weka*. Neste trabalho, optou-se pela contagem, por apresentar melhores resultados nas simulações de categorização. Em seguida, para a frequência em documentos, as opções são novamente três: aplicar a IDF, aplicar *Smooth IDF* (similar à IDF, mas adiciona 1 no denominador para não haver divisão por zero) e sem aplicação alguma. Para replicar o obtido no *Weka*, optou-se pela aplicação da IDF tradicional - o que também gerou em melhor acurácia nas previsões. Com isso, a atribuição numérica resultante foi a multiplicação desses dois valores.

Por fim, a regularização consiste em selecionar apenas os pontos relevantes para a categorização, procurando evitar *underfitting* (considerar informação “de menos”) ou *overfitting* (considerar informação “de mais”, como ruidos e anomalias). Optou-se por não realizar a regularização dos documentos e fazer essa “filtragem” apenas com o ranqueamento, etapa a ser descrita a seguir.

A última parte do pré-processamento, assim como feito no *Weka*, foi realizar a seleção de *features* com base no seu ranqueamento de relevância à categorização. A Figura 6.22 apresenta a sua configuração. Utilizando-se do mesmo critério de avaliação do *Weka* (*Information Gain*), e estabelecendo um *threshold* mais baixo (de 0.007), resultou-se em considerar 360 *features*. Percebe-se na imagem que o atributo de maior relevância na categorização foi justamente a variável *screen_name*, o que mostra o porquê de inclui-la. Em compensação, as variáveis de curtidas (*favorite_count*) e compartilhamentos (*retweet_count*) não agregaram conteúdo o suficiente para estarem acima do limite.

Figura 6.22 - Resultado do ranqueamento de termos no *Orange*



Fonte: arquivo pessoal

O resultado final do pré-processamento no *Orange* é apresentado na Figura 6.23. A coluna "*{...}*" indica os valores das *features* presentes no documento indicado pela linha, como se fosse a composição das linhas de uma matriz *BoW*. Nela, é possível notar que muitos dos *tokens* estão sem a terminação (o que mostra a aplicação da *stemmização*), bem como alguns deles são formados por dois termos (como "*fort abrac*").

Figura 6.23 - Resultado final do pré-processamento no *Orange*

bow-feature hidden skip-normalization	target	text	{...}
1	pos	forte abraço...	screen_name=jairbolsonaro, bolsonar=2.647, brasil=1.634, abrac=4.300, fort abrac=4.615
2	neut	última live ...	screen_name=jairbolsonaro, bolsonar=2.647, liv=4.300, youtub=5.303, últim=3.714, últim liv..
3	neut	hoje realiz...	screen_name=jairbolsonaro, liv=4.300, hoj=2.632, liv facebook=4.900, facebook=4.900, hoj ..
4	pos	obrigado c...	screen_name=jairbolsonaro, obrig=3.045, abrac=4.300, fort abrac=4.615
5	pos	obrigado ...	screen_name=jairbolsonaro, obrig=3.045, vam=1.923, apoi=3.320, junt=3.172, vam junt=3.9..
6	neut	ratinho fala ...	screen_name=jairbolsonaro, brasil=1.634, sen=4.394
7	pos	hoje manhã...	screen_name=jairbolsonaro, hoj=2.632, abrac=4.300, fort abrac=4.615
8	neg	fortes pouc...	screen_name=jairbolsonaro, brasil=1.634, vam=1.923, apoi=3.320, sofr=4.900, campanh=3...
9	pos	vencermos ...	screen_name=jairbolsonaro, país=2.322, comec=3.621, govern=2.364
10	pos	capazes re...	screen_name=jairbolsonaro, brasil=1.634, país=2.322, conhec=3.320, problem=4.063, grand.
11	pos	tempo bras...	screen_name=jairbolsonaro, bolsonar=2.647, brasil=1.634, corrupçã=3.932, represent=3.93..
12	neut	jair bolsona...	screen_name=jairbolsonaro, bolsonar=2.647, campanh=3.229
13	neut	live hoje b...	screen_name=jairbolsonaro, bolsonar=2.647, liv=4.300, hoj=2.632
14	pos	chegando ...	screen_name=jairbolsonaro, govern=2.364, cheg=2.833, grand=3.119, sistem=3.816, brasil!..
15	neut	entrevista r...	screen_name=jairbolsonaro, entrev=4.300
16	neg	brasil gigan...	screen_name=jairbolsonaro, brasil=1.634, popul=3.387, junt=3.172, merec=5.081, govern=2..
17	neg	imagens im...	screen_name=jairbolsonaro, juiz=4.394
18	neut	live hoje	screen_name=jairbolsonaro, liv=4.300, hoj=2.632
19	neut	hoje out qui...	screen_name=jairbolsonaro, hoj=2.632, divulg=3.764, hor=3.577, assist=4.215, estar=3.872

Fonte: arquivo pessoal

Uma vez que o pré-processamento foi concluído, a próxima etapa do projeto era a aplicação e avaliação de diferentes métodos de categorização, conforme apresentados na Seção 4.2.6, para escolher um a ser utilizado na construção do modelo final. Essa atividade será descrita na próxima seção.

6.4 Classificação

De acordo com o apresentado na Seção 4.2, existem diversas técnicas na literatura de mineração de opiniões para realizar a classificação dos documentos presentes no *corpus*. Para este projeto, foram avaliadas aquelas pertencentes a duas grandes áreas (com léxicos e com aprendizado de máquina), bem como alguns métodos próprios da ferramenta *Orange*. Ao final desta seção, será explicado o processo de escolha para o método escolhido para construção do modelo.

6.4.1 Classificação baseada em dicionários

As primeiras tentativas de categorização foram feitas com uso de dicionários (ou léxicos) através da ferramenta R. Para isso, o computador deve procurar na lista de documentos os termos avaliados no dicionário e, por meio da soma das polaridades das palavras identificadas, calcular a polaridade total do texto. Portanto, com isso, os *tweets* são classificados por meio de notas, em que números negativos representam documentos com sentimento negativo e números positivos, sentimentos positivos. Além disso, quanto maior o valor absoluto da nota, maior a intensidade do sentimento.

Inicialmente, foram testados os resultados com dicionários genéricos fornecidos por bibliotecas da própria ferramenta. Em seguida, foi construído um dicionário próprio para ser usado como léxico.

6.4.1.1 Dicionários genéricos

A maneira mais simples de categorizar documentos de forma automatizada é com uso de dicionários genéricos, cuja aplicação independe do contexto. A ferramenta R possui dois exemplos desse tipo de léxicos. Um, chamado *OpLexicon*, possui mais de 32 mil termos avaliados e os primeiros (em ordem alfabética) estão demonstrados na Figura 6.24; o outro, chamado *SentiLex*, pouco mais de 7 mil, e suas primeiras palavras são exibidas na Figura 6.25.

Figura 6.24 – Termos do dicionário *OpLexicon*

term	type	polarity
ab-rogar	vb	-1
ababadar	vb	0
ababelar	vb	-1
ababelar-se	vb	1
abacanar	vb	1
abacinar	vb	1
abafada	adj	-1
abafadas	adj	-1
abafado	adj	-1

Fonte: arquivo pessoal

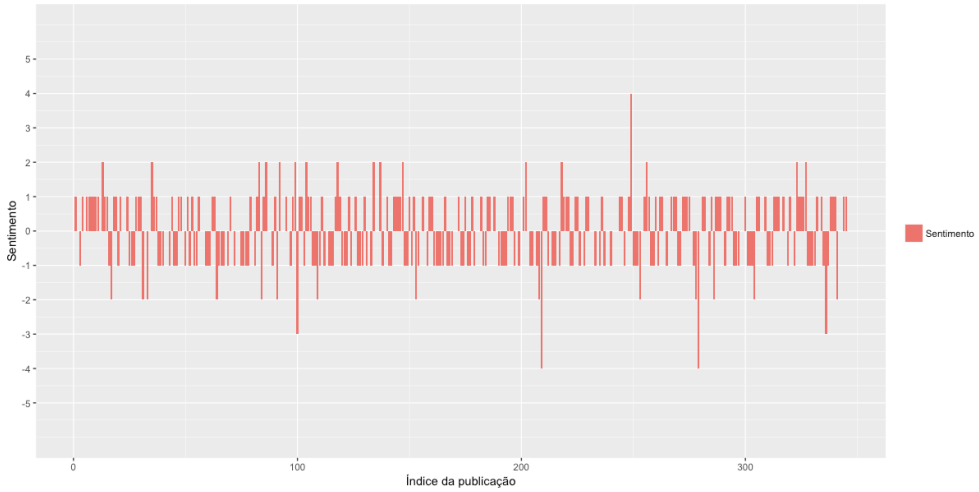
Figura 6.25 – Termos do dicionário *SentiLex*

term	grammar_category	polarity
a-vontade	N	1
abafado	Adj	-1
abafante	Adj	-1
abaixado	Adj	-1
abalado	Adj	-1
abalizado	Adj	1
abalroado	Adj	-1

Fonte: arquivo pessoal

Após a importação dos dicionários no *R*, a partir dos *tweets* pré-processados em R (ou seja, sem normalização, bigramas ou ranqueamento), foi feita a procura dos termos pré-avaliados em cada documento. O primeiro dicionário avaliado foi o *OpLexicon*. A Figura 6.26 apresenta as notas dadas às 758 publicações de um dos candidatos.

Figura 6.26 – Avaliação dos *tweets* de um candidato com uso do *OpLexicon*



Fonte: arquivo pessoal

De imediato, notam-se alguns problemas com essa avaliação. Primeiramente, a grande maioria dos *tweets* não receberam nota alguma: menos da metade do total (quase 350) receberam alguma nota e estão presentes no gráfico. Além disso, dentre os avaliados, quase todos possuem notas entre -1 e 1, sendo poucos aqueles que demonstram diferentes intensidades. Além disso, apesar do dicionário usado possuir uma grande quantidade de palavras, são poucas as que estão presentes no *corpus*. Uma diferença notável é a palavra “corrupção”, ausente dos dois dicionários, conforme apresentado na Figura 6.27.

Figura 6.27 – Avaliação do termo “corrupção” nos diferentes dicionários genéricos

```
> get_word_sentiment("corrupção")
$oplexicon_v2.1
[1] "Word not present in dataset"

$oplexicon_v3.0
[1] "Word not present in dataset"

$sentilex
[1] "Word not present in dataset"
```

Fonte: arquivo pessoal

O mesmo foi notado com os demais candidatos e seus *tweets*. A Tabela 1 apresenta a quantidade publicada por cada um e quantos desses foram realmente avaliados.

Tabela 1 – Relação entre *tweets* publicados e avaliados com léxico próprio

Candidato	Publicações	Avaliados
1	1187	582
2	1316	663
3	758	346
4	1716	790
5	813	271
6	943	345
7	660	288
8	385	177

Fonte: arquivo pessoal

Além desses problemas, os dois dicionários apresentam contradições em relação a certos termos. A Figura 6.28 apresenta um exemplo. A palavra “temer”, por exemplo, uma das mais ditas pelos candidatos, é avaliada como positiva no *OpLexicon*, mas negativa no *SentiLex*.

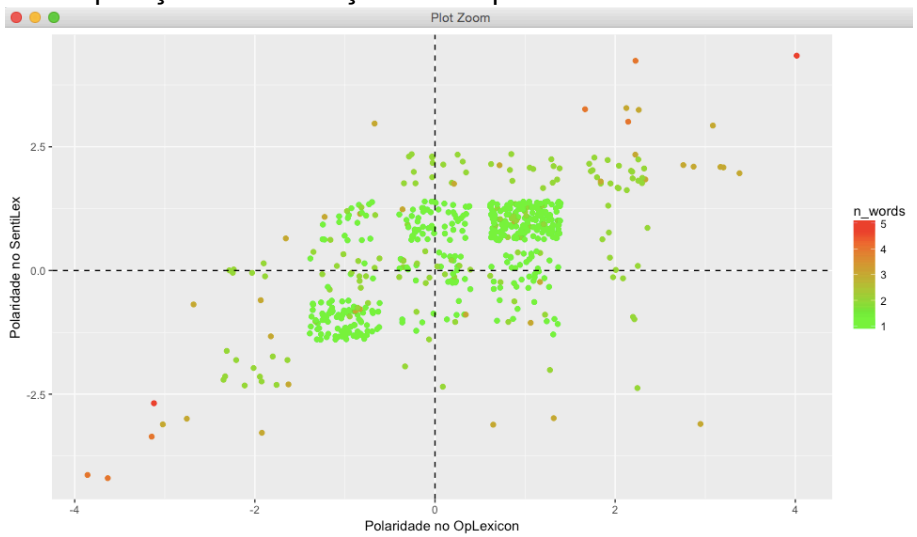
Figura 6.28 – Avaliação do termo “temer” nos diferentes dicionários

```
> get_word_sentiment("temer")
$oplexicon_v2.1
       term type polarity
28711 temer   vb        1

$oplexicon_v3.0
       term type polarity polarity_revision
30160 temer   vb        1                 A

$sentilex
      term grammar_category polarity polarity_target polarity_classification
6546 temer               V       -1          N0:N1                       MAN
```

Fonte: arquivo pessoal

Para avaliar a gravidade desses problemas, foram elaborados gráficos comparativos entre os dicionários para todos os candidatos. A Figura 6.29 apresenta a análise para um dos postulantes, em que cada ponto é um *tweet*. Nela, o valor do eixo y representa o sentimento com uso do *Sentilex*; já no eixo x, o sentimento através do *OpLexicon*; a cor de cada ponto indica a quantidade de palavras do documento que estão presentes no *OpLexicon*.

Figura 6.29 – Comparação entre avalições feitas por dois dicionários



Fonte: arquivo pessoal

A presença de alguns pontos no segundo e no quarto quadrantes indicam a presença de contradições. Ou seja, enquanto para um dicionário, o sentimento de um *tweet* é positivo, para o outro, é negativo. Esse comportamento, novamente, foi observado com todos os candidatos.

Com a comprovação desse conjunto de pontos negativos observados com o uso de dicionários genéricos, optou-se pela construção de um dicionário próprio a ser usado na categorização.

6.4.1.2 Dicionário próprio

Visando obter um resultado melhor com uma classificação com léxicos, optou-se pela construção de um dicionário próprio com vocabulário mais adequado para o contexto dos documentos: a campanha presidencial de 2018. Para isso, foram criados dois arquivos em formato .TXT, um contendo termos com sentimento positivo e o outro, negativo. Sua construção foi feita, novamente, com auxílio de um Professor do Departamento de Ciência Política da Universidade. Ao total, foram levantadas 345 palavras positivas e 309 negativas, totalizando apenas 654 termos. Entretanto, apesar da baixa quantidade de vocábulos avaliados (menos de 10% dos dicionários genéricos), os resultados se mostraram melhores.

Feita em R, o processo de classificação iniciou, assim como no caso anterior, com a importação do léxico para a ferramenta. Após a leitura dos dois documentos (com a devida codificação, para conseguir identificar caracteres com acentos), foi preciso juntar as expressões em uma tabela única e verificar a existência de termos repetidos.

A Figura 6.30 exibe os primeiros termos do léxico, em ordem alfabética, após ser importado na ferramenta.

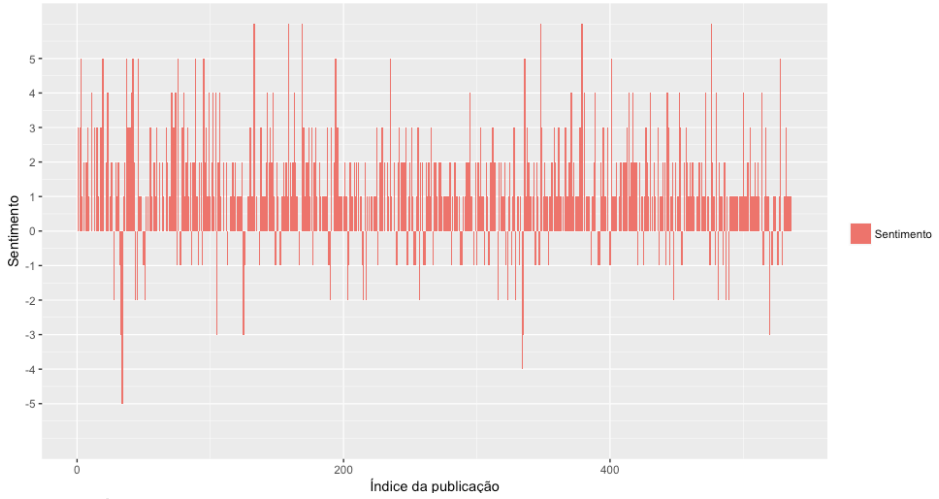
Figura 6.30 – Exemplos de termos de léxico criado

term	polaridade	tipo	sentiment
a favor	1	palavra	positivo
aberta	1	palavra	positivo
abertas	1	palavra	positivo
aberto	1	palavra	positivo
abertos	1	palavra	positivo
abertura	1	palavra	positivo
aborrecente	-1	palavra	negativo

Fonte: arquivo pessoal

Em seguida, utilizando do mesmo algoritmo para categorização com dicionários genéricos, foi feita a avaliação dos *tweets* do mesmo candidato avaliado na Figura 6.26. A Figura 6.31 apresenta os resultados obtidos. Percebe-se que mais *tweets* foram avaliados e com mais variações de intensidade se comparado ao caso anterior.

Figura 6.31 - Avaliação dos *tweets* de um candidato com uso de léxico próprio (1)



Fonte: arquivo pessoal

A Tabela 2 apresenta a relação de *tweets* publicados pelos candidatos com a quantidade avaliada por esse novo método.

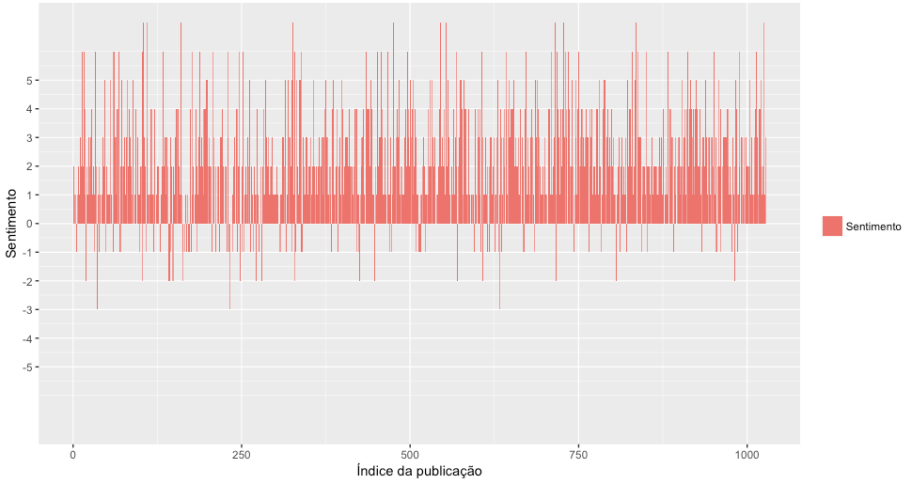
Tabela 2 – Relação entre *tweets* publicados e avaliados com léxico próprio

Candidato	Publicações	Avaliados
1	1187	1053
2	1316	1212
3	758	562
4	1716	1334
5	813	527
6	943	573
7	660	427
8	385	303

Fonte: arquivo pessoal

Entretanto, ao analisar o conteúdo de todos os candidatos, percebe-se que, em geral, foram identificadas mais publicações com sentimentos positivos do que negativos, apesar de uma quantidade aproximada de termos em cada lista. A Figura 6.32 apresenta a avaliação feita com outro candidato, repetindo o mesmo padrão da Figura 6.31.

Figura 6.32 - Avaliação dos *tweets* de um candidato com uso de léxico próprio (2)



Fonte: arquivo pessoal

Portanto, nota-se uma possível falha no dicionário construído (mesmo sendo feito com auxilio de especialistas), assumindo que os candidatos publicariam textos com maior diversidade de sentimentos do que o descoberto pela avaliação. Como foram conteúdos de redes sociais, sem alvo bem definido e, alguns, com ironias e/ou gírias, entende-se o porquê da falha na avaliação. Outro problema foi ao tratar candidatos falando de seus concorrentes: se um candidato X fala do partido do candidato Y, a tendência é que o sentimento seja ruim; entretanto, se o candidato Y fala de seu próprio partido, a tendência é que seja positivo. Com análise por léxicos, fazer esse tipo de ponderação não é possível. Sendo assim, optou-se por testar ferramentas mais inteligentes oferecidas pela própria *Orange*, a serem detalhadas a seguir.

6.4.2 Classificação com métodos automáticos do *Orange*

A ferramenta *Orange* possui uma biblioteca de *text mining* e que, portanto, oferece duas funções de análise de sentimentos de textos. A primeira (chamada de *Sentiment Analysis*) oferece uma avaliação continua, com uma nota entre -1 e 1; a segunda, por sua vez, é chamada de *Tweet Profiler* e, como seu nome diz, é especifica para *tweets*, categorizando-os conforme um conjunto de emoções, como alegria, tristeza e confiança. Ambas, entretanto, funcionam apenas para conteúdos na língua inglesa.

Sendo assim, foi necessário, primeiramente, desenvolver um sistema que traduzisse os *tweets* para o inglês. Assim como feito para o *web crawler*, o método em cascata foi aplicado:

- Comunicação: levantamento de requisitos
  - Conexão estável com a Internet;
  - Possuir chave de acesso à API da plataforma Google Cloud;
  - Traduzir os tuítes pré-processados para a língua inglesa;
  - Utilizar inglês americano como alvo para tradução;
  - Utilizar português brasileiro como idioma padrão dos *tweets*;
  - Criar arquivo em formato .CSV para posterior importação no *Orange*;
- Planejamento:
  - Duração esperada de um dia de trabalho;
  - Etapas:
    - Estabelecer conexão com API do Google Cloud Platform;
    - Testar função de tradução para um texto qualquer;
    - Testar função de tradução para um *tweet* qualquer;
    - Aplicar tradução em todos os *tweets;*
    - Criação de arquivo .CSV;
- Modelagem: pode ser reduzida a três etapas, conexão com API, tradução dos tuítes e criação de arquivo .CSV;
- Construção: criação do código em si. Para isso, foi usada a ferramenta R, a biblioteca *translate* e a API do Google Cloud;
- Implantação: no caso desse projeto, a implantação foi feita para traduzir os tuítes, conforme seu objetivo, sendo necessária a sua realização apenas uma vez.

A Figura 6.33 apresenta parte do arquivo gerado pelo programa. Observa-se, nele, as traduções feitas para o inglês. Esse documento seria, em seguida, importado para o *Orange*.

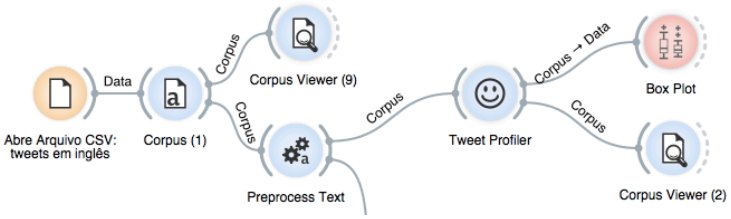
Figura 6.33 – Parte de arquivo CSV com *tweets* traduzidos para o inglês

10/7/18 01:45	Haddad_Fernand	meet plan government haddadpresidente	3248	1375
10/6/18 22:17	Haddad_Fernand	project bolsonaro will make population feel longing for government fear to cut rights prc	8008	1951
10/6/18 22:16	Haddad_Fernand	workers rights	3623	909
10/6/18 22:14	Haddad_Fernand	I can believe I can count on winning vote last minute I believe violence dictatorship lack f	3199	861
10/6/18 22:05	Haddad_Fernand	promote false news in addition costs money opponent committed little bids attack	968	420
10/6/18 21:58	Haddad_Fernand	teacher two months ago giving lesson live salary give lesson act donation teacher act de	8383	2002
10/6/18 21:57	Haddad_Fernand	god gave talent to the world it fits to offer opportunity to develop talent voteporlulavote	1319	343
10/6/18 21:55	Haddad_Fernand	world, deserves, respect, protection, government, government, puts, order, this, governn	1275	359

Fonte: arquivo pessoal

O primeiro teste foi ao aplicar a função de *Tweet Profiler*. A Figura 6.34 apresenta o diagrama de blocos para tal. Nota-se que, após a importação do arquivo e sua conversão para *corpus*, foi necessário fazer um novo pré-processamento, mas com as configurações para a língua inglesa. Entretanto, isso era o suficiente para aplicar o *corpus* na função, sem precisar criar a matriz *Bag of Words*.

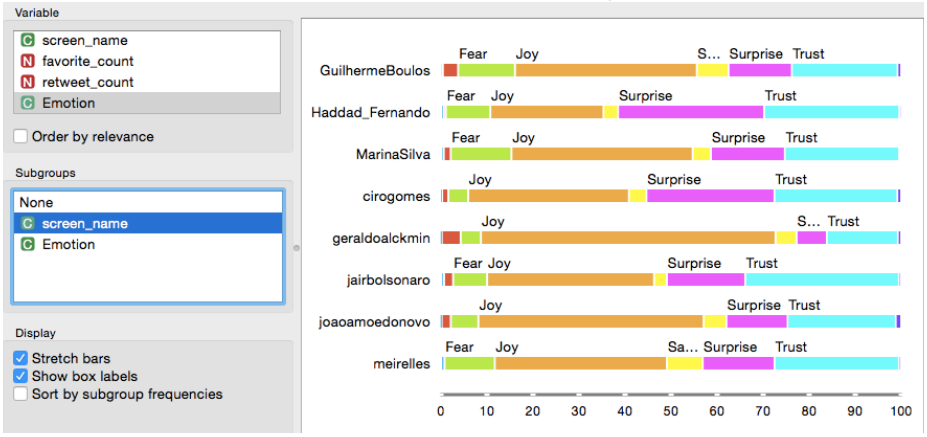
Figura 6.34 - Diagrama de blocos para uso da função *Tweet Profiler* no *Orange*



Fonte: arquivo pessoal

O resultado obtido, entretanto, é pouco conclusivo e está demonstrado na Figura 6.35. Nota-se que todos os candidatos possuem distribuição similar entre as emoções, com a maioria dos *tweets* transmitindo alegria (*joy*), confiança (*trust*) ou surpresa (*surprise*). Com isso, é difícil diferenciar estratégias de campanhas de cada um e, principalmente, se as mensagens transmitidas afetam o voto ou não.

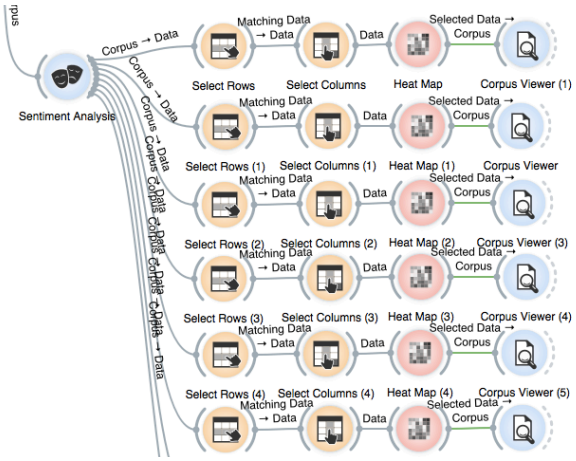
Figura 6.35 - Resultado da função *Tweet Profiler* no *Orange*



Fonte: arquivo pessoal

Em seguida, a função de *Sentiment Analaysis* foi verificada. Ela, que dá uma nota contínua entre -1 e 1, demanda as mesmas configurações de utilização que a *Tweet Profiler*, ou seja, apenas o bloco de pré-processamento. A Figura 6.36 apresenta o diagrama de blocos de seu uso (após o pré-processamento), onde cada linha do diagrama consiste na separação dos *tweets* de um candidato específico.

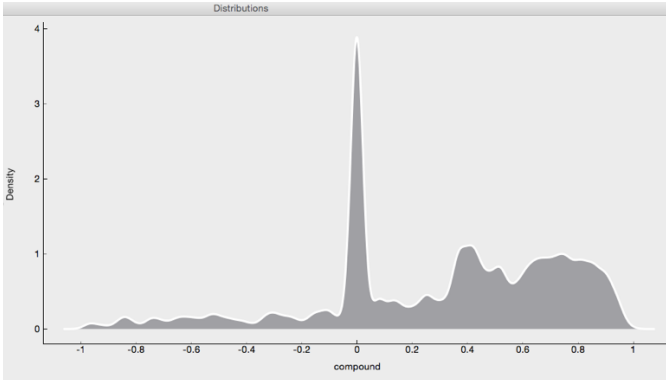
Figura 6.36 - Diagrama de blocos para uso da função *Sentiment Analysis* no *Orange*



Fonte: arquivo pessoal

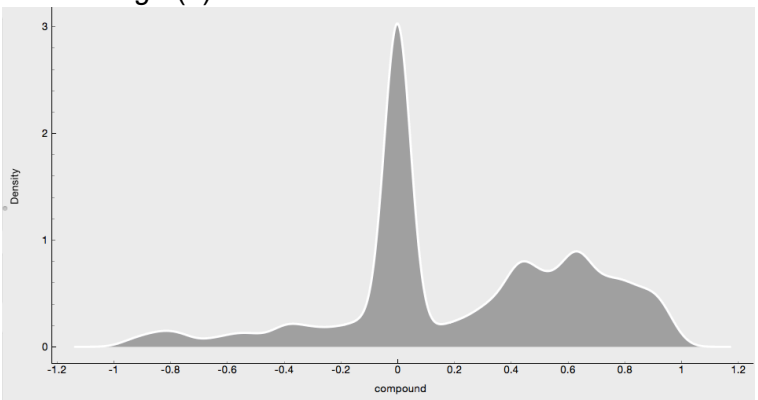
Entretanto, para todos os candidatos, a grande maioria dos *tweets* recebeu nota zero, o que indica a incapacidade de avaliar certos conteúdos por conta da função do *Orange*. As Figuras 6.37 a 6.40 apresentam a distribuição das avaliações de alguns candidatos, onde observa-se a concentração em torno do zero.

Figura 6.37 – Distribuição de avaliações de *tweets* de um candidato com uso da função *Sentiment Analysis* no *Orange* (1)



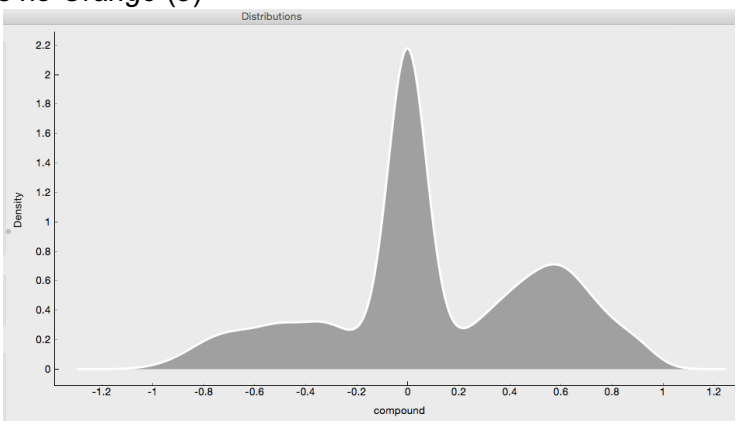
Fonte: arquivo pessoal

Figura 6.38 – Distribuição de avaliações de *tweets* de um candidato com uso da função *Sentiment Analysis* no *Orange* (2)



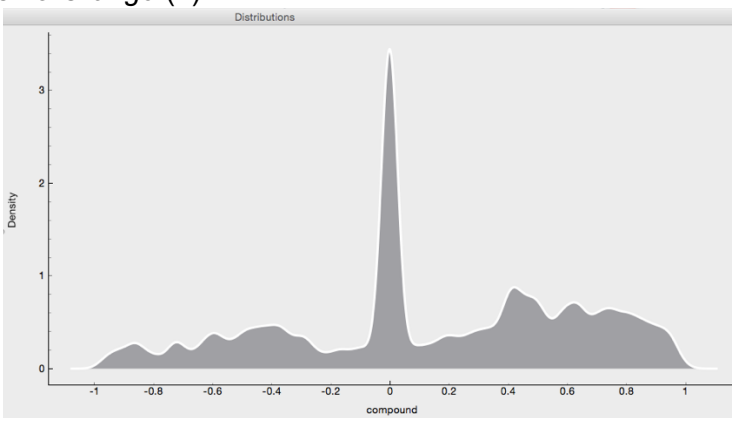
Fonte: arquivo pessoal

Figura 6.39 – Distribuição de avaliações de *tweets* de um candidato com uso da função *Sentiment Analysis* no *Orange* (3)



Fonte: arquivo pessoal

Figura 6.40 – Distribuição de avaliações de *tweets* de um candidato com uso da função *Sentiment Analysis* no *Orange* (4)



Fonte: arquivo pessoal

Sendo assim, nenhuma das funções oferecidas pela ferramenta mostraram-se adequadas para serem usadas em previsibilidade eleitoral. Além de classificaram textos de forma genérica (o que apresenta os mesmos problemas das categorizações com léxicos genéricos), a tradução automática para o inglês certamente resulta na diminuição da qualidade de seus resultados. Por isso, novas técnicas de classificação foram testadas, dessa vez com uma abordagem por aprendizado de máquina - a ser detalhada na próxima seção.

6.4.3 Classificação com aprendizado de máquina

Visando obter classificações mais confiáveis e que pudessem ser utilizadas na previsibilidade eleitoral, optou-se pela construção de um modelo com base em aprendizado de máquina. Este modelo seria classificado de forma discreta, com três sentimentos: negativo, neutro e positivo. Como apresentado na Seção 4.2, existem diversos algoritmos que criam modelos de classificação com aprendizado de máquina. Portanto, tornou-se necessário avaliar os resultados de alguns para optar por um na

construção do modelo. Essa avaliação, feita na ferramenta *Weka*, ocorreu com base na acurácia de cada modelo.

A partir dos dados pré-processados, diversos algoritmos foram aplicados. Em aprendizado de máquina, há diversas estratégias para executar o "treinamento" a partir dos dados fornecidos. O foco do treinamento é de demonstrar ao classificador exemplos de modo a permitir o aprendizado sobre os textos. Dentre as técnicas descritas na literatura, optou-se pelo uso da *Cross Validation K-Fold*.

A *Cross Validation K-Fold*, ou Validação Cruzada, usa o conceito de *folds* (pastas), dividindo o conjunto de treinamento (neste caso, 10% da amostra) em *k* conjuntos. Deste total, apenas um deles é usado para validar o modelo criado (chamado de conjunto de teste) e os restantes compõem o treinamento. O processo é repetido *k* vezes, cada conjunto sendo utilizado pelo menos uma vez como teste. No final, obtém-se a média de desempenho do classificador em todas as iterações.

Notadamente, quanto maior o *k*, mais custoso será para a máquina, afinal, tornam-se necessárias mais iterações. Entretanto, como a amostra deste projeto é relativamente pequena (no *Weka*, seriam 605 *tweets*), pode-se usar um valor alto para *k* sem prejudicar o desempenho da máquina. Sendo assim, optou-se por deixa-lo como 20, o valor máximo permitido no *Orange*.

O primeiro algoritmo testado foi o K-vizinhos mais próximos. O *Weka* permite que o projetista determine o valor de *k* de forma dinâmica (fornecer um intervalo e o algoritmo encontra o valor ótimo) ou estática (forneça um valor fixo para ele), além de especificar o método de calculo de distância entre vizinhos. O método escolhido foi o Euclidiano, por conta das referências na literatura. Quanto ao *k*, foram testados, estatisticamente e dinamicamente, todos os valores entre 1 e 10. A melhor performance (tanto acurácia quanto *F-measure*) foi obtida com k=1.

A Tabela 3 abaixo apresenta a matriz de confusão obtida, ou seja, relação entre as classificações feitas pelo algoritmo com as reais (feitas pelo especialista). As colunas indicam o que foi obtido pelo algoritmo (as previsões), enquanto as linhas indicam a classificação real. Portanto, apenas os elementos da diagonal principal são os corretamente categorizados.

Tabela 3 – Matriz de confusão do modelo criado pelo algoritmo de k-vizinhos mais próximos

		Classe prevista		
		Negativa	**Neutra**	**Positiva**
Classe real	**Negativa**	76	22	55
	Neutra	10	85	57
	Positiva	16	57	217

Fonte: arquivo pessoal

Com estes números, obteve-se uma acurácia de 64,13% e a média ponderada das *F-measures* das três categorias de 0,638.

Em seguida, foi avaliado o algoritmo C4.5 (árvore de decisão). Optou-se por partir de uma única raiz (no caso, o termo com maior ganho de informação, segundo

o ranqueador), o que resultou em uma árvore com 87 nós e 44 folhas. A matriz de confusão deste algoritmo é apresentada na Tabela 4.

Tabela 4 – Matriz de confusão do modelo criado pelo algoritmo C4.5

		Classe prevista		
		Negativa	**Neutra**	**Positiva**
Classe real	**Negativa**	68	9	76
	Neutra	14	63	85
	Positiva	29	27	234

Fonte: arquivo pessoal

Com estes números, obteve-se uma acurácia de 60,33% e a média ponderada das *F-measures* das três categorias de 0,587.

O terceiro algoritmo avaliado foi a floresta randômica. Para o *Weka*, sua construção é feita por meio de um conjunto de árvores com C4.5 e a quantidade delas é definida pelo usuário - neste caso, optou-se por 20. Cada árvore, naturalmente, possui sua quantidade específica de nós e folhas. O resultado obtido da classificação é exibido na Tabela 5.

Tabela 5 – Matriz de confusão do modelo criado pelo algoritmo de floresta aleatória

		Classe prevista		
		Negativa	**Neutra**	**Positiva**
Classe real	**Negativa**	85	14	54
	Neutra	13	92	57
	Positiva	23	36	231

Fonte: arquivo pessoal

Com estes números, obteve-se uma acurácia de 67,44% e a média ponderada das *F-measures* das três categorias de 0,669.

O quarto algoritmo utilizado foi o de Redes Neurais. No *Weka*, o tipo de rede neural disponível é o de *perceptron* multicamadas (ou MLP), ou seja, possui mais de uma camada de neurônios em alimentação direta. O *Weka* permite configurar os fatores de correção para a realimentação do algoritmo, após o cálculo do erro. O modelo elaborado pela ferramenta possuía 96 nos, cada um com pesos diferentes dados aos atributos. Com isso, obteve-se a matriz de confusão demonstrada na Tabela 6.

Tabela 6 – Matriz de confusão do modelo criado pelo algoritmo de redes neurais

		Classe prevista		
		Negativa	**Neutra**	**Positiva**
Classe real	**Negativa**	94	15	44
	Neutra	28	84	50
	Positiva	51	28	211

Fonte: arquivo pessoal

Com estes números, obteve-se uma acurácia de 64,30% e a média ponderada das *F-measures* das três categorias de 0,641.

Por fim, o quinto e ultimo algoritmo avaliado foi o de Naïve-Bayes. O *Weka* permite poucas configurações dele, sendo a principal delas usar (ou não) do estimador de Kernel (ou EDK). Entretanto, como o *Orange* não oferece essa possibilidade, optou-se pela simulação sem seu uso (apesar de a acurácia, com ele, melhorar em aproximadamente 1 p.p.). A Tabela 7 apresenta a matriz de confusão resultante.

Tabela 7 – Matriz de confusão do modelo criado pelo algoritmo de Naïve-Bayes

		Classe prevista		
		Negativa	**Neutra**	**Positiva**
Classe real	**Negativa**	117	4	32
	Neutra	12	97	53
	Positiva	22	23	245

Fonte: arquivo pessoal

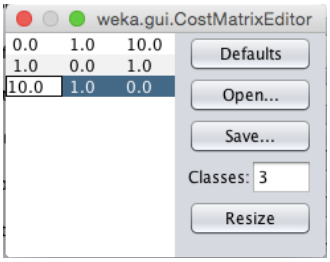
Com estes números, obteve-se uma acurácia de 75,87% e a média ponderada das *F-measures* das três categorias de 0,755.

O *Weka* permite também acrescentar uma consideração de custo no classificador, chamada sensibilidade de custos. Isso significa que alguns erros "custam" mais do que outros e, assim, o algoritmo procura evita-los mais do que os demais. Seu uso também foi testado, adicionando-o ao algoritmo de Naïve-Bayes (que teve os melhores resultados dentre os algoritmos considerados). A Figura 6.41 e a Figura 6.42 apresentam a configuração desses custos. A matriz de custo foi criada de tal forma que *tweets* com sentimentos negativos, mas avaliados como positivos, e seu contrário (positivos mas previstos como negativos) tivessem custo 10 vezes maior do que os demais erros.

Figura 6.41 – Configuração da sensibilidade de custos no *Weka* (1)

Fonte: arquivo pessoal

Figura 6.42 – Configuração da sensibilidade de custos no *Weka* (2)



Fonte: arquivo pessoal

Com isso, o resultado esperado era de que esses erros (que antes eram 32 e 22) reduzissem. Em compensação, é possível que o algoritmo compense ao aumentar os outros erros (principalmente, com *tweets* de sentimento neutro). A Tabela 8 apresenta o obtido:

Tabela 8 – Matriz de confusão do modelo criado com sensibilidade de custos

		Classe prevista		
		Negativa	**Neutra**	**Positiva**
Classe real	**Negativa**	80	62	11
	Neutra	6	136	20
	Positiva	5	111	174

Fonte: arquivo pessoal

Com estes números, obteve-se uma acurácia de 64,46% e a média ponderada das *F-measures* das três categorias de 0,657. Nota-se que os erros mais custosos foram reduzidos substancialmente (de um total de 52 para 16). Entretanto, as demais métricas de qualidade do algoritmo mostram que ele “sobre-compensou” em outras classes.

Tendo todos os testes realizados, não apenas com os algoritmos de aprendizado de máquina, mas também com as outras abordagens, foi necessário optar por um dos métodos para ser o utilizado na construção final do modelo - e, por consequência, para avaliar a teoria das estratégias emocionais.

6.4.3 Escolha de um método

A escolha foi por aplicar uma técnica de aprendizado de máquina, mais especificamente o algoritmo de Naïve-Bayes, sem ponderação de custos.

Os métodos com uso de dicionários mostraram-se falhos. Aqueles com uso de léxicos genéricos falham ao não terem vocabulário específico para o contexto (campanhas presidenciais de 2018) e por nem conseguirem categorizar metade dos *tweets*. Já o léxico personalizado não consegue levar em consideração parte do contexto (como o candidato) e mostrou-se imparcial, ao categorizar a grande maioria dos conteúdos como positivos.

As funções oferecidas pelo *Orange*, por sua vez, também não apresentam a confiabilidade necessária. Enquanto a *Tweet Profiler* não permitiu diferenciar estratégias entre candidatos, a *Sentiment Analysis* também não conseguiu avaliar boa

parte dos *tweets*. Além disso, a necessidade da tradução apresenta um risco à qualidade da classificação, já que fazê-la de forma automática pode causar mudanças consideráveis ao texto.

Por fim, já sabendo que um algoritmo com aprendizado de máquina seria a melhor solução, bastou apenas selecionar aquele com melhor desempenho. A Tabela 9 abaixo apresenta as métricas de qualidade dos 6 algoritmos avaliados.

Tabela 9 – Comparação de desempenho entre algoritmos

Algoritmo	**Acurácia**	***F-measure***
K-vizinhos mais próximos	64,13%	0,638
C4.5	60,33%	0,587
Floresta Randômica	67,44%	0,669
Redes Neurais	64,30%	0,641
Naïve-Bayes	75,87%	0,755
Naïve-Bayes com custo ponderado	64,46%	0,657

Fonte: arquivo pessoal

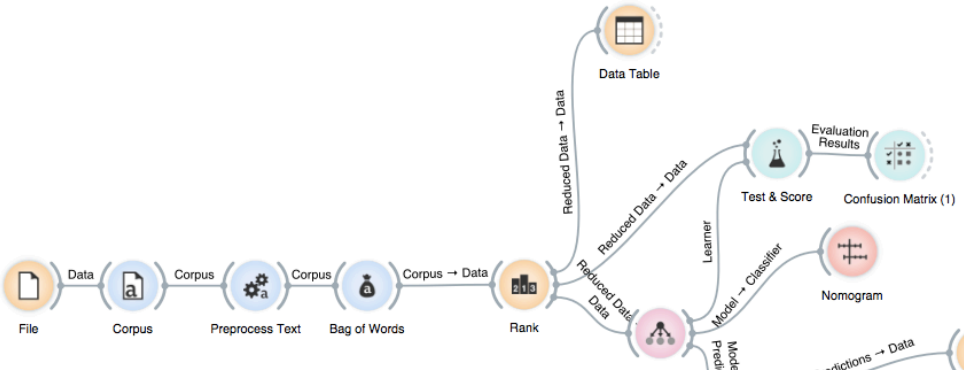
A partir dela, nota-se que o algoritmo de Naïve-Bayes, sem matriz de custo, apresenta o melhor desempenho - e, por isso, foi o escolhido para a construção do modelo, a ser detalhado na próxima seção.

6.5 Modelo construído

Com a confirmação de que o algoritmo de Naïve-Bayes era o escolhido para a construção do modelo de classificação, tornou-se necessária a construção em si. Como a ferramenta é mais visual, *user friendly* e permite a aplicação de modelos em outras amostras de dados, optou-se pelo uso do *Orange* para tal. Para isso, o pré-processamento foi feito conforme indicado na Seção 6.3. Lembrando que, no *Orange*, as publicações de todos os 8 candidatos selecionados foram analisadas.

Em seguida, foi necessário criar o diagrama de blocos exibido na Figura 6.43. Primeiramente, foi utilizada uma função fornecida pela ferramenta e específica para aplicação do algoritmo de Naïve-Bayes. Essa função não tem parâmetro algum a ser configurado pelo usuário e seu objetivo é construir um modelo de categorização a partir dos dados fornecidos. Tanto os dados pré-processados quanto o modelo criado por essa função foram conectados à função *Test & Score*, que aplica a validação cruzada (novamente, com 20 *folds*). Em seguida, o bloco *Confusion Matrix* constroi a matriz de confusão, enquanto o *Nomogram* permite visualizar o modelo construído.

Figura 6.43 – Diagrama de blocos feito no *Orange* para construção do modelo final



Fonte: arquivo pessoal

Com este diagrama, obteve-se, como resultado, a matriz de confusão apresentada na Tabela 10.

Tabela 10 – Matriz de confusão do modelo final

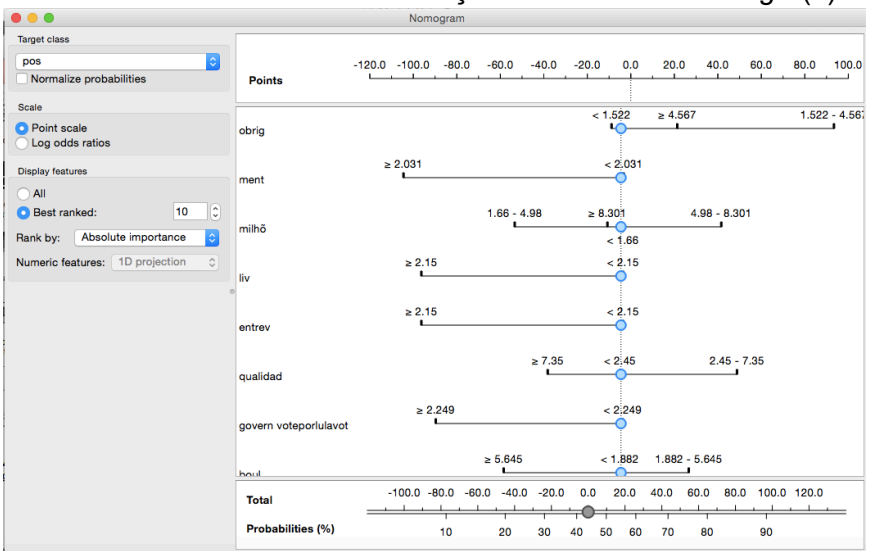
		Classe prevista		
		Negativa	**Neutra**	**Positiva**
Classe real	**Negativa**	195	15	30
	Neutra	21	150	38
	Positiva	47	50	234

Fonte: arquivo pessoal

Com estes números, obteve-se uma acurácia de 75,1% e a média ponderada das *F-measures* das três categorias de 0,749. Com isso, considerou-se que este modelo estava próximo o suficiente daquele construído pelo *Weka*.

A Figura 6.44 apresenta a maneira do *Orange* apresentar o modelo construído. No canto superior esquerdo, é indicado o sentimento avaliado (no caso, positivo). Cada linha representa um termo, enquanto a escala ao seu lado e o ponto azul indicam a pontuação do termo daquela linha no documento. Com base nas pontuações dos termos, na parte de baixo da janela, é indicada a probabilidade de um documento (com essas pontuações) ter sentimento positivo.

Figura 6.44 – Parte do modelo final de classificação construído no *Orange* (1)

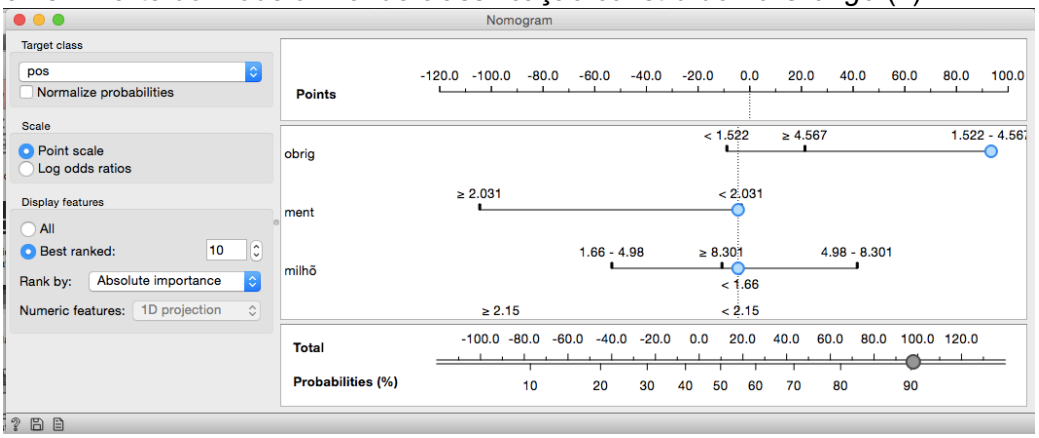


Fonte: arquivo pessoal

Neste caso, observa-se que com as seguintes pontuações dos termos em um documento ("*obrig*"<1.522; "*ment*"<2.031…), a probabilidade é de aproximadamente 45% que ele seja positivo.

A Figura 6.45 demonstra o caso em que a pontuação de "*obrig*" passa para o intervalo entre 1.522 e 4.567. Nota-se que, apenas com essa alteração, a probabilidade de que o documento seja positivo passa a 90%.

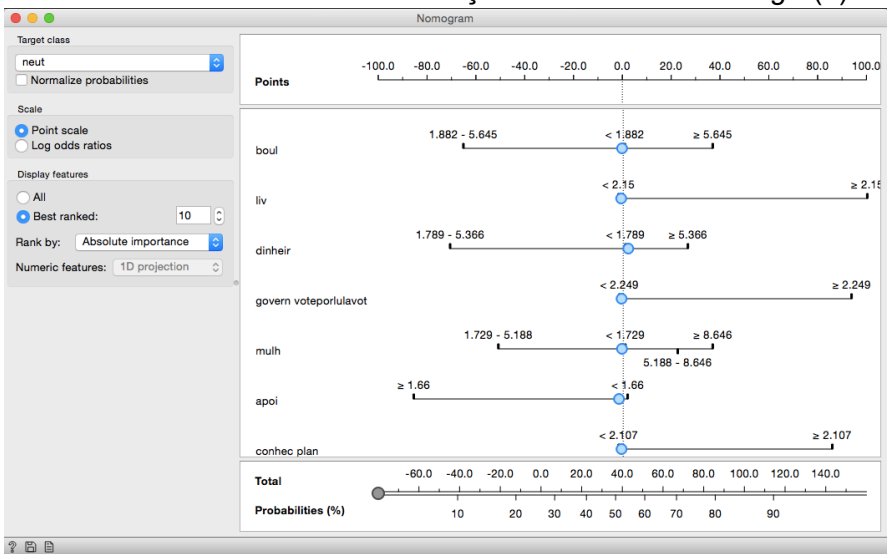
Figura 6.45 – Parte do modelo final de classificação construído no *Orange* (2)



Fonte: arquivo pessoal

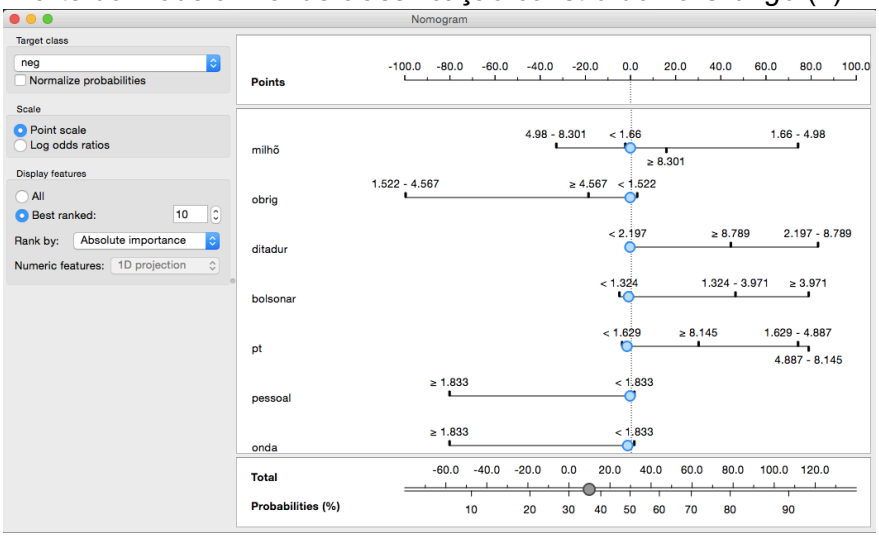
A mesma análise pode ser feita para os demais sentimentos (neutro, negativo), conforme apresentado nas Figuras 6.46 e 6.47.

Figura 6.46 – Parte do modelo final de classificação construído no *Orange* (3)



Fonte: arquivo pessoal

Figura 6.47 – Parte do modelo final de classificação construído no *Orange* (4)



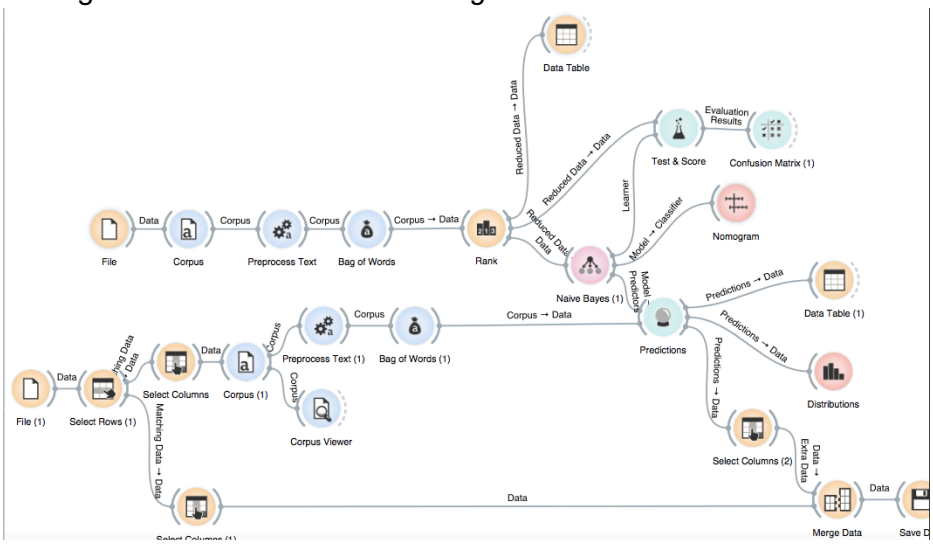
Fonte: arquivo pessoal

Nota-se que, para cada classe, há termos mais e menos relevantes, conforme já esperado e pré-analisado no *Weka*.

6.6 Visualização

A última etapa em um projeto de mineração de opiniões é visualizar as classificações obtidas de forma que as devidas análises possam ser feitas. Para isso, primeiramente, deve ser feita a aplicação do modelo construído aos demais *tweets* para identificar seus sentimentos. A Figura 6.48 apresenta o diagrama de blocos, no *Orange*, para que isso fosse feito.

Figura 6.48 - Diagrama de blocos final do *Orange*

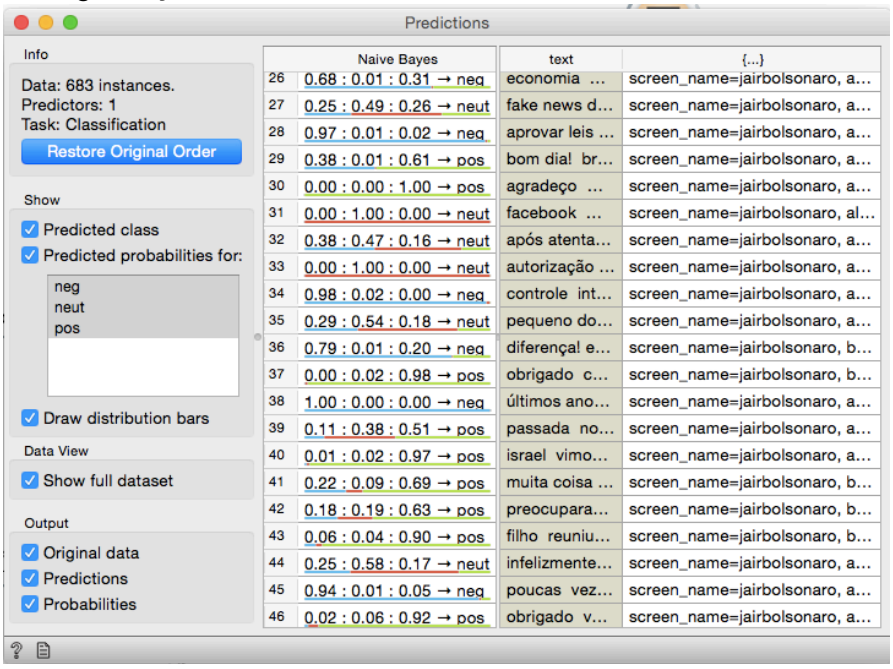


Fonte: arquivo pessoal

Nota-se, na imagem, que a parte superior do diagrama é a construção do modelo em si e que a parte de baixo é a categorização dos novos dados. Percebe-se, também, que o pré-processamento e a criação da matriz *BoW* são novamente utilizados.

A Figura 6.49, por sua vez, apresenta como o *Orange* categoriza os documentos. Para cada um, são analisados seus atributos (*features*) e, com base no modelo criado, são calculadas as probabilidades de ele pertencer a cada uma das classes possíveis. Aquela com maior porcentagem é apontada como a categoria do *tweet*.

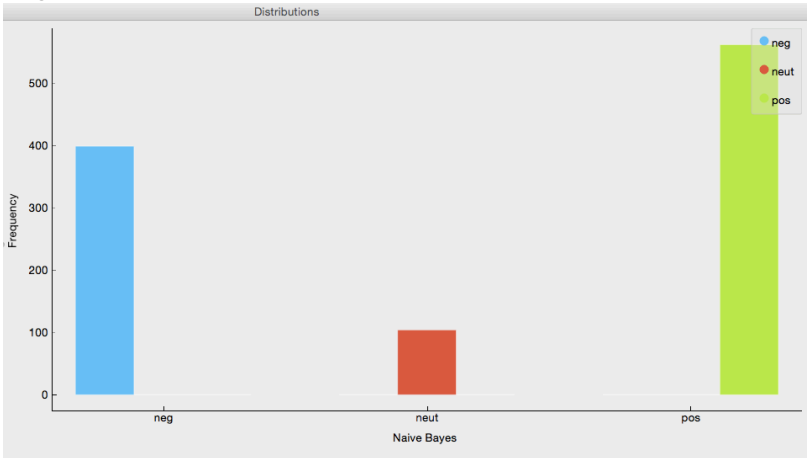
Figura 6.49 - Categorização dos *tweets* restantes



Fonte: arquivo pessoal

Em seguida, a primeira visualização feita, para cada candidato, foi a quantidade de cada sentimento. A Figura 6.50 apresenta os resultados obtidos para o Candidato 1.A cor azul é utilizada para os sentimentos negativos; a vermelha, para os neutros; e a verde, para os positivos. O eixo vertical, por sua vez, indica a quantidade de *tweets* em cada uma.

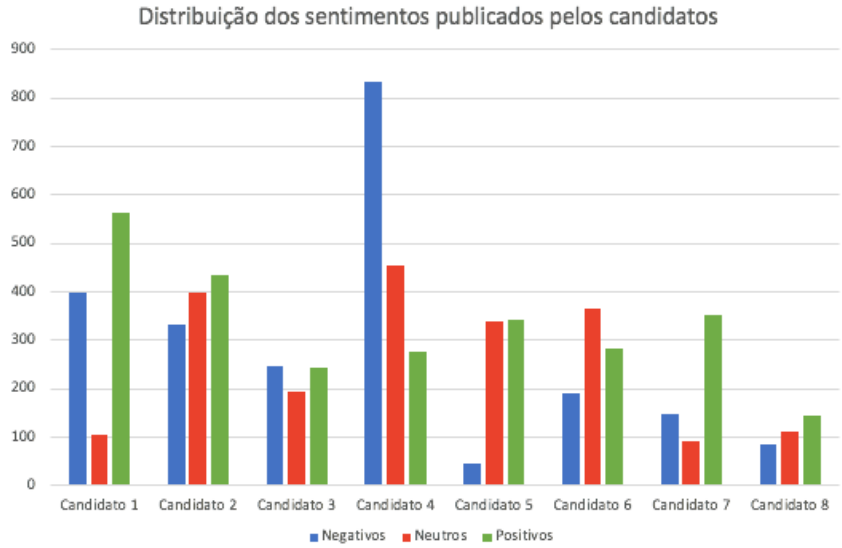
Figura 6.50 - Categorização dos *tweets* restantes do Candidato 1



Fonte: arquivo pessoal

A Figura 6.51 apresenta essa avaliação, para todos os candidatos, em um mesmo gráfico.

Figura 6.51 - Categorização dos *tweets* restantes de todos os candidatos



Fonte: arquivo pessoal

A partir desses gráficos, entretanto, não é possível fazer análises minuciosas a respeito das campanhas de cada um e como os conteúdos afetaram os votos. Por isso, optou-se também por fazer uma visualização temporal das publicações, acompanhando a quantidade de publicações feitas e a variação de seus sentimentos no decorrer da campanha.

Como o período analisado é muito grande, o que significa uma grande quantidade de publicações, optou-se pela divisão da distribuição dos sentimentos em dois períodos: pré-campanha (antes de 16 de agosto) e durante a campanha (entre 16 de agosto e 06 de outubro). O Candidato 8, por sua vez, teve poucos *tweets* publicados que cumprissem com os pré-requisitos estabelecidos anteriormente e, portanto, a análise de seu conteúdo não será feita.

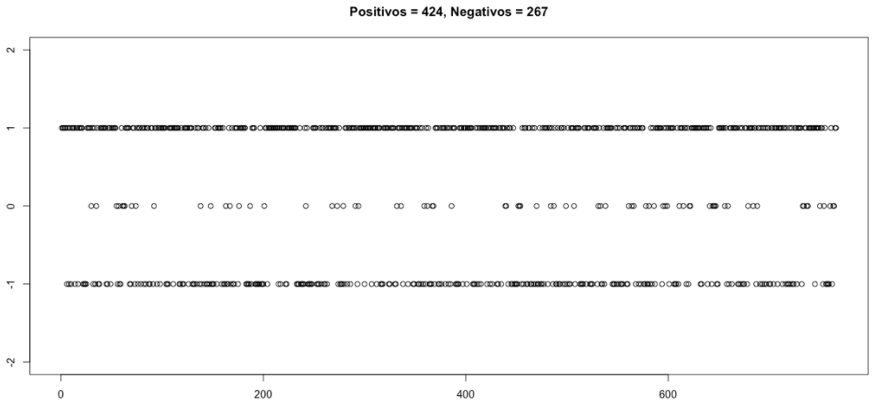
7 ANÁLISE DOS RESULTADOS OBTIDOS

Neste capítulo, serão discutidos os resultados obtidos com a aplicação do modelo de categorização criado e apresentado no capítulo anterior. Para isso, cada candidato será discutido individualmente em uma seção (ou seja, a Seção 7.1 será dedicada à análise do Candidato 1, e assim sucessivamente). Ao final, na Seção 7.8 será feita uma recapitulação geral dos resultados.

7.1 Análise do Candidato 1

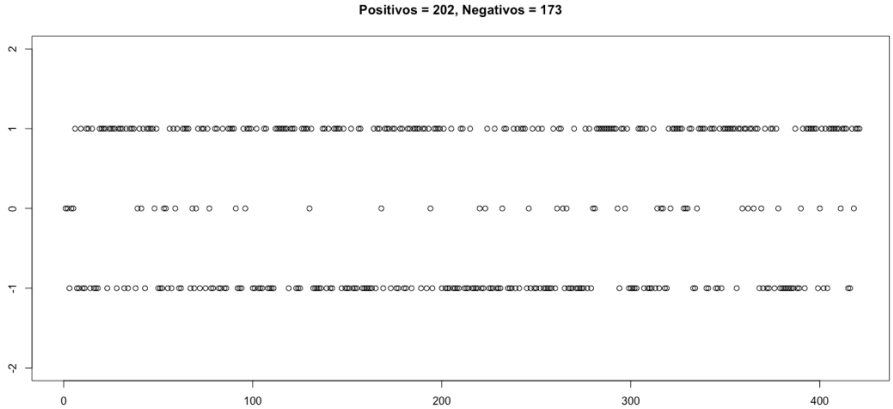
Nota-se, na Figura 6.51, que o candidato publicou mais conteúdos com sentimento positivo do que negativo. Entretanto, ao comparar a distribuição de sentimentos da antes da campanha (Figura 7.1, que contém a distribuição dos sentimentos dos *tweets* desse período. Neste caso, o eixo vertical indica os seguintes valores: +1 representa sentimentos positivos; 0, neutros; e -1, negativos. O eixo horizontal indica o *ID* do *tweet*, em ordem cronológica) com o período eleitoral (Figura 7.2) percebe-se que o candidato mudou o "tom" de seu discurso com o início da campanha. O que antes apresentava-se mais positivo, passou a conter mais publicações com sentimento negativo. Isso significa que o candidato, possivelmente, optou por uma estratégia emocional de ataque a seus concorrentes e a disseminar mais medo.

Figura 7.1 - Categorização temporal de *tweets* do Candidato 1, antes da campanha



Fonte: arquivo pessoal

Figura 7.2 - Categorização temporal dos *tweets* do Candidato 1, durante a campanha



Fonte: arquivo pessoal

Outra mudança com o início da campanha foi na frequência de publicações, que aumentou com o começo do período eleitoral. Entretanto, o curioso foi que o crescimento se tornou exponencial ao longo da campanha, indicando, possivelmente, uma mudança de foco na estratégia de comunicação do candidato, passando a focar mais em suas redes sociais.

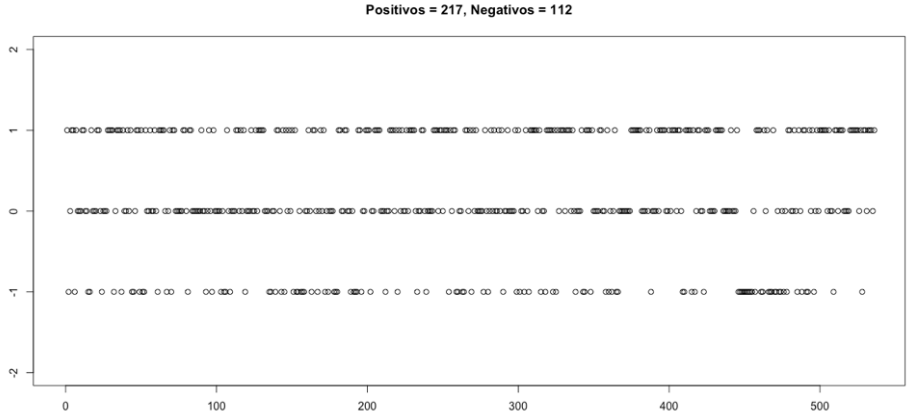
Observa-se também os efeitos de eventos externos em seus conteúdos no período da campanha (Figura 7.2). Após o atentado sofrido por um outro candidato no dia 06 de setembro (ou seja, ID = 120 no eixo horizontal da imagem), as publicações do Candidato 1 tornaram-se majoritariamente positivas por um período. Isso durou até 6 de setembro (ou ID = 200), quando foi divulgado o resultado de uma nova pesquisa de opinião feita pelo Ibope. Nela, o Candidato 1 apresentou queda, enquanto um de seus oponentes (e com quem disputava uma vaga no segundo turno) disparou à frente na segunda colocação. Esse resultado gerou uma série de publicações com ataques a este candidato - e, portanto, com sentimento negativo.

Até o final da campanha, houveram mais duas sequências consideráveis com *tweets* consecutivamente positivos. Uma, por volta de 24 de setembro, quanto o Ibope lançou nova pesquisa apresentando crescimento do Candidato 1 (o que certamente gerou entusiasmo de sua equipe). A segunda foi percebida em todos os candidatos: nos dois dias que antecederam as eleições (dias 5 e 6 de outubro), a tendência foi que os candidatos publicassem conteúdos positivos, motivando os eleitores a votarem neles, com palavras de esperança de bons resultados nas urnas.

7.2 Análise do Candidato 2

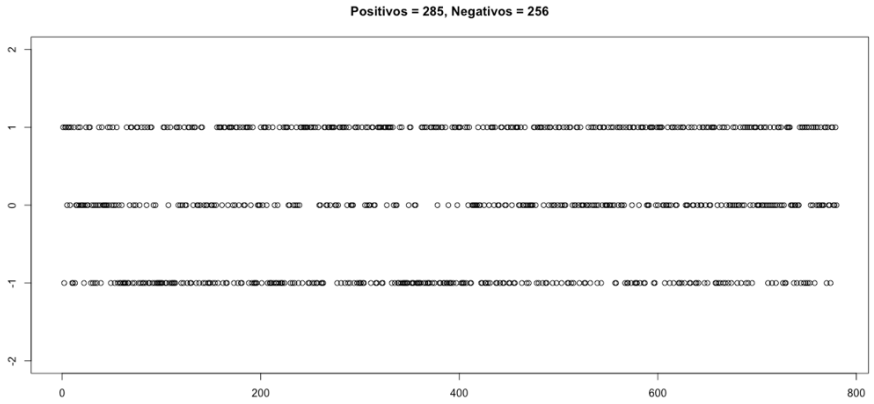
Para o Candidato 2, não há muitas particularidades e nem análises possíveis de serem feitas. Ao longo da campanha, a distribuição entre *tweets* com sentimento positivo e negativo foi praticamente equivalente (ver Figura 7.4), sem picos para um ou outro. Esse perfil é uma leve alteração em relação ao período pré-campanha (ver Figura 7.3), o qual tendia a ser mais positivo.

Figura 7.3 - Categorização temporal dos *tweets* do Candidato 2, antes da campanha



Fonte: arquivo pessoal

Figura 7.4 - Categorização temporal dos *tweets* do Candidato 2, durante a campanha



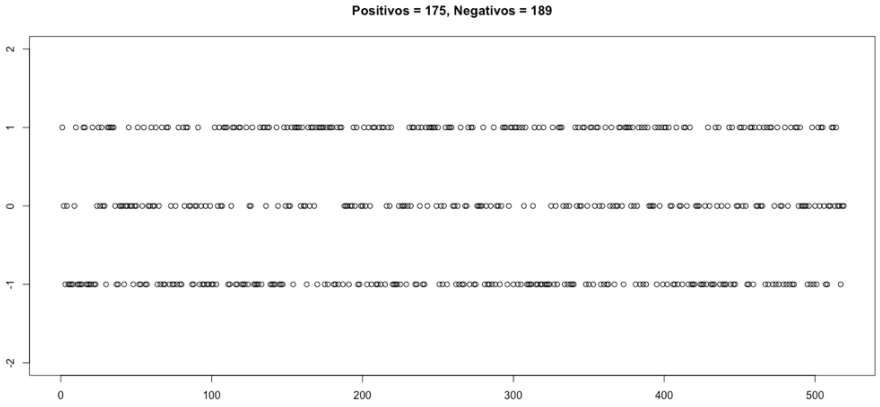
Fonte: arquivo pessoal

A frequência de publicações, por sua vez, também se manteve relativamente constante com o início do período eleitoral, apenas aumentando na última semana, com a proximidade do pleito. No caso dele, não se notou efeitos de eventos externos em suas publicações.

7.3 Análise do Candidato 3

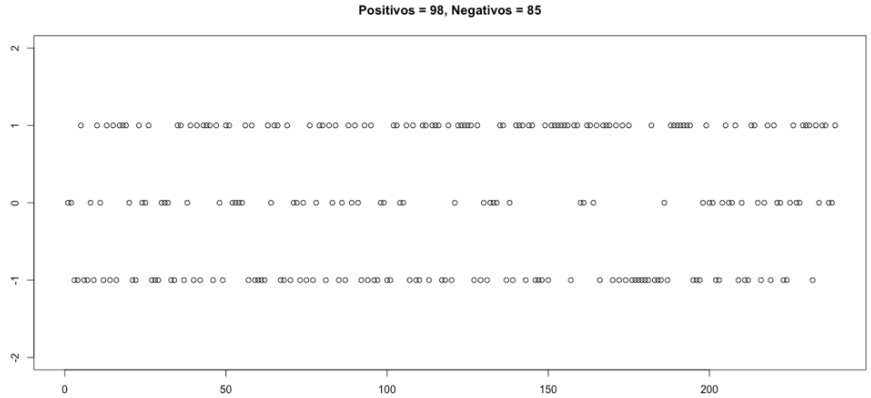
Para o 3° Candidato, nota-se, novamente, uma distribuição similar entre os sentimentos (ver Figura 6.51). Um diferencial, entretanto, foi o contraste entre o perfil de sentimentos antes e durante a campanha: enquanto antes era mais frequente a publicação de conteúdos com sentimentos negativos (com críticas a outros candidatos e dados negativos do país como um todo, ver Figura 7.5), no decorrer do período eleitoral, notou-se uma maior presença de sentimentos positivos (agradecendo o apoio de constituintes e focando em falar de mudanças a serem feitas, ver Figura 7.6).

Figura 7.7 - Categorização temporal dos *tweets* do Candidato 3, antes da campanha



Fonte: arquivo pessoal

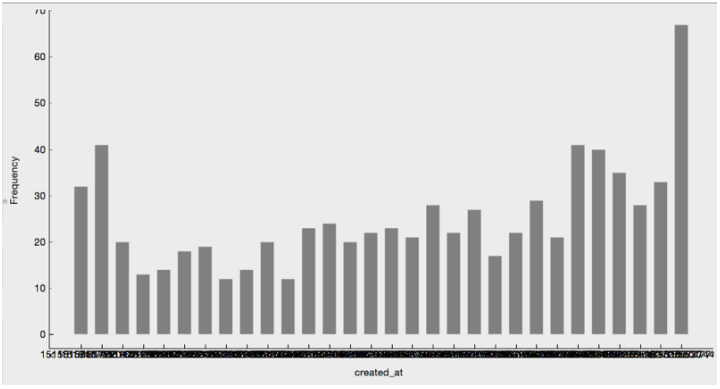
Figura 7.6 - Categorização temporal dos *tweets* do Candidato 3, durante a campanha



Fonte: arquivo pessoal

Em compensação, entre todos os candidatos, foi o que apresentou maior continuidade na frequência de publicações (ver Figura 7.7, que apresenta a distribuição temporal da frequência de publicações), apesar do início da campanha (com exceção da última semana, assim como os demais). Isso pode ser um indício de que o candidato optou por outros meios de comunicação com os constituintes, como o *Facebook* ou o *WhatsApp*, ou até que preferiu não intensificar as comunicações durante a campanha, procurando evitar risco de críticas ao que publicava.

Figura 7.7 - Distribuição temporal dos *tweets* do Candidato 3



Fonte: arquivo pessoal

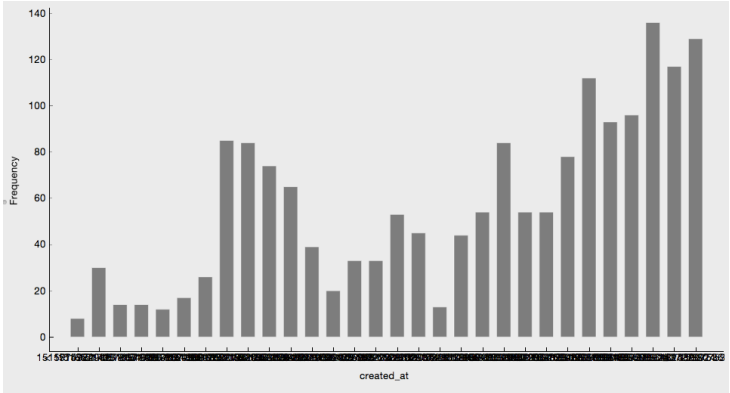
Na Figura 7.6, notam-se duas sequências relevantes que comprovam os efeitos de eventos externos na sua estratégia de comunicação. Após o dia 6 de setembro (aproximadamente, ID = 100 no eixo horizontal), quando o candidato sofreu um atentado enquanto fazia campanha, seguiu-se uma série de publicações com sentimento positivo, em sua maioria agradecendo o apoio dos eleitores e o trabalho das equipes médicas que o atenderam. Entretanto, a partir de 29 de setembro (ID = 170), quando foi divulgada uma pesquisa de opinião feita pelo Datafolha, que apontava estagnação do Candidato 3, enquanto outro candidato se aproximava rapidamente de sua pontuação, além de a *Revista Veja* ter divulgado um escândalo envolvendo sua ex-esposa (G1, 2018), seguiu-se uma sequência considerável de publicações com sentimento negativo. Nesse caso, elas consistiram tanto em acusações contra a mídia e negações das informações publicadas, quanto em críticas feitas ao candidato que o alcançava nas pesquisas.

Após esses eventos, já com o pleito se aproximando, repetiu-se o comportamento dos outros candidatos: aumentou a quantidade de publicações, em sua maioria transmitindo sentimentos positivos.

7.4 Análise do Candidato 4

O Candidato 4 apresenta algumas particularidades em relação aos demais. Primeiramente, quanto à frequência de publicações (indicado nas Figuras 6.51 e 7.8), é de longe o que mais compartilhou *tweets* entre todos os candidatos, tanto no período anterior ao lançamento da campanha quanto em seu decorrer. Ao total, ele publicou pelo menos 25% de *tweets* a mais que os demais postulantes. Além disso, apresentou uma forte constância dos sentimentos por ele transmitidos: em sua esmagadora maioria, sentimentos negativos, como visto na Figura 6.51. Isso mostra que o candidato tinha uma pauta a ser transmitida ao longo das eleições, independente dos resultados de pesquisas e dos eventos externos que pudessem ocorrer, similar ao Candidato 2, mas com foco de sentimentos diferentes.

Figura 7.8 - Distribuição temporal dos *tweets* do Candidato 4

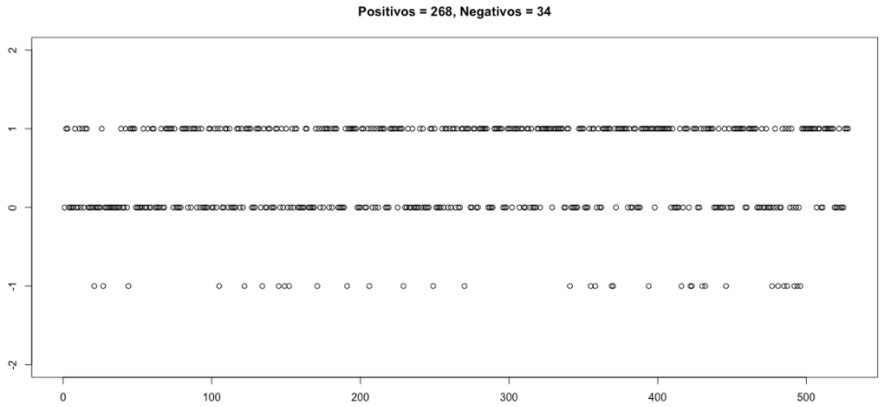


Fonte: arquivo pessoal

7.5 Análise do Candidato 5

Dentre todos os candidatos, o 5° foi o que mais publicou conteúdos com sentimentos positivos, em proporção com os com sentimentos negativos, conforme visto na Figura 6.51. Nesse caso, foram quase sete vezes mais *tweets* positivos do que negativos. Com isso, eventos externos tiveram pouca influência sob as emoções transmitidas. A exceção ocorreu mais ao final da campanha, logo após o último debate entre os candidatos, organizado pela *Rede Globo* no dia 04 de outubro (ID = 480). Nesse período, por poucos dias, o foco do candidato foi criticar as respostas dos demais e, com isso, a tendência foi transmitir sentimentos negativos (ver Figura 7.9).

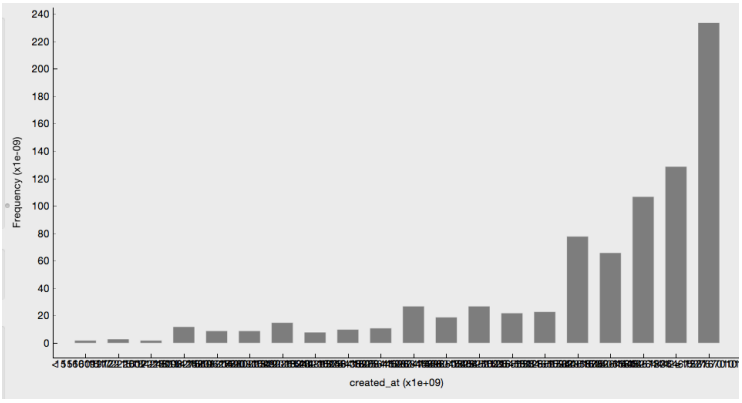
Figura 7.9 - Categorização temporal dos *tweets* do Candidato 5, durante a campanha



Fonte: arquivo pessoal

Entretanto, esse perfil pouco durou. A partir do dia 05 desse mês, com a proximidade do pleito, o candidato iniciou uma campanha com a frase "*vira vira*", tentando causar mudanças nos votos de eleitores. Com isso, além da quantidade de *tweets* ter aumentado exponencialmente, conforme pode ser visto na Figura 7.10, o sentimento dessas publicações voltou a ser positivo. Por fim, ainda a respeito da frequência de publicações, nota-se que o candidato era pouco ativo em sua rede social antes da campanhao que mudou drasticamente a partir de 16 de agosto.

Figura 7.10 - Distribuição temporal dos *tweets* do Candidato 5

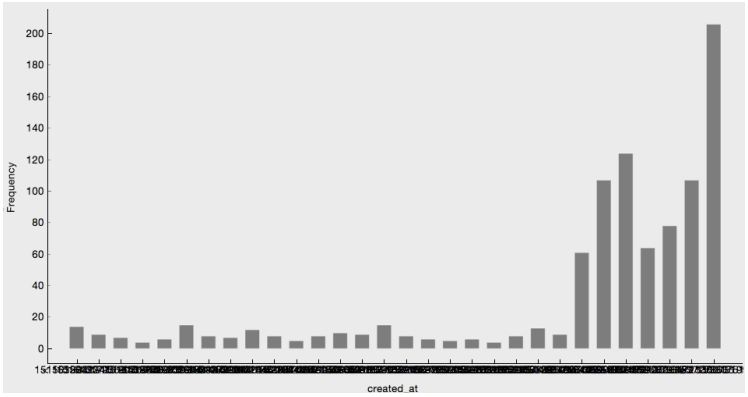


Fonte: arquivo pessoal

7.6 Análise do Candidato 6

Primeiramente, nota-se uma mudança no perfil de publicações do candidato entre o período anterior ao começo da campanha com o seu decorrer. Enquanto antes de 16 de agosto, o candidato pouco compartilhava conteúdos na rede social, a partir do início da campanha, a frequência aumentou consideravelmente (ver Figura 7.11), e constantemente também.

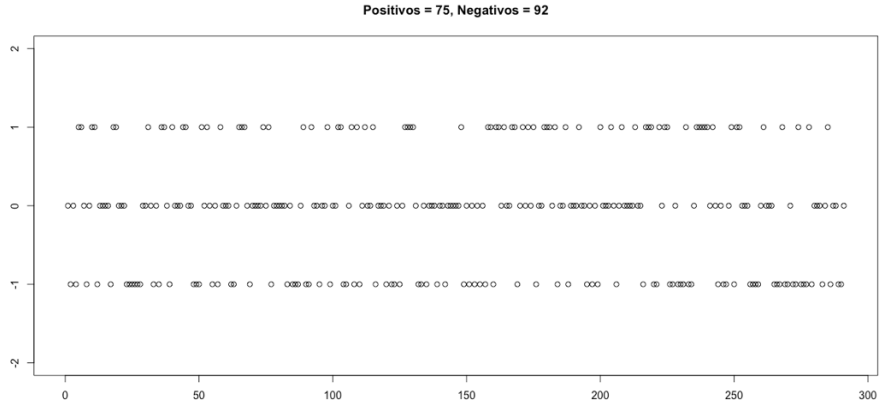
Figura 7.11 - Distribuição temporal dos *tweets* do Candidato 6



Fonte: arquivo pessoal

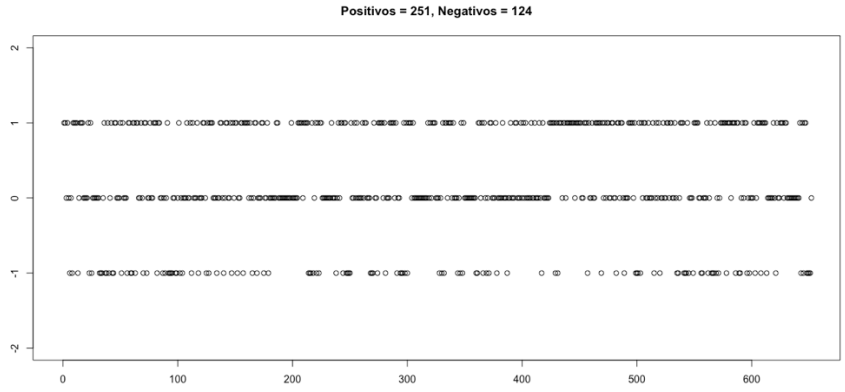
Além disso, a tendência, antes do período eleitoral, era que o sentimento transmitido pelo candidato fosse mais negativo, o que mudou com o início da campanha, conforme visto nas Figuras 7.12 e 7.13.

Figura 7.12 - Categorização temporal dos *tweets* do Candidato 6, antes da campanha



Fonte: arquivo pessoal

Figura 7.13 - Categorização temporal dos *tweets* do Candidato 6, durante a campanha



Fonte: arquivo pessoal

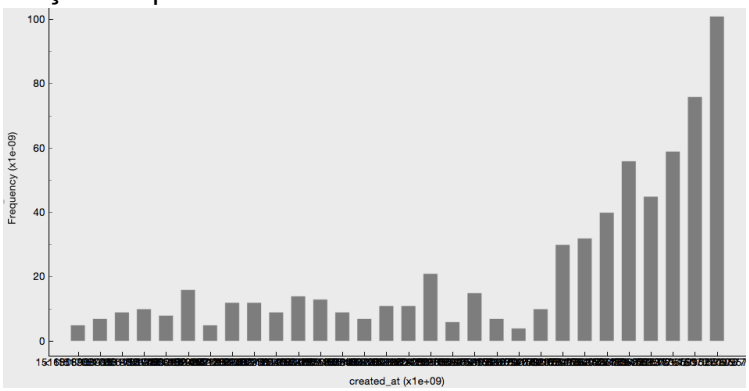
Entretanto, este Candidato teve uma particularidade em sua campanha: ele, *a priori*, seria o postulante à Vice-Presidência. Entretanto, com a proibição do candidato "original", no dia 11 de setembro, foi oficializada sua candidatura à Presidência da República (G1, 2018), o que teve seus efeitos nas publicações do candidato. A partir

desse dia (ID = 200 no eixo horizontal da Figura 7.13), houve uma sequência de *tweets* com sentimentos positivos (além do aumento natural da frequência de postagem). Além disso, no dia 19 de setembro (ID = 400), é divulgada uma pesquisa do Datafolha (G1, 2018) em que o candidato dispara à frente de seus concorrentes para assumir a segunda posição nas pesquisas. Novamente, isso gerou uma sequência de sentimentos positivos. Entretanto, a partir do dia 04 de outubro, com seu *status* nas pesquisas de opinião já consolidado na segunda colocação e após o debate organizado pela *Rede Globo*, houve uma mudança para nas emoções por ele transmitidas: o foco passou a ser negativo, possivelmente com críticas ao candidato com quem já se esperava que fosse disputar o segundo turno.

7.7 Análise do Candidato 7

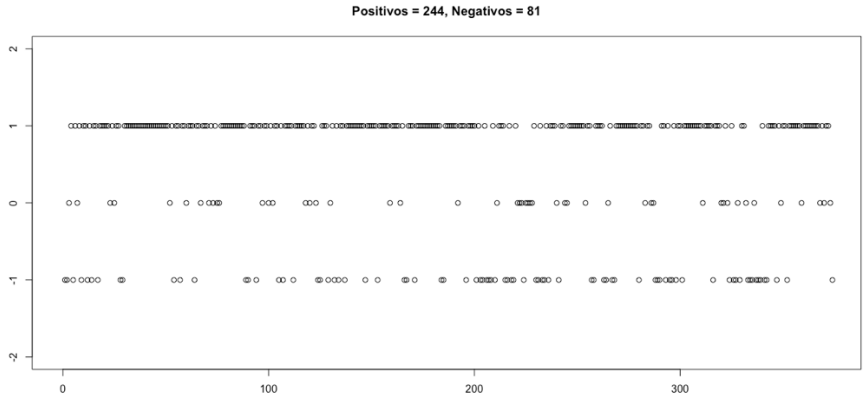
O Candidato 7 também permite poucas análises a partir dos conteúdos por ele compartilhados no *Twitter*. Foi, junto com o Candidato 5, um dos que mais tendia a disseminar sentimentos positivos pela rede social (ver Figura 6.51), tanto antes quanto durante a campanha eleitoral. Assim como o Candidato 5, esteve praticamente ausente do *Twitter* antes das eleições, com aumento da frequência de publicações praticamente exponencial durante esse período (ver Figura 7.14) e a tendência foi compartilhar conteúdo com sentimentos negativos após o debate da *Rede Globo* e retornar ao positivo nos últimos dias (ver Figura 7.15). Com isso, não houve sinais de influências de eventos externos, como as pesquisas de opinião.

Figura 7.14 - Distribuição temporal dos *tweets* do Candidato 7



Fonte: arquivo pessoal

Figura 7.15 - Categorização temporal dos *tweets* do Candidato 7, durante a campanha



Fonte: arquivo pessoal

7.8 Análise geral

Com base nas análises individuais, pode ser feita uma análise geral procurando relações entre os diferentes perfis (frequência, sentimento) de publicações dos candidatos com os resultados das eleições do primeiro turno, realizadas no dia 7 de outubro.

Inicialmente, destaca-se que os candidatos que se mostraram mais flexíveis a eventos externos (principalmente, os resultados de pesquisas de opinião), ou seja, que adaptavam os sentimentos por ele transmitidos conforme esses acontecimentos, tiveram bons resultados nas eleições. No total, foram três que apresentaram esse perfil (Candidatos 1, 3 e 6) e que, ao final, estiveram entre as 4 primeiras colocações (incluindo as duas primeiras).

Outra conclusão importante é que não há relação direta entre quantidade de publicações e a influência nos votos. Muitos candidatos estiveram relativamente ausentes da rede social ao longo do período pré-campanha e, com o início do período eleitoral, aumentaram consideravelmente a quantidade de *tweets*. Mesmo assim, isso não se refletiu em resultados. Além disso, todos os candidatos buscaram expressar um discurso mais positivo nos dias que antecederam às eleições, com o intuito de "puxar" os últimos votos do eleitorado. Outro ponto de análise é que, dos quatro candidatos que mostraram crescimento entre as primeiras pesquisas de opinião e o resultado final, três apresentaram sentimentos balanceados ao longo da campanha, ou seja, com distribuição aproximada entre publicações com emoções negativas e positivas.

Por fim, os dois candidatos que obtiveram mais votos destacaram-se por trocar de emoções com o início da campanha: enquanto ambos focavam mais em sentimentos negativos (críticas, demonstrar medo, focar no negativo do país e nos outros candidatos) no começo do ano, durante o período eleitoral, isso passou a ser mais balanceado, com o sentimento positivo (foco em mudanças, propostas e em si mesmo) predominando.

Dito isso, nota-se que a teoria de estratégias emocionais da Neuropolítica possui influência nas campanhas eleitorais, mesmo que inconscientemente. Cada candidato possuiu um perfil específico de emoções a serem transmitidas, com o objetivo de angariar mais votos: uns focam em repassar sentimentos positivos, como confiança e esperança; outros, procuram transmitir sentimentos negativos, como medo e críticas, posicionando-se como a solução aos problemas identificados. Além disso, há aqueles que adaptam seu discurso conforme a realidade atual, ou seja, de acordo com os resultados de pesquisas de intenção de voto e notícias veiculadas.

Para avaliar essa teoria, o modelo construído com base no algoritmo de Naïve-Bayes de aprendizado de máquina mostra-se apropriado. Ele pôde categorizar com boa acurácia os conteúdos publicados por todos os candidatos, principalmente após a aplicação de inúmeras técnicas de pré-processamento dos textos, e permitiu que o autor fizesse as devidas análises, posicionando este trabalho como um dos pioneiros no Brasil. O mesmo não ocorreu com a classificação com base no uso de léxicos e

com as funções prontas do *Orange*, o que comprova que, para documentos (assim como observado em outras aplicações), os algoritmos de aprendizado de máquina costumam apresentar melhor desempenho no reconhecimento de padrões.

A escolha de trabalhar com redes sociais para a análise também se mostrou válida. Além dos dados apresentados pelo Datafolha (2018), que mostram uma tendência de que o eleitorado brasileiro dependa cada vez mais da Internet, e não da televisão, para receber informações, os próprios candidatos estudados neste trabalho apresentaram essa mudança. Dos oito analisados, seis apresentaram grande aumento na frequência de publicações com o início da campanha. A opção pelo uso do *Twitter* como base, entretanto, não se mostrou ideal, já que os conteúdos compartilhados nessa rede possuem limitação de caracteres e também pois uma pequena parcela do eleitorado é usuário ativo dela (DATAFOLHA, 2018).

Ao avaliar o trabalho como uma proposta de solução para a previsibilidade eleitoral, os resultados obtidos preliminarmente mostram-se não conclusivos, mesmo com a identificação de alguns pontos em comum entre os candidatos à Presidência que lideraram as votações do dia 7 de outubro de 2018. Com os resultados deste trabalho, não é possível afirmar, por exemplo, que discursos com sentimentos positivos apelam mais aos eleitores do que aqueles com sentimentos negativos, conforme descoberto por Lavareda (2011). Contudo, há indícios de que, em trabalhos futuros, este projeto possa ser referência devido a seu caráter pioneiro na área.

8 CONSIDERAÇÕES FINAIS E PERSPECTIVAS

O presente trabalho possuiu, como objetivo, propor um novo método para previsibilidade eleitoral.

Após as previsões incorretas de pleitos no Reino Unido e nos Estados Unidos, em 2016, havia uma preocupação para que o mesmo ocorresse nas eleições presidenciais brasileiras de 2018. Outro fator de forte influência no cenário eleitoral do Brasil foi a Reforma Político-Eleitoral de 2017, que limitou os gastos de campanhas e reduziu suas durações, o que demanda novas abordagens por parte dos candidatos (como o uso mais intenso de redes sociais), tanto em 2018, quanto em eleições seguintes. A Ciência Política, por sua vez, apresenta diferentes teorias que procuram prever o comportamento eleitoral. Entretanto, estas possuem métodos de pesquisa que se limitavam, em sua maioria, à aplicação de *surveys*, sem o uso de técnicas modernas de aquisição e processamento de dados, como *machine learning*, e sem considerarem os impactos da Internet no direcionamento dos votos.

Dessa forma, propôs-se um novo método para previsão de eleições com base na teoria de estratégias emocionais da Neuropolítica, uma abordagem mais recente, pouco abordada na literatura e que avalia os sentimentos transmitidos pelos candidatos em suas campanhas. Como fonte dos dados, optou-se pelo uso de uma rede social, algo ainda pouco explorado na Ciência Política e que se mostrava cada vez mais impactante no cenário eleitoral brasileiro, com uma grande quantidade de votantes participando ativamente da rede. Dentre as diferentes possibilidades de rede social, a escolhida foi o *Twitter*, devido a restrições de acesso às demais.

Foram aplicadas diferentes técnicas de *data mining* (*web scraper,* pré-processamento de textos, tradução automática) em cima dos *tweets* e avaliadas diversas formas de analisar os conteúdos publicados pelos candidatos. Aquela que apresentou melhores resultados foi com uso do algoritmo de aprendizado de máquina de Naïve-Bayes. Após a criação de um modelo de categorização e sua aplicação em todos os tuítes, pôde-se analisar as estratégias utilizadas por todos os oito candidatos escolhidos. A partir dos resultados obtidos, foi possível diferenciar as estratégias emocionais de cada candidato.

A opção por usar das redes sociais como base de análise se mostrou válida. As limitações impostas pela Reforma Político-Eleitoral certamente aumentaram a relevância do uso da Internet pelos candidatos nas eleições de 2018 e a perspectiva é de que isso aumente progressivamente nas próximas eleições. Como a Ciência Política ainda não passou a estudar quantitativamente os efeitos das redes, este trabalho mostra-se pioneiro ao fazê-lo. Além disso, usar de *machine learning* em estudos de previsibilidade eleitoral também foi algo inovador e os resultados preliminares obtidos são promissores, apesar de não terem sido conclusivos quanto à previsão do pleito.

Uma das oportunidades para melhorias em trabalhos futuros é usar do modelo de classificação desenvolvido em outros tipos de conteúdos, como propagandas de televisão e publicações em outras redes sociais. Para campanhas feitas na televisão, já houveram outros estudos feitos até mesmo no Brasil, principalmente por Lavareda

(2011). Portanto, basear-se em publicações de outra rede social aparenta mais promissor. O *Twitter* foi a ferramenta escolhida apenas por ser a única viável a curto prazo, já que as demais não autorizaram a aquisição de seus dados. Entretanto, ela só é utilizada por uma pequena parte do eleitorado brasileiro: estima-se algo entre 10 e 15% do total, segundo pesquisa do Datafolha (2018). Essa mesma pesquisa apresentou porcentagem bem maior para outras redes, como *Facebook* (60%)*, WhatsApp* (66%) e *Instagram* (30%), tornando-as fontes mais representativas da realidade dos eleitores. Outro fator que favorece outras redes, principalmente *Facebook* e *Instagram*, é a quantidade máxima de caracteres de *tweets*, limitada a 280, enquanto *posts* dessas redes não possuem essa restrição. Com publicações com mais caracteres, estima-se possuir mais conteúdo a ser analisado, o que favorece o desempenho (acurácia) de um modelo criado com *machine learning*.

Outra possibilidade é a de considerar as demais teorias clássicas da Ciência Política, junto da teoria de estratégias emocionais, através da inclusão de outras fontes de dados, o que também seria pioneiro no Brasil. O processo de decisão do voto é muito complexo e, possivelmente, uma função multivariável. Não é apenas o discurso de um candidato que vai decidir o voto de uma pessoa, assim como não é apenas sua classe social que o fará. Dentro da Ciência Política, não há consenso quanto à teoria mais adequada, e todas possuem evidências de sua validade. Sendo assim, novos estudos poderiam ser feitos, incluindo, por exemplo, dados fornecidos pelo TSE (perfil dos eleitores e resultados das votações) ou pelo IBGE (através de seu Censo Demográfico) e, portanto, aspectos da Teoria Sociológica na análise. Neste caso, a avaliação das estratégias emocionais pode ser mantida com o método proposto por este trabalho, com uso do algoritmo de aprendizagem de máquina de Naïve-Bayes, já que apresentou resultados promissores.

Tratando de um ponto de vista pessoal, o trabalho atingiu as expectativas ao permitir que o autor desenvolvesse um projeto desafiador em uma área de interesse e que exige o domínio de complexas ferramentas de Engenharia de Controle e Automação.

REFERÊNCIAS

ACHEN, Christopher; BARTELS, Larry. **Democracy for Realists:** Why elections do not produce responsive government. Princeton: Princeton University Press, 2016.

AMENDOLA, Gilberto. Marina Silva lidera corrida eleitoral em site de apostas internacional. **O Estado de S. Paulo.** São Paulo, p. 1-2. 05 mar. 2018. Disponível em: <https://politica.estadao.com.br/noticias/geral,marina-silva-lidera-corrida-eleitoral-em-site-de-apostas-internacional,70002213850>. Acesso em: 17 nov. 2018.

AMORIM, Felipe. **Reprovação ao governo Temer sobe e tem pior índice para um presidente desde 1986, diz Ibope.**Disponível em: <https://noticias.uol.com.br/politica/ultimas-noticias/2018/06/28/reprovacao-ao-governo-temer-sobe-e-tem-pior-indice-para-um-presidente-desde-1986-diz-ibope.htm>. Acesso em: 17 nov. 2018.

BECKER, Karin; TUMITAN, Diego. **Introdução à Mineração de Opiniões: Conceitos, Aplicações e Desafios**. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 28., 2013, Recife**.** Porto Alegre: Sbbd, 2013. p. 1 - 26. Disponível em: <http://sbbd2013.cin.ufpe.br/Proceedings/artigos/pdfs/sbbd_min_02.pdf>. Acesso em: 17 nov. 2018.

CALEGARI, Luiza. **O que vai mudar com a reforma política aprovada no Congresso.** 2017. Disponível em: <https://exame.abril.com.br/brasil/o-que-vai-mudar-com-a-reforma-politica-aprovada-no-congresso/>. Acesso em: 18 nov. 2018.

CAPPRA, Ricardo. “Estratégia digital política no Brasil é ridícula”, diz gaucho que trabalhou nas campanhas de Barack Obama. Veja, São Paulo, 10 dez. 2012. Entrevista concedida a Renata Honorato.

CARREIRÃO, Yan de Souza. **A decisão do voto nas eleições presidenciais brasileiras (1989 a 1998).** 2000. 219 f. Tese (Doutorado) - Curso de Ciência Política, Universidade de São Paulo, São Paulo, 2001.

CLÉSIO, Flávio. **A toda poderosa floresta aleatória.** Disponível em: <https://mineracaodedados.wordpress.com/2014/02/15/a-toda-poderosa-floresta-aleatoria/>. Acesso em: 17 nov. 2018.

COHN, Nate. **Não culpe pesquisas pela surpresa com a saída britânica da UE.** Disponível em: <https://www1.folha.uol.com.br/mundo/2016/06/1785364-nao-culpe-pesquisas-pela-surpresa-com-a-saida-britanica-da-ue.shtml#_=_>. Acesso em: 17 nov. 2018.

COUTO, César Francisco de Moura. **Capítulo 2: Processos de Software.** Belo Horizonte: Texto, 2011. 53 slides, P&B. Disponível em:

<https://homepages.dcc.ufmg.br/~cesarfmc/classes/es/Capitulo_02.pdf>. Acesso em: 18 nov. 2018.

DATAFOLHA: quantos eleitores de cada candidato usam redes sociais, leem e compartilham notícias sobre política. Disponível em: <https://g1.globo.com/politica/eleicoes/2018/eleicao-em-numeros/noticia/2018/10/03/datafolha-quantos-eleitores-de-cada-candidato-usam-redes-sociais-leem-e-compartilham-noticias-sobre-politica.ghtml>. Acesso em: 17 nov. 2018.

EISENBERG, José M.; VALE, Teresa Cristina de S. C.. Simulação Eleitoral: uma nova metodologia para a ciência política. **Opinião Pública**, Campinas, v. 15, n. 1, p.190-223, jun. 2009.

FELDMAN, Ronen; SANGER, James. Text Mining Preprocessing Techniques. In: FELDMAN, Ronen; SANGER, James. **The text mining handbook:** Advanced approaches in analyzing unstructured data. Nova Iorque: Cambridge, 2006. p. 57-63.
FERNANDES, Cláudio; SILVA, Daniel Neves. **Brexit: a saída do Reino Unido da União Europeia**; *Brasil Escola*. Disponível em <https://brasilescola.uol.com.br/historiag/brexit-ou-saida-inglaterra-uniao-europeia.htm>. Acesso em: 17 nov. 2018.

FIGUEIREDO, Marcus Faria. **A decisão do voto:** Democracia e racionalidade. 2. ed. Belo Horizonte: Editora UFMG, 2008. 239 p.

FLOR, Daniela. **Crise nas pesquisas: prever eleições está cada vez mais difícil.** Disponível em: <https://veja.abril.com.br/mundo/crise-nas-pesquisas-prever-eleicoes-esta-cada-vez-mais-dificil/>. Acesso em: 17 nov. 2018.

FLORES, Paulo. **Redes sociais e TV: qual o peso de cada meio nas eleições de 2018.** 2018. Disponível em: <https://www.nexojornal.com.br/expresso/2018/03/18/Redes-sociais-e-TV-qual-o-peso-de-cada-meio-nas-elei%C3%A7%C3%B5es-de-2018>. Acesso em: 18 nov. 2018.

FREITAS, Cláudia. Sobre a construção de um léxico da afetividade para o processamento computacional do português. **Revista Brasileira de Linguística Aplicada**, [s.l.], v. 13, n. 4, p.1031-1059, 19 nov. 2013. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s1984-63982013005000024.

GALHARDO, Ricardo; BRAMATTI, Daniel. **Campanha à reeleição custa R$ 318 milhões.** 2014. Disponível em: <https://politica.estadao.com.br/noticias/geral,campanha-a-reeleicao-custa-r-318-milhoes,1597905>. Acesso em: 17 nov. 2018.

GONÇALVES, Alexandre L.; TODESCO, José Leomar; LEMOS, Robson Rodrigues. **Unidade 2: Classificação A.** Florianópolis: Material de aula, 2018. 75 slides, color.

GONÇALVES, Alexandre L.; TODESCO, José Leomar; LEMOS, Robson Rodrigues. **Unidade 2: Classificação B.** Florianópolis: Material de aula, 2018. 23 slides, color.

GONÇALVES, Alexandre L.; TODESCO, José Leomar; LEMOS, Robson Rodrigues. **Unidade 2: Classificação A.** Florianópolis: Material de aula, 2018. 75 slides, color.

GONÇALVES, Alexandre L.; TODESCO, José Leomar; LEMOS, Robson Rodrigues. **Unidade 3: Recuperação de Informação.** Florianópolis: Material de aula, 2018. 111 slides, color.

GURNEY, Kevin. Neural networks—an overview. In: GURNEY, Kevin. **An introduction to neural networks.** Londres: Ucl Press, 1997. p. 12-19.

G1: Após surpresa nos EUA, analistas tentam explicar falhas de pesquisas. Disponível em: <http://g1.globo.com/mundo/eleicoes-nos-eua/2016/noticia/2016/11/apos-surpresa-nos-eua-analistas-tentam-explicar-falhas-de-pesquisas.html>. Acesso em: 17 nov. 2018.

G1: Brasileiros não se sentem representados por políticos em exercício, aponta pesquisa. Disponível em: <https://g1.globo.com/politica/noticia/brasileiros-nao-se-sentem-representados-por-politicos-em-exercicio-aponta-pesquisa.ghtml>. Acesso em: 17 nov. 2018.

G1: terça-feira, 25 de setembro. Disponível em: <https://g1.globo.com/resumo-do-dia/noticia/2018/09/25/terca-feira-25-de-setembro.ghtml>. Acesso em: 17 nov. 2018.

HUFFPOST BRASIL. **Horário eleitoral: Como é feita a divisão do tempo de TV.** 2018. Disponível em: <https://www.huffpostbrasil.com/2018/07/20/entenda-como-e-feita-a-divisao-do-tempo-de-tv-no-horario-eleitoral_a_23485856/>. Acesso em: 17 nov. 2018.

KAUER, Anderson Uilian. **Análise de Sentimentos baseada em Aspectos e Atribuição de Polaridade.** 2016. 76 f. Dissertação (Mestrado) - Curso de Ciência da Computação, Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016.

KOTHARI, Keval. **What are the biggest differences between web crawling and web scraping?** Disponível em: <https://www.quora.com/What-are-the-biggest-differences-between-web-crawling-and-web-scraping>. Acesso em: 17 nov. 2018.

LAVAREDA, Antonio. Neuropolítica: O papel das emoções e do subconsciente. **Revista Usp**, São Paulo, v. 1, n. 90, p.120-146, maio 2011.

LAZARSFELD, Paul F.; BERELSON, Bernard R.; GAUDET, Hazel. **Erie County Study, 1940**. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, Disponível em <https://doi.org/10.3886/ICPSR07204.v1>. Acesso em: 17 nov. 2018.

LEI N° 13.488, de 19 de setembro de 2017. Altera a Lei nº 9.096, de 19 de setembro de 1995 (Lei dos Partidos Políticos), a Lei nº 9.504, de 30 de setembro de 1997 (Lei das Eleições), a Lei nº 4.737, de 15 de julho de 1965 (Código Eleitoral), a Lei nº 13.165, de 29 de setembro de 2015 (Minirreforma Eleitoral de 2015), e a Lei nº 5.768, de 20 de dezembro de 1971, com o fim de promover ampla reforma no ordenamento político-eleitoral. **Reforma Político-eleitoral**. Brasília, 06 out. 2017. Disponível em: <http://www2.camara.leg.br/atividade-legislativa/discursos-e-notas-taquigraficas/discursos-em-destaque/pl-8612-de-2017-promove-ampla-reforma-politico-eleitoral/projeto-de-lei-no-8-612-de-2017-reforma-politico-eleitoral>. Acesso em: 17 nov. 2018.

LEITE, Jair C. **Processo de Software.** Natal: Texto, 2008. 14 slides, color. Disponível em: <https://www.dimap.ufrn.br/~jair/ES/slides/ProcessoDeSoftware.pdf>. Acesso em: 17 nov. 2018.

MARTINS JUNIOR, José Paulo. MODELO SOCIOLÓGICO DE DECISÃO DE VOTO PRESIDENCIAL NO BRASIL 1994-2006. **Revista Debates**, Porto Alegre, v. 3, n. 2, p.68-96, dez. 2009. Disponível em: <https://core.ac.uk/download/pdf/26691795.pdf>. Acesso em: 17 nov. 2018.

NAGPAL, Anuja. **Over-fitting and Regularization.** Disponível em: <https://towardsdatascience.com/over-fitting-and-regularization-64d16100f45c>. Acesso em: 17 nov. 2018.

PIMENTEL, Jairo Tadeu Pires. **Razão e Emoção no Voto:** O caso da Eleição Presidencial de 2006. 2007. 129 f. Dissertação (Mestrado) - Curso de Ciência Política, Departamento de Ciência Política da Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2007.

RAMOS, Ricardo Argenton. **Modelos de Processo de Software.** Cuiabá: Texto, 2012. 84 slides, color. Disponível em: <http://araguaia2.ufmt.br/professor/disciplina_arquivo/100/20140922403.pdf>. Acesso em: 18 nov. 2018.

RENNÓ, Lucio; AMES, Barry. PT no purgatório: ambivalência eleitoral no primeiro turno das eleições presidenciais de 2010. **Opinião Pública**, Campinas, v. 20, n. 1, p.1-25, abr. 2014.

RIBEIRO, Alessandro Costa. **MODELO DE RECONHECIMENTO DE PADRÕES EM IDEIAS USANDO TÉCNICAS DE DESCOBERTA DE CONHECIMENTO EM TEXTOS.** 2018. 172 f. Dissertação (Mestrado) - Curso de Engenharia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2018.

SABIN, Bruce. **Review of Philip E. Converse's "The Nature of Belief Systems in Mass Publics".** Disponível em: <http://www.brucesabin.com/nature_of_belief_systems.html>. Acesso em: 17 nov. 2018.

SANTANA, Felipe. **Como Funciona o Algoritmo Naive Bayes.** 2017. Disponível em: <http://minerandodados.com.br/index.php/2017/12/14/naive-bayes-machine-learning/>. Acesso em: 17 nov. 2018.

SCHMIDT, Kiersten; ANDREWS, Wilson. **A Historic Number of Electors Defected, and Most Were Supposed to Vote for Clinton.** Disponível em: <https://www.nytimes.com/interactive/2016/12/19/us/elections/electoral-college-results.html>. Acesso em: 17 nov. 2018.

SHALDERS, André. **Eleições 2018: como partidos estão usando R$ 1,7 bilhão do fundo eleitoral para reeleger políticos tradicionais.** 2018. Disponível em: <https://www.bbc.com/portuguese/brasil-45429132>. Acesso em: 17 nov. 2018.

SHEN, Lucinda. **Citibank Says Donald Trump Now Has a Less than 1-in-3 Chance of Winning the Election.** Disponível em: <http://fortune.com/2016/10/11/presidential-election-citibank-clinton-trump/>. Acesso em: 17 nov. 2018.

SINGER, André Vitor. Raízes sociais e ideológicas do lulismo. In: SINGER, André Vitor. **Os sentidos do lulismo.** São Paulo: Companhia das Letras, 2012. Cap. 1. p. 1-41.

SOUZA, Karine França de; PEREIRA, Moisés Henrique Ramos; DALIP, Daniel Hasan. UniLex: Método Léxico para Análise de Sentimentos Textuais sobre Conteúdo de Tweets em Português Brasileiro. **Abakós**, Belo Horizonte, v. 5, n. 2, p.79-96, maio 2017.

STACK, Liam. **The 2016 Presidential Race, Explained.** 2016. Disponível em: <https://www.nytimes.com/2016/11/08/us/politics/trump-clinton-2016-presidential-race-explainers.html>. Acesso em: 18 nov. 2018.

TCKESKISS, L. A.. Lição I: Fenômenos Sociais e Acontecimentos Históricos. In: TCKESKISS, L. A.. **O Materialismo Histórico em 14 Lições.** São Paulo: Calvino Filho, 1934. p. 7-11.

TERRON, Sonia Luiza; SOARES, Gláucio Ary Dillon. As bases eleitorais de Lula e do PT: do distanciamento ao divórcio. **Opinião Pública**, Campinas, v. 16, n. 2, p.310-337, nov. 2010.

TSYTSARAU, Mikalai; PALPANAS, Themis. Survey on mining subjective data on the web. **Data Mining And Knowledge Discovery**, [s.l.], v. 24, n. 3, p.478-514, 19 out. 2011. Springer Nature. http://dx.doi.org/10.1007/s10618-011-0238-6.

VENTURINI, Lilian. **Por que o Congresso deve se renovar pouco, segundo este analista.** Disponível em: <https://www.nexojornal.com.br/expresso/2018/08/19/Por-que-o-Congresso-deve-se-renovar-pouco-segundo-este-analista>. Acesso em: 17 nov. 2018.

ZEENG Dashboard - Eleições 2018. 2018. Disponível em: <http://blog.zeeng.com.br/2018/09/25/dashboard-eleicoes/>. Acesso em: 18 nov. 2018.