UNIVERSIDADE FEDERAL DE GOIÁS
PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA ELÉTRICA
E DE COMPUTAÇÃO

MINERAÇÃO DE OPINIÃO EM MÍDIAS SOCIAIS COM APRENDIZADO DE MÁQUINA

Jhonathan de Godoi Brandão

[UFG] & [EMC]
[Goiânia - Goiás - Brasil]
19 de outubro de 2020

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

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[X] Dissertação [ ] Tese

2. Nome completo do autor

Jhonathan de Godoi Brandão

3. Título do trabalho

"Mineração de Opinião em Mídias Sociais com Aprendizado de Máquina"

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **JHONATHAN DE GODOI BRANDÃO**, **Discente**, em 01/02/2021, às 19:37, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Wesley Pacheco Calixto**, **Usuário Externo**, em 01/02/2021, às 21:35, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1843689** e o código CRC **8C26464A**.

UNIVERSIDADE FEDERAL DE GOIÁS
PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA ELÉTRICA
E DE COMPUTAÇÃO

MINERAÇÃO DE OPINIÃO EM MÍDIAS SOCIAIS COM APRENDIZADO DE MÁQUINA

Jhonathan de Godoi Brandão

Dissertação apresentada à Banca Examinadora como exigência parcial para a obtenção do título de Mestre em Engenharia Elétrica e de Computação pela Universidade Federal de Goiás (UFG), Escola de Engenharia Elétrica, Mecânica e de Computação (EMC), sob a orientação do Prof. Dr. Wesley Pacheco Calixto

[UFG] & [EMC]
[Goiânia - Goiás - Brasil]
19 de outubro de 2020

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

Brandão, Jhonathan de Godoi
Mineração de Opinião em Mídias Sociais com Aprendizado de Máquina [manuscrito] / Jhonathan de Godoi Brandão. - 2020.
70 f.

Orientador: Prof. Dr. Wesley Pacheco Calixto.
Dissertação (Mestrado) - Universidade Federal de Goiás, Escola de Engenharia Elétrica, Mecânica e de Computação (EMC), Programa de Pós-Graduação em Engenharia Elétrica e de Computação, Goiânia, 2020.
Bibliografia.
Inclui lista de figuras, lista de tabelas.

1. Análise de sentimentos. 2. Mineração de opinião. 3. Aprendizado de máquina. I. Calixto, Wesley Pacheco, orient. II. Título.

CDU 62+004+005

UNIVERSIDADE FEDERAL DE GOIÁS

ESCOLA DE ENGENHARIA ELÉTRICA, MECÂNICA E DE COMPUTAÇÃO

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **11/2020** da sessão de Defesa de Dissertação de Jhonathan de Godoi Brandão, que confere o título de Mestre em **Engenharia Elétrica e de Computação**, na área de concentração em **Engenharia de Computação.**

Aos **dezenove dias do mês de outubro de dois mil e vinte**, a partir das **14h00min**., realizou-se a sessão pública de Defesa de Dissertação intitulada **"Mineração de Opinião em Mídias Sociais com Aprendizado de Máquina"**. Os trabalhos foram instalados pelo Orientador Professor Doutor **Wesley Pacheco Calixto (EMC/UFG)** com a participação dos demais membros da Banca Examinadora: Professor Doutor **Luiz Eduardo Bento Ribeiro (ENGPROD/IFG)**, membro titular externo; Professor Doutor **Márcio Rodrigues da Cunha Reis (ENGPROD/IFG)** membro titular externo**, cuja participações ocorreram através de videoconferência**. Durante a arguição os membros da banca **não fizeram** sugestão de alteração do título do trabalho**.** A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pelo Professor Doutor **Wesley Pacheco Calixto**, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos **dezenove dias do mês de outubro de dois mil e vinte**.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Luiz Eduardo Bento Ribeiro**, **Usuário Externo**, em 20/10/2020, às 21:04, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **JHONATHAN DE GODOI BRANDÃO**, **Discente**, em 21/10/2020, às 08:36, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Wesley Pacheco Calixto**, **Usuário Externo**,

Documento assinado eletronicamente por **MÁRCIO RODRIGUES DA CUNHA REIS**, **Usuário Externo**, em 23/11/2020, às 09:21, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1621227** e o código CRC **239295A8**.

"Mais do que de máquinas, precisamos de humanidade. Mais do que de inteligência, precisamos de afeição e doçura. Sem essas virtudes, a vida será de violência e tudo será perdido."

Charles Chaplin
em *"O Grande Ditador"*, 1940.

Dedico este trabalho a meus pais, irmã, afilhadas e cunhado. Em especial, dedico a minha esposa Laís que sempre acreditou e esteve ao meu lado.

AGRADECIMENTOS

Aos iniciar novos projetos elaboramos planos perfeitos ao nossos olhos, mas muitas vezes o percusso para alcançar esse objetivo é mais árduo do que imaginamos. Passamos por processo de descoberta contínua. Nossas qualidades e defeitos são expostos com o surgimento das dificuldades. Grandes obstáculos se impõe em nosso caminho.

Em determinados momentos duvidei do encerramento com sucesso deste projeto. Por isso agradeço ter em minha vida pessoas que não me deixaram desistir nos momentos mais críticos. Gratidão incomensurável à minha querida e amada esposa Laís Ferreira pelos conselhos e apoio incondicional, não foram momentos fáceis. Meus agradecimentos ao meu orientador e mestre Wesley Pacheco por demonstrar grande coração e empatia em todo o processo. Sua orientação e dedicação tornou o trabalho possível, mas também me ensinou muito sobre como me tornar um ser humano melhor.

Agradeço aos meus pais por ter dado a criação que me ensinou os valores que me conduz todos os dias. Por me ensinarem o poder da educação e me incentivarem a buscar as oportunidades que eles não tiveram.

Meus agradecimentos à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte financeiro.

RESUMO

O objetivo deste trabalho é desenvolver ferramenta que otimize modelos de aprendizado supervisionado de máquina para classificar polaridade de opiniões em tuítes. São utilizados cinco conjuntos de dados distintos que são preparados, preprocessados e então utilizados como entrada para a etapa de treinamento e avaliação dos modelos de aprendizado de máquina. Os melhores resultados de acurácia obtidos no treinamento e avaliação dos modelos são de 82,45% para os dados sem preprocessamento × 78,83% com todos os preprocessamentos propostos para o conjunto de dados utilizando classificador Naive Bayes. Por fim, é realizada otimização hiperparamétrica dos classificadores e seleção do modelo que obtém a melhor acurácia. O modelo otimizado obtém acurácia maior que 90% para alguns conjuntos de dados. As técnicas de aprendizado supervisionados apresentam dependência de dados rotulados para treinamento, o método proposto produz desempenhos semelhantes para conjuntos de dados de tamanhos variados, o que possibilita o desenvolvimento de modelos de classificação otimizados com quantidade reduzida de dados rotulados.

OPINION MINING ON SOCIAL MEDIA WITH MACHINE LEARNING

ABSTRACT

The aim of this work is to develop a tool for that optimizes supervised machine learning models in order to classify polarity of opinions in tweets. Five different datasets are used, which are prepared, preprocessed and then used as input for the training and evaluation stage of machine learning models. The best accuracy results obtained in the training and evaluation of the models are 82.45% for the data without preprocessing $\times$ 78.83% with all the proposed preprocessing for the dataset using the Naive Bayes classifier. Finally, hyperparametric optimization of the classifiers and selection of the model that obtains the best accuracy is performed. The optimized model achieves an accuracy greater than 90% for some data sets. The supervised learning techniques depend on labeled data for training, the proposed method produces similar performances for datasets of varying sizes, which allows the development of optimized classification models with reduced amount of labeled data.

SUMÁRIO

Pág.

LISTA DE FIGURAS

Pág.

LISTA DE TABELAS

Pág.

CAPÍTULO 1

INTRODUÇÃO

Na internet as pessoas podem realizar atividades como comprar ingressos, produtos e serviços, comunicar-se em redes sociais, realizar vídeo-conferências, compartilhar ideias e conteúdos em *blogs* e *microbloggins*. Os dados gerados a partir destas interações sociais são do tipo não-estruturados, pois não possuem estrutura de organização, além de serem flexíveis e dinâmicos. Devido a quantidade e velocidade em que são gerados, surge a necessidade de ferramentas e técnicas que possam extrair, transformar, armazenar e analisá-los. A mineração de dados da *web* e de texto são combinadas para analisar diferentes tipos de dado em aplicações da vida real.

E-commerces, *blogs*, fóruns, redes sociais, *sites* de notícias são exemplos de plataformas presentes na *web* onde as pessoas expressam suas opiniões, que podem ser utilizadas para entender o que consumidores e o público em geral pensam sobre eventos sociais, políticos, marcas, produtos, serviços, campanhas de *marketing* e estratégias empresariais. Pesquisas em mineração de opinião surgem em meados dos anos 2000 com o intuito de avaliar computacionalmente opiniões, sentimentos, emoções e atitudes (RAVI; RAVI, 2015).

Mineração de opinião, também conhecida como análise de sentimentos, é a tarefa de detectar, extrair e classificar opiniões, sentimentos e atitudes sobre diferentes tópicos que estejam expressos em textos. De acordo com Liu (2012) o termo mineração de opinião foi utilizado primeiramente em Dave et al. (2003). Deste então, pesquisas são realizadas para analisar sentimentos em documentos, artigos, notícias, avaliações de produtos e serviços. A opinião de outros consumidores influencia diretamente na decisão de compra dos usuários que estão buscando e selecionando produtos ou serviços.

Hemmatian e Sohrabi (2017) afirmam que a análise de sentimento pode ser realizada a nível de documento, sentença, aspecto ou conceito. A nível de documento, o sentimento positivo ou negativo é sumarizado a partir da análise do inteiro teor do documento. Em diversas situações este nível não é adequado pois as opiniões contidas no documento são relativas a diversas entidades. Neste sentido, surgem as abordagens a nível de sentença. Nesta, primeiramente determina se é subjetiva (opinião) ou objetiva (fato) e posteriormente avalia-se a polaridade. Na abordagem a nível de aspecto o foco é a opinião expressa, independente de qual constructo será considerado: documento, sentença ou oração. O nível conceitual é baseado na inferên-

cia de informações conceituais sobre emoções e sentimentos associados à linguagem natural.

Pang et al. (2002) realizam a análise de sentimentos a nível de documento em resenhas de filmes obtidas no *site imdb.com* (IMDb - *Internet Movie Database*) classificando-os em positivo ou negativo. O intuito do estudo é comparar o desempenho da análise de sentimentos para categorização de documentos, utilizando o método baseado em tópicos, onde os documentos são categorizados de acordo com seu assunto, como esportes, economia, turismo entre vários outros. Para a análise de sentimentos, os autores utilizam as técnicas de aprendizado de máquina Naive Bayes, máxima entropia e máquinas de vetores de suporte que resultam em acurácias de 72,8% a 82,9%. Os resultados obtidos não superam a categorização baseada em tópicos, que obtém 90% de acurácia em outros trabalhos relacionados. Os autores consideram a tarefa de análise de sentimentos como campo de estudo desafiador.

No estudo de Turney (2002) é utilizado aprendizado de máquina para analisar sentimentos a nível de documento em resenhas de usuários da internet. O autor desenvolve classificador que prediz a polaridade, positiva ou negativa, de avaliações do *site epinions.com*. A acurácia obtida é de 74% para a base de 410 resenhas em quatro diferentes domínios (automóveis, bancos, filmes e destinos turísticos). O autor avalia o desempenho do algoritmo para cada base isoladamente, a menor acurácia obtida é para a base de filmes com 66% e a maior 84% para a base de automóveis. O motivo para o baixo desempenho em resenhas de filme, é que o todo da opinião avaliada não necessariamente é a soma de suas partes.

Outra abordagem para a mineração de opinião é a nível de sentença, em contrapartida aos métodos a nível de documentos apresentados em Pang et al. (2002) e Turney (2002). Nasukawa e Yi (2003) abordam a análise de sentimentos associados a polaridades positivas e negativas de assuntos específicos contidos em documento. É utilizada análise semântica com analisador sintático e léxico de sentimento com o intuito de identificar as relações semânticas entre o assunto e as expressões de sentimento. Os resultados de precisão deste estudo são de 75% a 95%, demonstrando que há utilidade na avaliação dos sentimentos na maioria dos textos analisados pela abordagem a nível de sentença.

Com o aumento de *sites* onde os usuários podem avaliar e escrever resenhas de produtos ou serviços, surge a necessidade de identificar avaliações que não apresentem opinião expressa sobre a qualidade do produto ou avaliações com ruídos. Este tipo de avaliação não contribui para quem busca a opinião para decisão de compra. Liu

et al. (2007) propõem abordagem para tratar/filtrar a qualidade das avaliações utilizando aprendizado de máquina, na qual os autores obtêm acurácias de 72,81% a 83,94% na detecção e sumarização de resenhas de baixa qualidade.

A expansão das redes sociais, como *Twitter*[1], permite que as pessoas emitam opinião sobre os mais diversos assuntos, marcas, produtos ou serviços. Isso leva ao desenvolvimento de trabalhos como o de Barbosa e Feng (2010). Os autores analisam a polaridade das opiniões de tuítes dos usuários, mensagens do *Twitter*, através de aprendizado de máquina. Para o desenvolvimento dos classificadores eles utilizam o software *Waikato Environment for Knowledge Analysis* (Weka). Na metodologia proposta divide-se os textos em subjetivos ou objetivos, e posteriormente classificam a polaridade dos tuítes opinativos em **positiva** ou **negativa**. O estudo obteve acurácia de 55,5% a 88%. A principal limitação da abordagem proposta encontra-se nos casos em que as sentenças possuem sentimentos antagonistas[2].

Ahuja et al. (2019) avaliam o impacto da extração de atributos para análise de sentimentos. No estudo, os autores utilizam seis técnicas de preprocessamento e duas para a extração de atributos com o intuito de melhorar o desempenho dos classificadores árvore de decisão, máquina de vetores de suporte (*Support Vector Machine* - SVM), K vizinhos mais próximos (*K-Nearest Neighbor* - KNN), floresta aleatória, regressão logística e *Naive Bayes*. Nas simulações realizadas neste estudo, obteve-se resultados de acurácia de 46% a 57%. O classificador de regressão logística tem melhor desempenho comparada as outras abordagens utilizadas.

Estudos com o intuito de otimizar os resultados obtidos pelos classificadores de aprendizado de máquina como o de Singh et al. (2017) propõe utilizar as técnicas estatísticas de redução de dimensionalidade *document frequency* (DF), *mutual information* (MI) e *information gain* (IG) para selecionar os atributos de treinamento de classificadores de aprendizado de máquina. Três conjuntos de dados são utilizados para treinamento e teste de quatro classificadores, resultando em acurácias de 60,56% a 87,65%. Os autores sugerem, para estender este estudo, o uso de redes neurais para a representação das palavras na etapa de preprocessamento dos dados.

Souza et al. (2018) comparam o desempenho de três algoritmos de otimização por enxame de partículas com duas técnicas de aprendizado de máquina para o agrupamento de opiniões em três bases de dados com diferentes níveis de complexidade.

[1] www.twitter.com

[2] Que ou aquele que é e age contrário ou em sentido oposto a alguém ou alguma coisa: adversário, opositor.

As acurácias obtidas pelos algoritmos baseados em enxame de partículas são menores que as com classificadores de aprendizado supervisionado para os conjuntos de dados analisados. Porém, para os autores os algoritmos propostos são competitivos considerando que os dados não precisam ser rotulados para o treinamento e nem do uso de dicionários léxicos.

Ravi e Ravi (2015) expõe a necessidade do desenvolvimento de sistemas para o preprocessamento automático de dados não-estruturados, pois esta tarefa consome elevado tempo da atividade de mineração de opinião. O estudo também afirma que poucas pesquisas exploram o potencial das técnicas de otimização para a seleção de atributos. Os autores propõe a utilização de sistemas híbridos, que combinam técnicas de aprendizado de máquina e otimização. Para Ravi e Ravi (2015) o maior desafio é a análise de sentimentos em domínios cruzados. Sentenças como: **A tela é curva** tem polaridade positiva para televisores, mas negativa para dispositivos móveis.

Hemmatian e Sohrabi (2017) avaliam que são necessários estudos utilizando redes neurais convolucionais focados na identificação de comentários neutros e na melhoria do desempenho dos modelos de mineração de opinião em grandes conjuntos de dados. Os autores apontam para a necessidade de desenvolver estudos para a mineração de opinião a nível conceitual. Além de combinar esta abordagem com os outros três níveis de análise.

Vários são os trabalhos desenvolvidos com o intuito de analisar, estratificar e separar opiniões em textos. No entanto, ainda há a lacuna para desenvolver modelos otimizados que classifique polaridade de opiniões em tuítes. Os tuítes possuem limitação de 280 caracteres para cada mensagem com elevado fluxo de dados sobre os mais diversos assuntos, dificultando a análise e elevando o processamento. Somado a isso, tem-se novas ferramentas de programação com desempenho ainda não analisado/comparado com os conjuntos de dados disponíveis na literatura. Desta forma, o estudo comparativo de técnicas e ferramentas adicionais justificam este trabalho.

De posse dos conjuntos de dados com opiniões rotuladas dos tuítes, do conhecimento para aplicação de métodos que possam analisar, estratificar e classificar a polaridade das opiniões e ligado as novas ferramentas e bibliotecas de *software* existentes no mercado, descreve-se a hipótese primária deste trabalho: **se** é possível obter o conjunto de dados dos tuítes e **se** é possível realizar análise exploratória do desempenho de diversos classificadores, **então** é possível desenvolver modelo otimizado para classificar a polaridade da opinião dos usuários de mídia social utilizando aprendizado

supervisionado de máquina.

O objetivo geral deste trabalho é desenvolver modelo otimizado para aprendizado de máquina à ser aplicado na mineração de opinião em tuítes. Ainda como objetivos específicos têm-se: i) avaliar o desempenho de classificadores Naive Bayes, máquina de vetor de suporte, árvore de decisão e K-vizinhos mais próximos, ii) comparar os resultados obtidos em cinco bases de dados selecionadas em trabalhos correlacionados que utilizaram outros métodos e ferramentas para a classificação e iii) aplicar busca em grade para realizar otimização hiperparamétrica e selecionar modelo de aprendizado de máquina otimizado para a mineração de opinião.

A motivação para o trabalho surge a partir da necessidade de compreender como a opinião de indivíduos nas mídias sociais tem influência nas decisões de outros usuários e no comportamento coletivo. O monitoramento do sentimento das pessoas sobre marcas, produtos e serviços auxilia organizações a tomarem decisões estratégicas na elaboração de novas campanhas de *marketing*, posicionamento de marca, resolução de crises e no desenvolvimento de novos produtos ou serviços. A opinião dos usuários é importante não somente para empresas, mas políticos, pessoas e organizadores de eventos esportivos ou sociais podem utilizar da vantagem desta técnica, como por exemplo no monitoramento da opinião de eleitores sobre temas que podem impactar na decisão de voto em período eleitoral.

O estudo de Lima (2016) utiliza a análise de sentimentos como um dos aspectos para o que é definido como tríade da *persona virtual*. Traçar a *persona* dos consumidores é fator chave para definir estratégias de mercado bem sucedidas. A autora utiliza conjuntos de dados rotulados manualmente que estão disponíveis na literatura, entre eles Sanders, SS-Twitter e Sentiment140 que também são analisados neste trabalho. Para a classificação de polaridade de opiniões com aprendizado supervisionado de máquina é necessário conjunto de dados rotulados para o treinamento dos modelos. A anotação destes dados geralmente é realizada por especialistas do domínio que analisam as mensagens extraídas das mídias sociais e determinam se são positivas, negativas ou neutras de acordo com critérios definidos.

O trabalho está estruturado de maneira a apresentar o referencial teórico compreendido no desenvolvimento da metodologia proposta. No Capítulo 2 são apresentados os conceitos relativos a mineração de opinião e algumas bases de dados relevantes desta área. No Capítulo 3 é apresentada técnica de aprendizado de máquina e os classificadores utilizados no estudo. O Capítulo 4 apresenta a metodologia proposta e o Capítulo 5 os resultados obtidos a partir do que foi proposto. No Capítulo 6 as

conclusões e sugestões para trabalhos futuros.

CAPÍTULO 2

MINERAÇÃO DE OPINIÃO

Este capítulo apresenta e conceitua a mineração de opinião e análise de sentimentos, as áreas de aplicação e as técnicas utilizadas na abordagem dos problemas. São apresentados alguns conjuntos de dados de tuítes disponíveis na literatura e os resultados dos respectivos estudos.

2.1 Opinião

A opinião é o juízo que se forma de alguém ou de alguma coisa, é a manifestação das ideias individuais a respeito de alguém ou algo (MICHAELIS, 2019). As opiniões influenciam o comportamento das pessoas, sendo essenciais em diversas atividades humanas. Indivíduos querem saber a opinião de outras pessoas antes de tomar a decisão de comprar determinado produto ou ainda opiniões sobre candidatos políticos antes de tomar a decisão de voto em eleições. No passado, pessoas buscavam opiniões perguntando a amigos e familiares. Organizações conduziam pesquisas de opinião e grupos de foco para entender o sentimento de consumidores e do público em geral sobre seus produtos ou serviços (POZZI et al., 2016).

Atualmente existem diversos espaços na *web* onde pessoas compartilham avaliações e discutem suas opiniões sobre produtos. Desta maneira, pessoas e organizações tem informações disponíveis em abundância para auxiliar sua tomada de decisão. Contudo, encontrar e monitorar opiniões em *sites* na *web* é tarefa trabalhosa devido a rápida proliferação de novos *sites* a todo instante. Assim, sistemas automáticos de mineração de opinião se tornam necessários (HEMMATIAN; SOHRABI, 2017).

Mineração de opinião ou análise de sentimentos são nomes utilizados para se referir à tarefa de analisar sentimentos, avaliações, opiniões, atitudes e emoções de pessoas a cerca de produtos, serviços, políticos, eventos e organizações. Extração de opinião, mineração de sentimento, análise de subjetividade e análise de emoções são exemplos de nomes encontrados na literatura para o problema de análise de sentimentos. Contudo, atualmente todas as tarefas deste campo de estudo estão inclusos na análise de sentimento (LIU, 2012; POZZI et al., 2016).

O intuito da mineração de opinião é analisar opiniões que expressam ou insinuam sentimentos com polaridades **negativas**, **positivas** ou **neutras**, ou com diferentes graus de força e intensidade em escala entre -5 e +5. Liu (2012) define **opinião** como

quíntupla, (e, a, s, h, t) em que e é o nome da entidade[1], a é um ou mais aspectos da entidade e, s é o sentimento sobre os aspectos a da entidade e, h o detentor da opinião e t o tempo em que a opinião é expressa por h. Assim, a quíntupla refere-se ao sentimento de indivíduos, expresso por determinado termo polarizado, a cerca de determinada entidade e e suas características a em determinado tempo t (LIU, 2012; POZZI et al., 2016).

As opiniões são agrupadas em: i) regulares: que podem ser opiniões diretas ou indiretas, ou com base em seus efeitos a cerca da entidade e ii) comparativas: que são relação de similaridades ou diferenças entre duas ou mais entidades, ou ainda a preferência do detentor da opinião com base em aspectos compartilhados pelas entidades (POZZI et al., 2016). Além disto, diferenciam-se em explícitas e implícitas. As opiniões explícitas são declarações subjetivas que fornecem opiniões regulares ou comparativas. Por outro lado, as opiniões implícitas são declarações objetivas que implicam opinião regular ou comparativa que geralmente expressam fato desejável ou indesejável. Opiniões explícitas são fáceis de detectar e classificar em comparação com as implícitas (LIU, 2012).

Em geral a mineração de opinião é realizada em três níveis diferentes, que são definidos de acordo com a granularidade do texto escolhido para a análise: i) a nível de documento a tarefa é classificar a opinião, como positiva ou negativa, do texto considerando o documento inteiro, por exemplo, dada a avaliação sobre determinado hotel, o sistema avalia se todos expressam opinião positiva ou negativa, sem considerar os possíveis aspectos, ii) em nível de sentença cada mensagem é avaliada como positiva, negativa ou neutra, na qual a suposição é que na mensagem cada sentença representa única opinião sobre a entidade e iii) a nível de entidade ou aspecto, a granularidade que o texto é analisado é menor que em nível de documento e sentença. Esta última é baseada na ideia que a opinião consiste do sentimento e do alvo (POZZI et al., 2016).

2.2 Aplicações da mineração de opinião

A linguística e o processamento de linguagem natural (PLN) são áreas de pesquisa com longo histórico. Porém, o estudo sobre sentimento e opinião de pessoas, inicia e ganha força a partir dos anos 2000, tornando-se área de pesquisa ativa. O crescimento deve-se a diversos fatores, sendo o primeiro a aplicabilidade em diversos domínios como vendas, *marketing*, campanhas políticas, relações públicas, mercado financeiro,

[1]Produto, serviço, tópico, questão, pessoa, organização ou evento.

avaliação de produtos ou serviços entre outros. Isso leva ao segundo fator, que é o surgimento de novos e desafiantes problemas de pesquisa não estudados antes e o terceiro fator que é o surgimento das mídias sociais, que proporciona pela primeira vez na história da humanidade, elevados volumes de dados opinativos (LIU, 2012).

O início e rápido crescimento da mineração de opinião coincide com a proliferação das mídias sociais na *web*, tornando-se área central de pesquisas neste contexto. Desta maneira, a mineração de opinião não tem somente impacto em PLN, como também em ciências políticas, econômicas, administrativas, sociais e todas outras áreas afetadas por opiniões de indivíduos. A mineração de opinião em mídias sociais é o campo dotado de conceitos interdisciplinares que integra teorias sociais com métodos computacionais para analisar como os indivíduos interagem e como as comunidades se formam (LIU, 2012; ZAFARANI et al., 2014; NHACUONGUE, 2015).

Várias pesquisas aplicadas são realizadas em mineração de opinião e análise de sentimentos. Hong e Skiena (2010) estudam a relação entre apostas na Liga Esportiva Profissional de Futebol dos Estados Unidos e a opinião pública em *blogs* e *Twitter*. Os estudos de O'Connor et al. (2010), Araújo et al. (2013), Becker et al. (2013), Lima (2016) e Bordin Junior (2018) também têm foco na análise de mensagens publicadas nas redes sociais, em especial no *Twitter*.

Outras pesquisas avaliam a opinião de consumidores com o intuito de criar e aprimorar campanhas de *marketing*. Liu et al. (2007) propõe modelo para prever desempenho de vendas. Em McGlohon et al. (2010) avaliações são utilizadas para determinar a posição de vendas dos produtos e comerciantes. Bugeja (2014) e Rambocas e Pacheco (2018) estudam a análise de sentimentos no *Twitter* aplicada a pesquisas de *marketing*. Trabalhos como de Bollen et al. (2011), Feldman et al. (2011) e Mittal e Goel (2012) utilizam análise de sentimentos na tentativa de prever o comportamento do mercado financeiro para auxiliar a tomada de decisão de investidores.

É crescente o interesse em pesquisas com foco em processos eleitorais. Em Yano e Smith (2010) é proposto método para prever o volume de comentários em *blogs* políticos. Chen et al. (2010) utilizam opinião para prever ponto de vista político. Os trabalhos de Tumasjan et al. (2010), Wani e Alone (2015) e Salunkhe et al. (2017) estudam a predição do resultado de eleições com base na opinião pública sobre políticos e campanhas.

2.3 Mineração de opinião em tuítes

O uso de dispositivos móveis como segunda tela e aplicativos de redes sociais levam ao fenômeno social onde elevadas quantidades de dados são gerados pelos usuários de mídias sociais enquanto acompanham eventos sociais televisivos. Neste contexto, organizações encontram dificuldades em descobrir quais conteúdos são relevantes ao seu interesse. Para isso, são necessárias ferramentas que possam analisar esta massiva fonte de dados não processados e extrair padrões relevantes (ZAFARANI et al., 2014).

O *Twitter* se destaca como uma das principais redes sociais em que os usuários compartilham suas opiniões. As mensagens compartilhadas nesta rede se chamam tuítes e são na maioria do tipo textual. Sendo assim, as aplicações práticas de mineração de opinião em tuítes fazem parte também da área de mineração de textos e PLN. O intuito é extrair os textos publicados pelos usuários que contenham ideias e opiniões a respeito de determinado assunto, classificar e quantificar estas opiniões. Os textos extraídos possuem elevada variedade de emoções expressas com *emoticons*[2] ou símbolos, além de várias gírias (LIMA, 2016; BRITO, 2017).

2.4 Abordagens utilizadas

A classificação de polaridade em mineração de opinião pode ser dividida em abordagens de aprendizado de máquina, baseada em léxico e híbrida. Classificadores de aprendizado de máquina entregam maior acurácia, enquanto os baseados em léxico produzem maior generalização. A abordagem híbrida combina ambas abordagens e os léxicos de sentimento comumente apresentam papel fundamental nos métodos de mineração de opinião (MEDHAT et al., 2014; RAVI; RAVI, 2015).

2.4.1 Mineração de opinião baseada em aprendizado de máquina

Mitchell (1997) afirma que o aprendizado de máquina lida com a questão de construir programas de computador que melhoram automaticamente com a experiência. Na mineração de opinião através do aprendizado de máquina utiliza-se modelos para prever a polaridade de textos opinativos. As técnicas utilizadas são divididas em supervisionadas e não supervisionadas. Em soluções supervisionadas emprega-se mensagens rotuladas como entrada para o treinamento de classificadores, que posteriormente devem realizar a classificação de mensagens desconhecidas. Métodos não

[2] Representação das emoções (expressão facial) pela junção de ícones ou de caracteres que estão disponíveis no teclado do computador, utilizados em bate-papos e mensagens em redes sociais.

supervisionados são empregados quando tem-se dificuldade em encontrar mensagens rotuladas (MEDHAT et al., 2014).

A mineração de opinião é em essência problema de categorização de texto, porém ao invés de categorizar os textos por tópico, classifica-se a polaridade do texto. Sendo assim, pode-se usar qualquer método de aprendizado de máquina. Na literatura destacam-se metologias como: máquina de vetores de suporte (*Support Vector Machine* - SVM), Naive Bayes, máxima entropia (ME), K-vizinhos mais próximos (K-*Nearest Neighbor* - KNN), regressão logística (RL), floresta aleatória entre outras (LIU, 2012; BORDIN JUNIOR, 2018).

2.4.2 Mineração de opinião baseada em dicionários léxicos

Nesta abordagem, utiliza-se dicionários léxicos de sentimento para realizar a classificação. Estes dicionários são compostos por conjuntos de palavras e suas polaridades. Desta maneira, a polaridade é atribuída se o texto tem palavras com a respectiva polaridade. Os dicionários frequentemente possuem a intensidade do sentimento de cada palavra e o sentimento da mensagem é calculado com o somatório das polaridades, considerando o peso de cada palavra (LIMA, 2016; BORDIN JUNIOR, 2018).

2.5 Conjuntos de dados para mineração de opinião

São vários os conjuntos de dados disponíveis para utilização em trabalhos de mineração de opinião. Os amplamente utilizados são: i) **Sentiment140**, ii) **Stanford Sentiment (STS)** e iii) **Health Care Reform (HCR)**, criados e disponibilizados por Go et al. (2009), iv) **SentiStrenght Twitter Dataset (SS-Twitter)**, disponibilizada pelos desenvolvedores da ferramenta SentiStrenght, v) **Sanders**, desenvolvida por Nike Sanders (GO et al., 2009; LIMA, 2016). A Tabela 2.1 dispõe a síntese dos cinco conjuntos de dados considerando somente os rótulos **positivos**, **negativos** e **neutros**.

Tabela 2.1 - Síntese de alguns conjuntos de dados.

Conjuntos	Positivo	Negativo	Neutro	Total
Sentiment140	182	177	139	498
STS	108	75	33	216
HCR	211	279	124	614
SS-Twitter	1340	949	1953	4242
Sanders	519	572	2333	3424

O conjunto de dados Sentiment140 é dividido em duas partes: treinamento e teste. O conjunto de dados de treinamento contém 1,6 milhão de tuítes capturados entre 6 de abril a 25 de junho de 2009. O conjunto dados de teste possui 498 tuítes, coletados em 14 de junho de 2009, sendo 359 com sentimento **positivo** ou **negativo** (GO et al., 2009).

O conjunto de dados STS possui 216 tuítes com rótulos anotados manualmente, sendo *zero* para **negativo**, quatro para **positivo** e dois para **neutro**. Este conjunto de dados é também versão do corpo de teste da base Sentiment140. A diferença é que são extraídos do *Twitter* dados do dia 25 de maio 2009 (GO et al., 2009). O conjunto de dados Sanders possui total de 5513 tuítes que têm os rótulos de sentimentos atribuídos manualmente. Os termos utilizados para a busca são @apple, #google, #twitter e #microsoft. Este conjunto de dados possui 3424 tuítes com os rótulos **positivo**, **negativo** ou **neutro** (LIMA, 2016).

Para o conjunto de dados HCR, a pesquisa é realizada pela *hashtag* #hcr com tuítes coletados em março de 2010. Os rótulos são anotados manualmente pelos autores, sendo: **positivo**, **negativo**, **neutro**, **irrelevante** e **outros**. Os conjuntos de dados são divididos em dados de treinamento, desenvolvimento e teste. A base de testes tem 614 registros rotulados como **positivo**, **negativo** ou **neutro** (GO et al., 2009). O conjunto de dados SS-Twitter possui 4242 tuítes sem a descrição do período, sem os termos utilizados para a coleta e com mensagens com sentimentos positivo, negativo ou neutro (LIMA, 2016). Os atributos coletados nesta base são força média positiva, negativa e a mensagem do tuítes, em que a polaridade é determinada por um par de valências indicando a média de força positiva e negativa.

2.6 Considerações finais

A mineração de opinião é a área de pesquisa capaz de extrair informações relevantes em grandes conjuntos de dados. Para isto são necessárias técnicas de classificação capazes de realizar esta tarefa de maneira automática. No próximo capítulo serão apresentadas as técnicas de aprendizado de máquina comumente utilizadas para a de mineração de opinião.

CAPÍTULO 3

TÉCNICAS DE APRENDIZADO DE MÁQUINA

Neste capítulo são apresentadas técnicas de preprocessamento, extração e seleção de atributos utilizadas em mineração de opinião para o treinamento dos algoritmos de aprendizado de máquina. São descritas as técnicas Naive Bayes, máquina de vetores de suporte e K-vizinhos mais próximos. Também é descrita técnica de otimização de hiperparâmetros em grade para busca e seleção de modelo ótimo de classificação. Além disto, são apresentadas as principais métricas de avaliação de desempenho da classificação.

3.1 Preprocessamento de textos

É comum que os conjuntos de dados de tuítes não estejam prontos para que as técnicas de mineração de opinião sejam aplicados neles. Os tuítes são dados não-estruturados, por isso algumas ações de preprocessamento de texto são necessárias. O método de preprocessamento mais comum envolve quatro etapas: i) **tokenização**, ii) **remoção de *stopwords***, iii) ***stemming*** e iv) **extração de atributos**. O resultado da aplicação deste fluxo de preprocessamento é a matriz **atributo-valor** relevante ao problema, onde as linhas representam os documentos (tuítes) e cada coluna representa o termo com seu respectivo peso, calculado a partir da frequência que aparece na mensagem (CAMILO; SILVA, 2009; LIMA, 2016).

A tokenização é definida como processo de análise léxica. Neste sentido, a tokenização permite obter todas as palavras que foram utilizadas no conjunto de dados e agrupá-las em *tokens* ou termos. O termo pode ser representado por uma palavra (1-grama) ou conjuntos de palavras (2, 3 ,..., n-grama) na etapa de extração de atributos. Durante esta etapa pode ser necessário realizar a limpeza dos dados removendo caracteres como sinais de pontuação ou a substituição de termos como menções a usuários e *urls* que não tenham valor para a análise, sendo trocados pelos termos **URL** e **USUARIO**. Esta remoção ou substituição ocasionam impacto na redução de dimensionalidade (TORRES et al., 2012; LIMA, 2016).

O tratamento de negações pode ser realizado na tokenização. A maneira mais simples de realizar esta ação é a remoção das palavras na etapa de *stopwords*. Outra forma de tratar esta situação é a junção dos *tokens* de negação com as palavras que o sucedem, na qual a frase: **This is not good** é tokenizada como: **This, is, not_good**.

Depois dos dados tokenizados, pode-se realizar a técnica de remoção de *stopwords*,

que é utilizada com o intuito de remover palavras que guardam informações irrelevantes sobre o contexto e que são úteis apenas para a compreensão geral do texto. São exemplos de *stopwords* artigos, preposições, advérbios, conjunções, pronomes e pontuação da língua. A aplicação desta técnica implica na diminuição dos vetores de atributos, melhorando o desempenho na etapa de classificação. Existem diversas listas de *stopwords* disponíveis na internet, mas dependendo do contexto pode ser necessário elaborar lista própria para que não se perca a acurácia do método de mineração aplicado (LIMA, 2016; BRITO, 2017).

Na técnica de *stemming* é realizada a normalização linguística do termo, removendo os prefixos e sufixos para encontrar a raiz, eliminando plurais e tempos verbais. Com isto, é possível agrupar palavras com o mesmo significado conceitual, permitindo que os vetores de atributos sofram redução de dimensionalidade por conta da diminuição do número de palavras distintas, aumentando a frequência de cada termo. Em processos de *stemming* elaborados, que não mudam a essência original da palavra, é possível diminuir consideravelmente o esforço computacional e o tamanho do léxico, obtendo maior precisão dos resultados (BRITO, 2017).

3.2 Extração de atributos

A última etapa é a extração de atributos utilizando técnica para representação do texto (tuíte ou documento) no multiconjunto de suas palavras. A técnica mais utilizada é a saco-de-palavras (*bag-of-words*), onde cada texto é representado pelo conjunto de pares (termo, peso) agrupados em estrutura matricial na qual cada termo recebe determinado peso. Os termos são os *tokens* compostos por 1, 2, 3, ..., n-grama (SCHUTZE, 2008; LIMA, 2016).

O peso do termo no saco-de-palavras é o número de vezes que cada *token* aparece no documento. Porém, esta representação nem sempre avalia com exatidão a importância de cada palavra, impactando no desempenho do classificador. A técnica frequência do termo-inverso da frequência nos documentos (*term frequency-inverse document frequency* – TF-IDF), bastante utilizada na área de recuperação de informação também tem uso em aprendizado de máquina como apresentado por Joachims (1997). Esta técnica calcula a importância do termo no documento atribuindo determinado peso $w_{i,j}$, que é calculado por (AIZAWA, 2003; ROBERTSON, 2004):

$$w_{i,j} = tf_{i,j} \cdot idf_j, \tag{3.1}$$

onde $tf_{i,j}$ é a frequência da palavra j no documento i e idf_j é a frequência inversa da palavra j em todo o documento, dado por:

$$idf_j = \log \frac{|D|}{|(document \in D | j \in document)|}, \tag{3.2}$$

conceitualmente, idf_j é o logaritmo do número total de documentos dividido pelo número de documentos que contém a palavra j. A técnica atribui peso alto a palavras que são menos frequentes em todos os documentos, e que ao mesmo tempo, tem alta frequência no documento que é usada. Desta forma, palavras com valores de TF-IDF altos podem ser usadas como exemplos representativos dos documentos aos quais pertencem, enquanto *stopwords*, que são comuns em todos os documentos recebem pesos menores (ZAFARANI et al., 2014).

3.3 Seleção de atributos com algoritmo Qui-quadrado

As técnicas de seleção de atributos são utilizadas para encontrar os dados mais relevantes de determinado conjunto de dados. A aplicação destas técnicas reduz a dimensionalidade dos dados, permitindo diminuir o custo computacional de treinamento dos modelos de aprendizado de máquina, além de aumentar o desempenho da classificação. O intuito é selecionar os dados que mais contribuem para o desempenho do classificador de aprendizado de máquina, sem alterar as características originais dos dados.

O algoritmo Qui-quadrado (χ^2), utilizado em modelos de classificação, é baseado no método χ^2 que é também conhecido como teste de independência χ^2. Este método avalia a dependência entre variáveis estocásticas, retirando os atributos com maior probabilidade de serem independentes da classe objetivo e portanto, irrelevantes para a classificação. Através da tabela de contingência, ou tabela de dupla entrada, com os valores observados O_i e esperados E_i, é calculado o valor de χ^2 do atributo d, dado pela expressão (3.3). Dada a hipótese H_0 de que o atributo d e a classe y são independentes, valores altos de χ^2 indicam que H_0 não foi satisfeita e portanto, o atributo d é relevante e pode ser útil para a classificação (BARNARD, 1992; HUAN LIU; SETIONO, 1995).

$$\chi^2_d = \sum \frac{(O_i - E_i)^2}{E_i} \tag{3.3}$$

3.4 Algoritmos de aprendizado supervisionado de máquina

O estudo do aprendizado de máquina é parte do campo de pesquisa da inteligência artificial (IA). Os algoritmos de aprendizado de máquina são utilizados na geração de classificadores para determinado conjunto de amostras. Classificação é o processo de atribuir o rótulo da classe ao qual a informação pertence. Assim, as técnicas de aprendizado de máquina buscam produzir classificadores capazes de inferir a classe de amostras do mesmo domínio ao qual foi treinado (RUSSELL; NORVIG, 2016).

3.4.1 Naive Bayes

Os classificadores *baysianos* preveem a probabilidade de associações entre classes, assim como a probabilidade de determinada amostra pertencer a classe específica. Eles são derivados a partir do Teorema de Bayes dado por (3.4), na qual $P(H|X)$ é probabilidade *a posteriori* de H condicionado a X, X é a amostra de dados de classe desconhecida, H é a hipótese que X pertença à classe C e $P(H)$ é a probabilidade *a priori* condicionada a H (HAN et al., 2012):

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{3.4}$$

Supondo que $P(C_k|X)$ denota a probabilidade do objeto X pertencer à classe C_i. A função custo *zero-um* que representa o custo de associar X à classe incorreta, é minimizada se e somente se, X é associado à classe C_k a qual $P(C_k|X)$ é máxima. Este método é designado por estimativa *maximum a posteriori* (MAP). A classe que deve ser associada ao objeto X é dada por (FACELI et al., 2011):

$$C_{MAP} = \arg\max_i \; P(C_i|X) \tag{3.5}$$

na qual $\arg \max_i$ retorna a classe C_i com maior probabilidade de estar associada a X, que é aquela que possui o valor máximo para $P(C_i|X)$. Em (3.4), o teorema de Bayes é utilizada como método para calcular $P(C_i|X)$, dado por:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{3.6}$$

Considerando que os valores dos atributos de determinado objeto são independentes entre si dada a classe C, $P(X|C_i)$ pode ser decomposto no produto $P(X_1|C_i) \times$

$... \times P(X_j|C_i)$, em que X_j é o j-ésimo atributo do objeto X. Assim, a probabilidade $P(C_i)$ de determinado objeto X pertencer à classe C_i é proporcional a:

$$P(C_i|X) \ \alpha \ P(C_i) \prod_{j=1}^{d} P(X_j|C_i) \tag{3.7}$$

O classificador Naive Bayes é obtido a partir da função discriminante dada por (3.7) e pela regra de decisão dada por (3.5). A expressão do classificador é dada de forma aditiva, aplicando logaritmos em (3.7), tem-se:

$$\log(P(C_i|X)) \ \alpha \ \ \log(P(C_i)) + \sum_{j} \log(P(X_j|C_i)) \tag{3.8}$$

Para o caso particular de duas classes, (3.7) é reescrita como:

$$\log \frac{P(C_1|X)}{P(C_2|X)} \ \alpha \ \ \log \frac{P(C_1)}{P(C_2)} + \sum_{j} \log \frac{P(X_j|C_1)}{P(X_j|C_2)} \tag{3.9}$$

Em (3.9), o sinal de cada termo indica a contribuição de cada atributo para cada classe. Se o quociente $P(X_j|C_1)/P(X_j|C_2)$ é maior que 1, o logaritmo é positivo e o atributo contribui para a predição da classe C_1.

3.4.2 Máquina de vetores de suporte

O classificador máquina de vetores de suporte (*Support Vector Machine* - SVM) é do tipo de aprendizado supervisionado. A classificação é realizada através da busca pelo maior vetor de suporte de separação dos hiperplanos. Nesta técnica é utilizado o *kernel* para transformar a superfície não-linear em linear (HAN et al., 2012). É realizado mapeamento não-linear para transformar os dados de treinamento em dimensão superior. Nesta nova dimensão, procura-se a separação ótima do hiperplano entre as classes. Realizando-se o mapeamento apropriado para dimensão suficientemente elevada é possível encontrar a separação entre as classes (LIMA, 2016).

3.4.3 Árvore de decisão

Na técnica árvore de decisão a classificação é realizada com a construção de fluxograma de decisão formando a estrutura de árvore. Cada nó representa determinado teste sobre o atributo, cada ramo representa o resultado do teste e as regras de clas-

sificação são os caminhos entre a raiz e a folha. O conjunto de regras de classificação é dado pela árvore de decisão criada no processo. Sendo determinado objeto desconhecido x_i, este é classificado de acordo com as regras determinadas pela árvore (HAN et al., 2012).

3.4.4 *K*-vizinhos mais próximos

A técnica K-vizinhos mais próximos (*K-Nearest Neighbor* - KNN) é descrita inicialmente nos anos 1950 e realiza a classificação baseada na vizinhança dos objetos. Os k-vizinhos mais próximos do objeto x_i determina a classe, sendo que cada objeto representa determinado ponto no espaço n-dimensional. A medida de distância calculada a partir do atributos dá a proximidade entre os objetos (HAN et al., 2012).

Para a distância euclidiana de dois objetos $x_1 = \{x_{11}, x_{12}, ..., x_{1n}\}$ e $x_2 = \{x_{21}, x_{22}, ..., x_{2n}\}$ é dada por (HAN et al., 2012):

$$d(x_1, x_2) = \sqrt{\sum_{N}^{k}(a_{1k} - a_{2k})^2} \tag{3.10}$$

na qual N é o número de atributos do objeto, que é atribuída à classe mais comum entre os vizinhos, para o objeto desconhecido x_i.

3.5 Validação cruzada *k-fold*

Em casos onde a quantidade de dados não é suficiente, dividir a base para testes pode prejudicar a análise, dado que a base para treino pode ficar pequena e acabar não representando os dados reais. Para casos como estes, tem-se a técnica validação cruzada k-fold. Neste método de validação cruzada, o *corpus* é dividido de maneira randômica em k, nos chamados *folds*. Posteriormente são executados k turnos de treinamento e validação, assim, uma das partes é escolhida para teste enquanto $k-1$ são utilizados para treinamento. A medida final para desempenho é dada pela média dos k testes (HAN et al., 2012).

3.6 Avaliação de desempenho

Esta é a etapa final no processo de mineração de dados, onde são avaliados os resultados da análise, que depende de alguns fatores como a técnica escolhida. É necessário avaliar o desempenho do classificador com a utilização de métricas, que mensura a capacidade deste em avaliar novos conjuntos de dados, garantindo que os

resultados obtidos a partir do processo, apresentem padrões verdadeiros (LIU, 2012; LIMA, 2016; BRITO, 2017).

A maioria das métricas de avaliação utilizam a **matriz de confusão**, disposta na Tabela 3.1, na qual são relacionadas as classes reais do objeto de avaliação (se positivo ou negativo) com a escolha que o classificador realiza. Para as quatro situações possíveis têm-se: i) verdadeiro positivo (VP) que são os casos classificados como positivos e que realmente são da classe de positivos, ii) falso positivo (FP) que são os casos classificados como positivos e são da classe de negativos, iii) falso negativo (FN) que são os casos classificados como negativos e que são da classe de positivos e iv) verdadeiro negativo (VN) que são os casos classificados como negativos e realmente são. Assim, VP e VN são classificações realizadas corretamente ao passo que FN e FP são classificações realizadas equivocadamente (BRITO, 2017).

Tabela 3.1 - Matriz de confusão.

		Classe Prevista	
		Positiva	**Negativa**
Classe Real	**Positiva**	VP	FN
	Negativa	FP	VN

3.6.1 Acurácia

A acurácia mede o desempenho de acertos que o classificador realiza, dada por:

$$A = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{FN} + \text{VN}} \tag{3.11}$$

Porém a acurácia sozinha não indica a eficácia do classificador (LIMA, 2016).

3.6.2 Precisão

Precisão mensura as classificações realizadas corretamente como positivas dentre todos que foram classificados como positivo, dividindo o número de acertos VP pela quantidade total classificada como positivo, dada por:

$$P = \frac{\text{VP}}{\text{VP} + \text{FP}} \tag{3.12}$$

3.6.3 Sensitividade

Ao contrário de precisão, a sensitividade (*recall*), é a relação entre a quantidade classificada corretamente como positiva, dentre todos que são realmente positivos, dada por:

$$S = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{3.13}$$

3.6.4 Média harmônica da precisão e sensitividade

Apesar da utilidade das medidas precisão e sensitividade para avaliar o desempenho dos classificadores, em várias situações faz-se necessário a utilização de medida única, para realizar a comparação de maneira direta entres dois ou mais classificadores. Neste sentido, utiliza-se a média harmônica da precisão e sensitividade (*F-Measure*), dada por:

$$F_\beta = \frac{(1 + \beta^2) \cdot P \cdot S}{\beta^2 \cdot P + S} \tag{3.14}$$

na qual a constante β determina o peso para a precisão ($0 < \beta < 1$) ou para a sensitividade ($\beta > 1$).

3.7 Otimização hiperparamétrica via busca em grade

O termo hiperparâmetro é utilizado para as variáveis de ajuste dos algoritmos de aprendizado de máquina que governam o espaço de classificação do modelo. A otimização hiperparamétrica possibilita melhorar o desempenho de predição dos modelos, porém existe custo inerente à estimativa de valores ótimos. Os desafios da obtenção destes valores ótimos estão diretamente ligados ao conjunto de dados, aos algoritmos utilizados, à função de custo e entre outros fatores.

Na busca em grade, tradicionalmente utilizada na otimização de modelos com número reduzido de hiperparâmetros, cada variável que se deseja otimizar é delimitada em intervalo definido de busca, com a quantidade e valores que se considerem adequados. Neste método de busca é gerado espaço hiperparamétrico em grade do conjunto T dos arranjos candidatos. A função $\mathbf{f}$ avalia $\lambda_{1:T}$ nos treinamentos dos modelos de aprendizado através de função de perda $\mathcal{L}$. O objetivo é encontrar o ponto de configuração ótimo entre todas as configurações testadas (PROVINCE, 2015; ALVARENGA

JÚNIOR, 2018).

3.8 Considerações finais

Os classificadores apresentados com a técnica de aprendizado de máquinas são promissores para a área de mineração de opinião pela capacidade de extrair informações relevantes em grandes conjuntos de dados de maneira automática. Com a utilização da técnica de otimização em grade é possível obter modelo otimizado através da configuração de hiperparâmetros que produzem melhor desempenho de classificação. No próximo capítulo será apresentada a metodologia proposta para desenvolver os classificadores que minerarão algumas opiniões de tuítes.

CAPÍTULO 4

METODOGIA

Neste capítulo é apresentada a metodologia utilizada para o desenvolvimento do estudo proposto. É descrito como os conjuntos de dados são preparados, quais são os estágios de preprocessamento aplicados para a geração de novos subconjuntos, quais classificadores de aprendizado de máquina são utilizados, como os resultados são avaliados, como é realizada a seleção de atributos e método de busca em grade para otimização hiperparamétrica e seleção do modelo otimizado.

4.1 Contextualização

O desenvolvimento de modelos otimizados para a mineração de opinião de tuítes é necessário devido a elevada geração de novos dados sobre os mais diversos assuntos. Além disto, a limitação de 280 caracteres para os tuítes torna esta tarefa desafiante, pois eleva o uso de linguagem informal induzindo o uso de abreviações. Mineração de opiniões em tuítes é utilizada para entender o que consumidores e público em geral pensam sobre políticos, produtos, marcas, serviços e eventos sociais. As informações obtidas através da análise, estratificação e separação destas opiniões compõe, por exemplo, processos de elaboração de campanhas de *marketing* e estratégias empresariais.

A metologia proposta neste trabalho tem o objetivo de comparar o desempenho de classificadores de aprendizado de máquina para a mineração de opinião de tuítes. São utilizados quatro estágios de preprocessamento. Para cada conjunto de dados estes estágios são aplicados gerando novos subconjuntos que são insumo de treinamento dos classificadores para que se possa avaliar o impacto do preprocessamento dos dados no desempenho do modelo.

Os estágios são: i) **Estágio 1**: dados sem preprocessamento, ii) **Estágio 2**: dados limpos com padronização dos termos, iii) **Estágio 3**: dados limpos com padronização dos termos e remoção de *stopwords* e iv) **Estágio 4**: dados limpos com padronização dos termos, remoção de *stopwords* e *stemming*. Além disto, os atributos são extraídos com o uso das técnicas: i) saco-de-palavras e ii) TF-IDF. A Figura 4.1 ilustra o fluxo da metologia proposta.

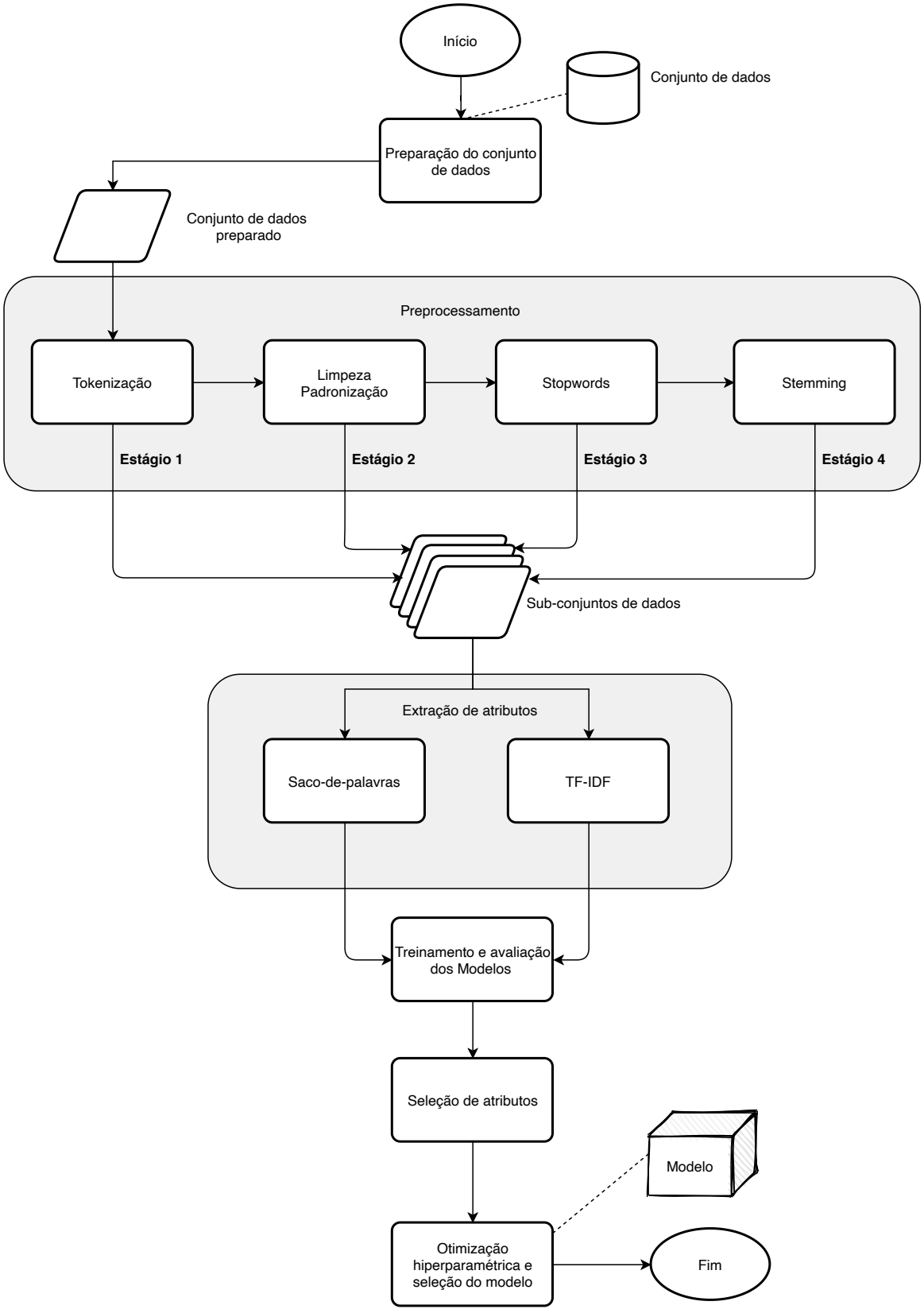

Figura 4.1 - Fluxo da metodologia proposta.

4.2 Preparação do conjunto de dados

Cada conjunto de dados é disponibilizado em formato bruto distinto. Por isto é necessária a padronização para que as tarefas sejam realizadas em todos eles sem a necessidade de personalização. São necessárias ações específicas como: remoção de colunas, alteração de cabeçalhos e padronização das anotações de sentimento. Ao fim, o conjunto deve ser armazenado em novo arquivo com formato *Comma-Separated-Values* (CSV), possuindo duas colunas, uma com o texto do tuíte e a outra com a anotação do sentimento: positivo, negativo ou neutro. A Figura 4.2 ilustra o conjunto de dados antes e depois da padronização.

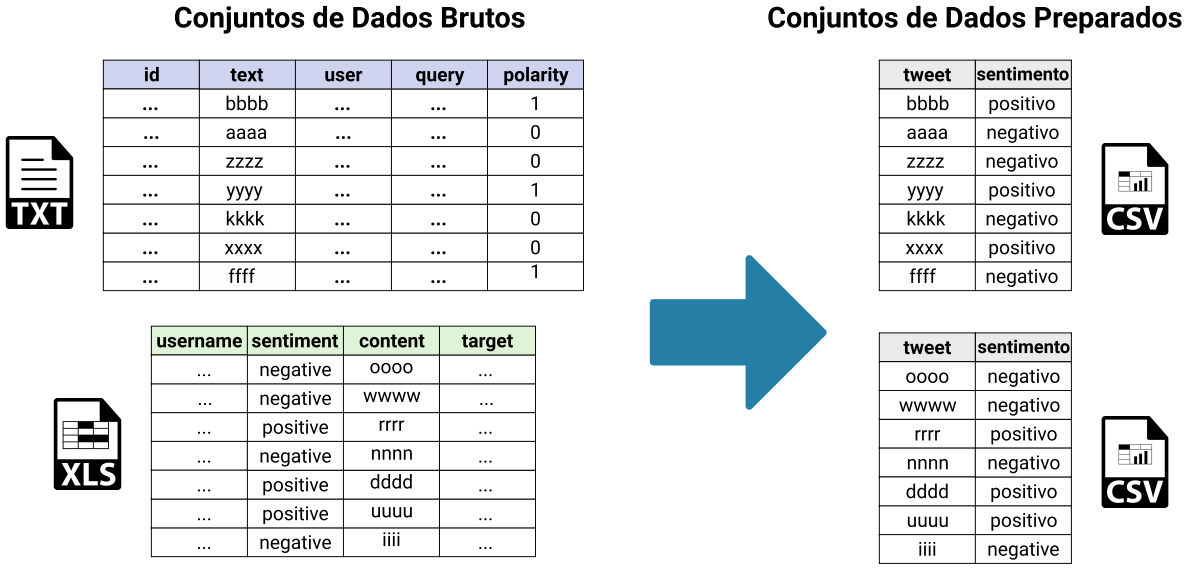

Figura 4.2 - Etapa de padronização dos conjuntos de dados.

4.3 Preprocessamento

A etapa de preprocessamento é o conjunto de modificações aplicadas aos dados antes do treinamento dos classificadores de aprendizado de máquina. Nesta etapa os dados são preparados, organizados, selecionados e estruturados para serem utilizados no treinamento do modelos de classificação de opinião. Propõe-se nesta etapa a execução dos seguintes estágios: i) **tokenização**, ii) **limpeza e padronização dos termos**, iii) **remoção de *stopwords*** e iv) ***stemming***. Na tokenização o texto de cada tuíte é dividido em lista de *tokens*, que são unidades mínimas de texto: palavras ou símbolos, como ilustrado na Figura 4.3. O critério utilizado para separação do texto são espaços em branco e a presença de caracteres especiais.

Figura 4.3 - Tokenização de dados.

Os *tokens* como acentos, vírgulas, pontuações e outros elementos não textuais como *hiperlinks* e manipulações do *Twitter* são considerados ruídos. O tratamento de ruídos também tem como benefício a redução de dimensionalidade dos dados que são utilizados no treinamento do classificador. Além disto, todos os tuítes têm seus textos convertidos para letras minúsculas. Esta ação é necessária para que os *tokens*, por exemplo: **House** e **house** sejam considerados o mesmo.

Após isto, é realizada a remoção de *stopwords*, outra etapa que impacta na redução de dimensionalidade dos dados. Existem listas de *stopwords* em diversos idiomas para auxiliar na realização desta etapa. Artigos, preposições, advérbios, conjunções e pronomes são exemplos de *stopwords* que guardam pouca ou nenhuma informação relevante sobre o contexto analisado e são úteis somente para a compreensão do texto. A etapa de *stemming* também impacta na redução de dimensionalidade, pois diminui o número de *tokens* distintos. Nesta etapa as palavras são reduzidas à sua raiz, removendo os prefixos, sufixos e eliminando os plurais e tempos verbais.

4.4 Extração dos atributos dos dados

Após realizadas as etapas de tokenização, remoção de *stopwords* e *stemming*, o próximo passo é extrair os atributos. Como o tratamento de palavras é tarefa exaustiva para os computadores, os dados precisam ser representados de maneira que possam ser processados pelos classificadores de aprendizado de máquina. São utilizadas duas técnicas para a extração de atributos: i) saco-de-palavras e ii) *term frequency-inverse document frequency* – TF-IDF.

Nas duas técnicas os *tokens* são analisados com as abordagens 1-grama, 2-grama

e 3-grama. Das três abordagem, apenas uma é escolhida para ser aplicada neste trabalho. Os fatores considerados para a escolha da abordagem são: i) redução da dimensionalidade dos dados e ii) desempenho da classificação. No saco-de-palavras o peso é dado pela frequência com que o termo aparece no documento. Na técnica TF-IDF avalia-se a importância do termo para o texto. Termos menos frequentes no *corpus* e que ao mesmo tempo têm alta frequência no documento em que aparecem recebem pesos maiores. Nas duas técnicas os dados são vetorizados e transformados em matriz esparsa com o comprimento de todo vocabulário com seu respectivo peso.

4.5 Aplicação e validação dos modelos de aprendizado de máquina

De posse do conjunto de dados organizado/preprocessado, pode-se aplicar os classificadores de aprendizado de máquina. São utilizados os classificadores: i) Naive Bayes, ii) máquina de vetor de suporte, iii) árvore de decisão e iv) K-vizinhos mais próximos. Estes classificadores foram escolhidos devido ao amplo uso na área de mineração de textos. Como as amostras de dados não são extensas para a divisão dos conjuntos para treinamento e testes, é utilizada toda base para a modelagem e avaliação de desempenho do classificador através da validação cruzada *k-fold*.

Na validação cruzada *k-fold* o conjunto de dados é divido em k conjuntos menores, k - 1 para treinamento do modelo e um para testes, como ilustrado na Figura 4.4. São realizados k turnos de treinamento e validação, sendo que as métricas de avaliação escolhidas são calculadas através da média e desvio padrão de todos os treinos realizados. A escolha da quantidade de *folds* é realizada de forma empírica, na tentativa de obter os melhores resultados. Os modelos são comparados e analisados a partir das métricas **acurácia** e **média harmônica da precisão e sensitividade**.

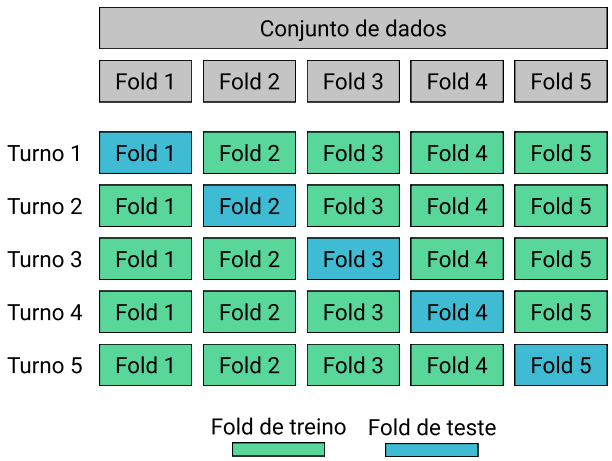

Figura 4.4 - Visualização hipotética da validação cruzada *k-folds* para $k = 5$.

4.6 Seleção de atributos com χ^2

Dado o objetivo do estudo de obter o classificador com melhor desempenho para cada conjunto de dados, opta-se por utilizar o algoritmo χ^2 para seleção dos atributos que menos contribuem para o desempenho do modelo e adicionalmente, com a redução da dimensionalidade dos dados, diminuindo o custo computacional da otimização hiperparamétrica e seleção do modelo.

São selecionados determinada porcentagem dos atributos P_{pa} provenientes do *stemming*, **Estágio 4** do preprocessamento, que recebem os maiores valores de χ^2. O valor de P_{pa} é escolhido a partir de testes empíricos na seleção dos atributos de um dos conjuntos de dados com valores de porcentagem de 50% à 100% em passos de 10%. Após obtido o P_{pa}, utiliza-se este valor nos demais conjuntos de dados pois, é proposto que todos passem pelas mesmas etapas.

4.7 Otimização hiperparamétrica e seleção do modelo com maior desempenho

A etapa final da metodologia proposta consiste em realizar a otimização hiperparamétrica dos classificadores de aprendizado de máquina, e então selecionar o que obtiver a maior acurácia. A otimização é realizada através da busca em grade da combinação das variações de configurações dos hiperparâmetros dos classificadores Naive Bayes, máquinas de vetores de suporte, K-vizinhos mais próximos e árvores de decisão.

4.8 Considerações finais

Este capítulo apresentou a metodologia para seleção de modelo otimizado para a mineração de opinião de tuítes. A partir de determinado conjunto de dados são gerados novos subconjuntos para avaliar o impacto das etapas de preprocessamento no desempenho dos quatro classificadores escolhidos para testes. De posse dos resultados é realizada seleção dos 90% melhores atributos resultantes do *stemming*, estágio do preprocessamento. Por fim, realiza-se busca em grade para obter e selecionar modelo com melhor desempenho. O próximo capítulo apresenta os resultados obtidos com a metodologia proposta.

CAPÍTULO 5

RESULTADOS

Neste capítulo são apresentados os resultados gerados a partir da metodologia proposta. É realizada mineração de opinião em cinco conjuntos de dados através das técnicas de aprendizado de máquina: Naive Bayes, máquinas de vetores de suporte, árvore de decisão e K-vizinhos mais próximos. Os resultados são avaliados pelas métricas de acurácia e média harmônica da precisão e sensitividade. Por fim, é realizada busca em grade para seleção e otimização paramétrica do modelo de aprendizado com melhor desempenho para cada um dos conjuntos de dados.

5.1 Tecnologias utilizadas

Os experimentos do trabalho são desenvolvidos com auxilio da linguagem de programação *Python*. Ela é lançada em 1991 por Guido Van Rossum, atualmente é desenvolvida em modelo comunitário e mantida pela organização sem fins lucrativos *Python Software Foundation*[1]. A escolha desta ferramenta se deu por ser linguagem de programação de código aberto e por possuir variedade de bibliotecas como *Numpy*[2] e *Pandas*[3] que auxiliam a análise, estruturação e manipulação de dados. Para o desenvolvimento dos modelos de classificação é utilizada a biblioteca *Scikit-learn*[4], descrita em Pedregosa et al. (2011). Esta biblioteca possui ferramentas que auxiliam o desenvolvimento de modelos de aprendizado de máquina.

5.2 Obtenção dos conjuntos de dados

No desenvolvimento e teste deste trabalho são utilizados os tuítes com rótulos positivos e negativos dos conjuntos de dados: i) **Sentiment140**, ii) **Stanford Sentiment** (STS), iii) **Health Care Reform** (HCR), iv) **SentiStrenght Twitter Dataset** (SS-Twitter) e v) **Sanders**. Os conjuntos STS e HCR são criados e disponibilizados por Go et al. (2009), o conjunto de dados SS-Twitter é disponibilizado pelos desenvolvedores da ferramenta *SentiStrenght* e o conjunto de dados Sanders é desenvolvido por Nike Sanders (GO et al., 2009; LIMA, 2016). A Tabela 2.1 dispõe os dados contidos nos conjuntos de dados utilizados para a tarefa de mineração de opinião.

[1] https://www.python.org/psf/
[2] https://numpy.org/
[3] https://pandas.pydata.org/
[4] https://scikit-learn.org/

5.3 Impacto da representação dos dados e da validação cruzada k-fold no desempenho dos classificadores

De posse dos conjuntos de dados, são avaliados o desempenho da classificação sem nenhum preprocessamento dos dados com os modelos NB, utilizando saco-de-palavras para extração de atributos e SVM com TF-IDF. O objetivo é avaliar o impacto da representação dos dados em 1-grama, 2-grama e 3-grama, e da validação cruzada com 10, 15 e 20-$folds$. O valor de $k = 10$ é a referência pois é comumente utilizado nas aplicações de aprendizado de máquina na literatura. Os resultados para o classificador SVM estão publicados em Brandão e Calixto (2019). Na Tabela 5.1 estão dispostos os resultados dos experimentos iniciais.

Em geral as melhores acurácias foram obtidas com a representação dos dados em 1-grama. Pode-se observar isto nos resultados para os conjuntos de dados Sentiment140, HCR, SS-Twitter e Sanders. Ao contrário, STS teve melhores acurácias com a representação em 2-grama com 73,77% para o modelo NB e 69,40% para o SVM, nas duas situações utilizando 10-*folds* para validação cruzada. Nesse conjunto teve-se maior discrepância do resultado para a representação em 3-grama com modelo SVM, sendo as menores acurácias obtidas com 63,93% para 10-$folds$ e 15-$folds$ e 65,57% para 15-$folds$. Outro ponto a destacar é que a representação em 3-grama produz as menores acurácias, com exceção ao classificador NB e 10-$folds$ para o conjunto de dados Sentiment140 que a menor acurácia foi com 1-grama e 2-grama, sendo 78,55%.

Os resultados iniciais indicam que aumentar os valores de k para validação cruzada pode produzir melhores acurácias, enquanto aumentar a quantidade de *tokens* na representação dos atributos em geral não representa melhorias no desempenho. Contudo, utilizar 20-$folds$ aumenta em demasia o custo computacional pois é necessário realizar o dobro de ciclos de treinamento e teste quando comparado à $k = 10$. Neste mesmo sentido, utilizar representação de 1-grama produz maior quantidade de atributos, o que também aumenta o custo de processamento do modelo de aprendizado de máquina. Portanto, nos testes iniciais não foram obtidos ganhos na maioria dos casos com o uso de mais *tokens* no termo e desta forma, a opção por 2-grama é realizada para obter maior custo-benefício entre redução de dimensionalidade e ganho de desempenho na classificação.

Tabela 5.1 - Acurácias obtidas com os classificadores Naive Bayes e máquinas de vetores de suporte.

Sentiment140						
	NB com saco-de-palavras			**SVM com TF-IDF**		
***k*-folds**	**1-grama**	**2-grama**	**3-grama**	**1-grama**	**2-grama**	**3-grama**
***k*=10**	78,55%	78,55%	79,11%	79,39%	78,83%	76,88%
***k*=15**	**79,67%**	78,83%	79,39%	**81,06%**	78,27%	78,55%
***k*=20**	**79,67%**	79,39%	79,10%	79,94%	79,67%	78,83%
STS						
	NB com saco-de-palavras			**SVM com TF-IDF**		
***k*-folds**	**1-grama**	**2-grama**	**3-grama**	**1-grama**	**2-grama**	**3-grama**
***k*=10**	70,49%	**73,77%**	70,49%	68,85%	**69,40%**	63,93%
***k*=15**	71,04%	72,68%	70,49%	67,76%	68,30%	65,57%
***k*=20**	69,95%	72,78%	71,58%	67,76%	67,21%	63,93%
HCR						
	NB com saco-de-palavras			**SVM com TF-IDF**		
***k*-folds**	**1-grama**	**2-grama**	**3-grama**	**1-grama**	**2-grama**	**3-grama**
***k*=10**	**70,20%**	68,78%	67,96%	75,51%	73,06%	71,63%
***k*=15**	70,00%	69,18%	67,35%	**75,92%**	72,86%	71,84%
***k*=20**	69,80%	67,76%	67,35%	73,47%	72,86%	72,45%
SS-Twitter						
	NB com saco-de-palavras			**SVM com TF-IDF**		
***k*-folds**	**1-grama**	**2-grama**	**3-grama**	**1-grama**	**2-grama**	**3-grama**
***k*=10**	72,56%	71,17%	69,33%	73,39%	72,35%	71,12%
***k*=15**	72,56%	71,69%	68,81%	73,00%	72,04%	71,04%
***k*=20**	**72,78%**	71,25%	68,68%	**73,61%**	72,13%	71,21%
Sanders						
	NB com saco-de-palavras			**SVM com TF-IDF**		
***k*-folds**	**1-grama**	**2-grama**	**3-grama**	**1-grama**	**2-grama**	**3-grama**
***k*=10**	76,26%	76,35%	74,79%	78,64%	76,35%	74,89%
***k*=15**	76,99%	75,89%	75,16%	79,65%	78,64%	76,99%
***k*=20**	**77,09%**	76,81%	76,35%	**80,48%**	78,92%	77,54%

5.4 Conjuntos de dados preparados

Nesta etapa os conjuntos de dados são preparados para que sejam manipulados em processos padronizados nas etapas seguintes. Para isto cada conjunto recebe tratamento único, pois são disponibilizados na literatura em formatos e características distintas. É disposto na Tabela 5.2 trechos do conjunto de dados HCR antes da preparação. Como resultado desta etapa obtém-se o conjunto de dados como disposto na Tabela 5.3. O mesmo processo de preparação é realizado para todos os outros conjuntos.

Tabela 5.2 - Trecho do conjunto de dados HCR antes da preparação.

tweet id	user id	username	content	sentiment	target	annotator id	comment
10250616855	21130327	KateRoseMe	Every1 deserves access 2 better healthcare like Sarah Palin got as a child!Sign Petition 4 #HCR http://bit.ly/3GsRB9 #p2	positive	hcr	aluckhardt	Also negative for Palin, as tweet implies hypocrisy on her part.
10255992832	23825051	webmiss007	RT @jodotcom: Folks, if we don't pass #hcr now, it will be 30 more yrs before we get another chance; let's don't mess it up.	positive	hcr	aluckhardt	
10272412959	25777000	kjlintner	Corey Haim is dead. I wonder how many Americans w/o Health Care will die today & go unmentioned? #p2 #tcot #hcr	positive	hcr	aluckhardt	
10290782310	41373006	sarahlee310	AP Health Care Poll: Only FOUR PERCENT Of Americans Don't Want Any Reform http://bit.ly/9rYbqR #HCR #p2	positive	hcr	aluckhardt	All caps emphasizes sentiment in this factual tweet. Could be negative for conservatives, who suggested that more than 4% didn't want reform.
10296651867	64816476	BlondeAmerican	The #teaparty protesters & Anti- #HCR protesters outnumbered the Obama supporters. I'm proud of the people of Missouri.	positive	teaparty	aluckhardt	
9946829766	17388022		RW so pissed over #hcr they think ppl will vote against Dems. Ppl would vote for ths who care abt "we the people" not "GOP the people" #p2	negative	gop	AMahmud	
10267241057	73699879	USLib	How he sleeps at night. RT @wolfblitzercnn: Karl Rove joins me Wednesday in SitRoom 5PM ET. What would you ask him?	negative	gop	athornton	Shows negative sentiment to Rove
10278968512	25941604	Contemplari	Uncovered Shocker: Sen. Byrd Single-Handedly Stopped President Clinton From Using Reconciliation http://tinyurl.com/yl252v5 MUST SEE #hcr	negative	hcr	athornton	sarcastic?, might be irrelevant since talking about Clinton
10281108398	7713202	GOPLeader	The Dem scheme to ram health care takeover through Congress now has a name: the Slaughter Solution http://bit.ly/aMxRVS #hcr	negative	dems	athornton	"ram"and "Slaughter"are negative, takeover itself could be positive
10281128016	16726846	bccohan	RT @GOPLeader: The Dem scheme to ram health care takeover through Congress now has a name: the Slaughter Solution http://bit.ly/aMxRVS #hcr	negative	dems	athornton	"ram"and "Slaughter"are negative, takeover itself could be positive, fact that its RT by a GOP shows possibly negative action

5.5 Estágios de preprocessamento das mensagens e extração de atributos

Cada conjunto de dados tem as **mensagens** dos tuítes estratificadas a partir de quatro estágios. No **Estágio 1** não recebem nenhum preprocessamento, como disposto na Tabela 5.4 para o conjunto de dados HCR. No **Estágio 2**, disposto na Tabela 5.5, são realizadas as tarefas de preprocessamento de **tokenização, limpeza e padronização dos termos**. A Tabela 5.6 dispõe as mensagens do **Estágio 3** depois da **remoção de *stopwords***. Por último, no **Estágio 4** é realizada tarefa de ***stemming***, o resultado é disposto na Tabela 5.7.

A partir dos estágios de preprocessamento são criados quatro *corpus* para cada conjunto de dados. Por sua vez, todos os *corpus* têm seus atributos extraídos utilizando as técnicas de saco-de-palavras e TF-IDF, que foram escolhidas por sua vasta utilização na literatura. Sendo assim, para cada conjunto de dados são gerados 32 modelos de classificação.

5.6 Modelagem dos classificadores de aprendizado de máquina

Os classificadores são modelados com o apoio da biblioteca *Scikit-learn*, descrita em Pedregosa et al. (2011). São mantidos os valores padrões dos parâmetros dos classificadores. Para treinamento e avaliação de desempenho utiliza-se a técnica de validação cruzada *k-fold* com $k = 10$. Este valor é escolhido de maneira empírica a partir dos testes realizados em Brandão e Calixto (2019) e disposto na Tabela 5.1, com o objetivo de obter o melhor custo-benefício entre desempenho da classificação e custo computacional.

São consideradas as métricas de acurácia e média harmônica da precisão e sensitividade (*F-Measure*) para os modelos gerados. A acurácia é a referência para determinar o melhor modelo por demonstrar quantos dos exemplos foram de fato classificados corretamente, independente da classe. Em caso de valores de acurácia iguais, a métrica *F-Measure* que leva em consideração a precisão e a revocação é utilizada. Esta métrica resume melhor a qualidade do modelo, pois valores altos representam que os modelos são capazes tanto de acertar as predições como recuperar os exemplos da classe de interesse.

Em análise geral, o classificador SVM obteve maiores resultados de acurácia e *F-Measure* para a maioria dos conjuntos de dados. Com exceção a Sentiment140, onde o melhor desempenho foi obtido pelo classificador Naive Bayes. Em contrapartida,

os experimentos com o classificador KNN teve os menores resultados absolutos para todos os conjuntos de dados.

Tabela 5.3 - Trecho do conjunto de dados HCR preparado.

sentimento	**mensagem**
positivo	Every1 deserves access 2 better healthcare like Sarah Palin got as a child!Sign Petition 4 #HCR http://bit.ly/3GsRB9 #p2
positivo	RT @jodotcom: Folks, if we don't pass #hcr now, it will be 30 more yrs before we get another chance; let's don't mess it up.
positivo	Corey Haim is dead. I wonder how many Americans w/o Health Care will die today & go unmentioned? #p2 #tcot #hcr
positivo	AP Health Care Poll: Only FOUR PERCENT Of Americans Don't Want Any Reform http://bit.ly/9rYbqR #HCR #p2
positivo	The #teaparty protesters & Anti- #HCR protesters outnumbered the Obama supporters. I'm proud of the people of Missouri.
negativo	RW so pissed over #hcr they think ppl will vote against Dems. Ppl would vote for ths who care abt "we the people" not "GOP the people" #p2
negativo	How he sleeps at night. RT @wolfblitzercnn: Karl Rove joins me Wednesday in SitRoom 5PM ET. What would you ask him?
negativo	Uncovered Shocker: Sen. Byrd Single-Handedly Stopped President Clinton From Using Reconciliation http://tinyurl.com/yl252v5 MUST SEE #hcr
negativo	The Dem scheme to ram health care takeover through Congress now has a name: the Slaughter Solution http://bit.ly/aMxRVS #hcr
negativo	RT @GOPLeader: The Dem scheme to ram health care takeover through Congress now has a name: the Slaughter Solution http://bit.ly/aMxRVS #hcr
negativo	Nancy Pelosi: We have to pass #hcr so you can find out what's in it http://is.gd/a5XoOhttp://is.gd/a5XBB #tcot #tlot #sgp #hhrs
negativo	RT @Senate_GOPs TPM: GOP To Dems: If You Think You'll Be More Popular After Health Care, Think Again http://bit.ly/bsdauM #tcot #hcr #sgp

Tabela 5.4 - Estágio 1 das mensagens do conjunto de dados HCR.

mensagem
Every1 deserves access 2 better healthcare like Sarah Palin got as a child!Sign Petition 4 #HCR http://bit.ly/3GsRB9 #p2
RT @jodotcom: Folks, if we don't pass #hcr now, it will be 30 more yrs before we get another chance; let's don't mess it up.
Corey Haim is dead.
AP Health Care Poll: Only FOUR PERCENT Of Americans Don't Want Any Reform http://bit.ly/9rYbqR #HCR #p2
The #teaparty protesters & Anti- #HCR protesters outnumbered the Obama supporters. I'm proud of the people of Missouri.
RW so pissed over #hcr they think ppl will vote against Dems. Ppl would vote for ths who care abt "we the people" not "GOP the people" #p2
How he sleeps at night. RT @wolfblitzercnn: Karl Rove joins me Wednesday in SitRoom 5PM ET. What would you ask him?
Uncovered Shocker: Sen. Byrd Single-Handedly Stopped President Clinton From Using Reconciliation http://tinyurl.com/yl252v5
The Dem scheme to ram health care takeover through Congress now has a name: the Slaughter Solution http://bit.ly/aMxRVS #hcr
RT @GOPLeader: The Dem scheme to ram health care takeover through Congress now has a name: the Slaughter Solution http://bit.ly/aMxRVS #hcr
Nancy Pelosi: We have to pass #hcr so you can find out what's in it http://is.gd/a5XoOhttp://is.gd/a5XBB #tcot #tlot #sgp #hhrs
RT @Senate_GOPs TPM: GOP To Dems: If You Think You'll Be More Popular After Health Care, Think Again http://bit.ly/bsdauM #tcot #hcr #sgp

Tabela 5.5 - Estágio 2 das mensagens do conjunto de dados HCR.

mensagem
deserves access better healthcare like sarah palin got as a child sign petition hcr http
rt jodotcom folks if we do pass hcr now it will be more yrs before we get another chance let do mess it up
corey haim is dead i wonder how many americans health care will die today go unmentioned tcot hcr
ap health care poll only four percent of americans do want any reform http hcr
the teaparty protesters hcr protesters outnumbered the obama supporters i proud of the people of missouri
rw so pissed over hcr they think ppl will vote against dems ppl would vote for ths who care abt we the people not gop the people
how he sleeps at night rt wolfblitzercnn karl rove joins me wednesday in sitroom et what would you ask him
uncovered shocker byrd stopped president clinton from using reconciliation http must see hcr
the dem scheme to ram health care takeover through congress now has a name the slaughter solution http hcr
rt gopleader the dem scheme to ram health care takeover through congress now has a name the slaughter solution http hcr
nancy pelosi we have to pass hcr so you can find out what in it http tcot tlot sgp hhrs
rt tpm gop to dems if you think you be more popular after health care think again http tcot hcr sgp

Tabela 5.6 - Estágio 3 das mensagens do conjunto de dados HCR.

mensagem
deserves access better healthcare like sarah palin got child sign petition hcr http
rt jodotcom folks pass hcr yrs get another chance let mess
corey haim dead wonder many americans health care die today go unmentioned tcot hcr
ap health care poll four percent americans want reform http hcr
teaparty protesters hcr protesters outnumbered obama supporters proud people missouri
rw pissed hcr think ppl vote dems ppl would vote ths care abt people gop people
sleeps night rt wolfblitzercnn karl rove joins wednesday sitroom et would ask
uncovered shocker byrd stopped president clinton using reconciliation http must see hcr
dem scheme ram health care takeover congress name slaughter solution http hcr
rt gopleader dem scheme ram health care takeover congress name slaughter solution http hcr
nancy pelosi pass hcr find http tcot tlot sgp hhrs
rt tpm gop dems think popular health care think http tcot hcr sgp

Tabela 5.7 - Estágio 4 das mensagens do conjunto de dados HCR.

mensagem
deserv access better healthcar like sarah palin got child sign petit hcr http
rt jodotcom folk pass hcr yr befor get anoth chanc let mess
corey haim dead wonder mani american health care die today go unment tcot hcr
ap health care poll onli four percent american want ani reform http hcr
teaparti protest hcr protest outnumb obama support proud peopl missouri
rw piss hcr think ppl vote dem ppl would vote th care abt peopl gop peopl
sleep night rt wolfblitzercnn karl rove join wednesday sitroom et would ask
uncov shocker byrd stop presid clinton use reconcili http must see hcr
dem scheme ram health care takeov congress ha name slaughter solut http hcr
rt goplead dem scheme ram health care takeov congress ha name slaughter solut http hcr
nanci pelosi pass hcr find http tcot tlot sgp hhr
rt tpm gop dem think popular health care think http tcot hcr sgp

5.7 Resultados para Sentiment140

Ao contrário que em outros conjuntos de dados, para Sentiment140 o maior resultado de acurácia é obtido com classificador Naive Bayes usando saco-de-palavras e **Estágio 1** de preprocessamento (82,45%), como disposto na Tabela 5.8. Este conjunto de dados tem a maior média de acurácias entre todos, com 73,44%. Novamente o classificador KNN produz modelo com menor acurácia, sendo 59,33% com saco-de-palavras e **Estágio 2** de preprocessamento.

Tabela 5.8 - Resultados para Sentiment140.

		Naive Bayes		**SVM**		**KNN**		**Árvore de Decisão**	
		Acurácia	*F-Measure*	*Acurácia*	*F-Measure*	*Acurácia*	*F-Measure*	*Acurácia*	*F-Measure*
Estágio 1	*Saco-de-palavras*	**82,45%**	82,45%	78,27%	78,21%	61,28%	59,93%	72,14%	72,03%
	TF-IDF	80,78%	80,64%	81,89%	81,87%	76,60%	76,35%	63,51%	63,32%
Estágio 2	*Saco-de-palavras*	78,27%	78,20%	73,26%	73,06%	59,33%	53,08%	71,03%	70,39%
	TF-IDF	77,99%	77,82%	77,16%	77,09%	76,32%	76,22%	64,35%	63,78%
Estágio 3	*Saco-de-palavras*	78,55%	78,46%	74,93%	74,60%	61,56%	56,06%	70,75%	69,73%
	TF-IDF	77,72%	77,45%	77,72%	77,59%	75,21%	75,13%	69,36%	68,99%
Estágio 4	*Saco-de-palavras*	78,83%	78,76%	74,09%	73,86%	61,56%	55,01%	67,97%	67,38%
	TF-IDF	78,83%	78,67%	79,39%	79,33%	77,44%	77,34%	71,59%	70,95%

5.8 Resultados para STS

Os modelos para STS tem menor média de acurácia entre todos conjuntos de dados, com 66,77%. A Tabela 5.9 dispõe os resultados dos modelos gerados e pode-se observar que dois deles tiveram acurácia de 57,92%, a menor para STS. Sendo os modelos utilizando TF-IDF e com **Estágio 1** e **Estágio 2** de preprocessamento. O modelo com **Estágio 2** é considerado o pior porque tem menor valor de *F-Measure*, com 54,53%. A maior acurácia é obtida com classificador SVM utilizando saco-de-palavras para extração de atributos e **Estágio 1** de preprocessamento.

Tabela 5.9 - Resultados para STS.

		Naive Bayes		**SVM**		**KNN**		**Árvore de Decisão**	
		Acurácia	*F-Measure*	*Acurácia*	*F-Measure*	*Acurácia*	*F-Measure*	*Acurácia*	*F-Measure*
Estágio 1	*Saco-de-palavras*	72,13%	72,10%	**76,50%**	74,77%	60,11%	44,07%	62,84%	57,85%
	TF-IDF	60,11%	41,08%	73,77%	69,42%	71,04%	68,05%	57,92%	55,58%
Estágio 2	*Saco-de-palavras*	71,58%	71,48%	72,68%	69,96%	60,66%	41,35%	67,21%	64,21%
	TF-IDF	63,39%	48,67%	69,95%	63,27%	68,85%	66,34%	57,92%	54,53%
Estágio 3	*Saco-de-palavras*	69,95%	69,82%	71,58%	65,85%	58,47%	36,90%	66,12%	61,23%
	TF-IDF	65,57%	54,59%	67,21%	59,70%	66,67%	64,99%	66,67%	60,56%
Estágio 4	*Saco-de-palavras*	68,31%	68,03%	74,32%	70,20%	59,56%	38,56%	68,31%	62,31%
	TF-IDF	65,57%	54,59%	67,21%	60,16%	69,40%	67,94%	65,03%	60,33%

5.9 Resultados para HCR

Na Tabela 5.10 estão dispostos os resultados obtidos para o conjunto de dados HCR. Observa-se que a maior acurácia é de 73,47% com o classificador SVM utilizando TF-IDF com o **Estágio 2** de preprocessamento. Para esta mesma configuração o modelo Naive Bayes produziu acurácia de 71,43%, sendo a melhor para este classificador. A média de acurácias deste conjunto de dados é 68,18%. O menor resultado de acurácia é de 55,92%, obtido com modelo do classificador KNN utilizando saco-de-palavras e **Estágio 3** de preprocessamento.

Tabela 5.10 - Resultados para HCR.

		Naive Bayes		**SVM**		**KNN**		**Árvore de decisão**	
		Acurácia	*F-Measure*	*Acurácia*	*F-Measure*	*Acurácia*	*F-Measure*	*Acurácia*	*F-Measure*
Estágio 1	*Saco-de-palavras*	70,61%	70,35%	73,06%	72,21%	63,06%	58,71%	66,53%	65,66%
	TF-IDF	70,20%	64,25%	72,45%	69,47%	69,18%	67,52%	64,08%	62,83%
Estágio 2	*Saco-de-palavras*	68,98%	68,83%	70,00%	69,07%	62,65%	59,87%	65,92%	64,91%
	TF-IDF	71,43%	66,86%	**73,47%**	70,64%	70,00%	69,02%	62,24%	61,24%
Estágio 3	*Saco-de-palavras*	70,41%	70,35%	70,61%	69,89%	55,92%	54,08%	67,14%	65,62%
	TF-IDF	70,41%	66,62%	72,65%	70,37%	68,98%	68,09%	65,92%	64,64%
Estágio 4	*Saco-de-palavras*	71,22%	71,12%	70,61%	70,04%	57,14%	56,03%	68,57%	67,03%
	TF-IDF	70,41%	66,62%	71,43%	69,04%	70,20%	69,12%	66,33%	64,83%

5.10 Resultados para SS-Twitter

Novamente os modelos desenvolvidos a partir do classificador KNN gerou o menor resultado de acurácia com 51,99%, como disposto na Tabela 5.11. Neste modelo é utilizado saco-de-palavras como método de extração de atributos e **Estágio 1** de preprocessamento. Neste conjunto de dados a maior acurácia é obtida com o modelo elaborado a partir do classificador SVM com TF-IDF e **Estágio 1** de preprocessamento, sendo de 75,27%.

Tabela 5.11 - Resultados para SS-Twitter.

		Naive Bayes		**SVM**		**KNN**		**Árvore de Decisão**	
		Acurácia	*F-Measure*	*Acurácia*	*F-Measure*	*Acurácia*	*F-Measure*	*Acurácia*	*F-Measure*
Estágio 1	*Saco-de-palavras*	74,62%	74,05%	**75,32%**	74,07%	54,26%	54,25%	65,09%	63,91%
	TF-IDF	62,43%	46,93%	75,27%	73,32%	66,67%	64,21%	61,12%	60,02%
Estágio 2	*Saco-de-palavras*	72,48%	71,80%	72,74%	71,24%	51,99%	51,09%	64,70%	63,48%
	TF-IDF	63,17%	48,76%	72,26%	69,76%	65,66%	62,90%	61,86%	60,47%
Estágio 3	*Saco-de-palavras*	71,56%	71,20%	72,87%	71,25%	48,01%	47,03%	66,36%	64,95%
	TF-IDF	65,14%	53,28%	72,35%	69,67%	67,67%	66,96%	65,88%	65,14%
Estágio 4	*Saco-de-palavras*	72,30%	71,87%	71,87%	70,15%	51,94%	51,77%	67,23%	65,06%
	TF-IDF	66,27%	55,63%	71,82%	69,25%	67,89%	65,81%	64,53%	62,70%

5.11 Resultados para Sanders

O modelos gerados para o conjunto de dados Sanders, dispostos na Tabela 5.12, obtiveram 71,38% de acurácia média. Os classificadores Naive Bayes e SVM apresentaram todos resultados de acurácia maior que a média geral. O modelo gerado com classificador KNN utilizando saco-de-palavras e **Estágio 2** de preprocessamento gerou o menor valor de acurácia, sendo 52,52%. Apesar da média de acurácias (71,11%) do classificador Naive Bayes ser maior que a média obtida pelos modelos SVM, este último apresentou o maior resultado com 78,83% utilizando TF-IDF e **Estágio 1** de preprocessamento.

Tabela 5.12 - Resultados para Sanders.

		Naive Bayes		**SVM**		**KNN**		**Árvore de Decisão**	
		Acurácia	*F-Measure*	*Acurácia*	*F-Measure*	*Acurácia*	*F-Measure*	*Acurácia*	*F-Measure*
Estágio 1	*Saco-de-palavras*	76,63%	76,61%	76,81%	76,65%	57,84%	57,83%	68,10%	67,96%
	TF-IDF	75,89%	75,19%	77,54%	77,17%	71,86%	71,19%	66,64%	66,52%
Estágio 2	*Saco-de-palavras*	**77,82%**	77,80%	75,25%	75,22%	53,53%	46,64%	67,28%	67,11%
	TF-IDF	75,99%	75,48%	76,44%	76,08%	71,59%	71,37%	65,44%	65,24%
Estágio 3	*Saco-de-palavras*	76,17%	76,16%	75,16%	75,13%	58,57%	55,49%	68,19%	68,11%
	TF-IDF	77,36%	77,24%	76,08%	75,64%	71,13%	70,92%	66,18%	66,03%
Estágio 4	*Saco-de-palavras*	77,73%	77,72%	76,99%	76,99%	58,11%	54,97%	69,29%	68,91%
	TF-IDF	77,73%	77,61%	77,91%	77,67%	72,87%	72,70%	66,45%	66,30%

5.12 Impacto da seleção de atributos com χ^2 na dimensionalidade dos dados

Como última etapa de preparação dos atributos dos conjuntos de dados antes da seleção e otimização de hiperparâmetro dos modelos de aprendizado de máquina, realiza-se a seleção dos atributos através do algoritmo χ^2. São selecionados $P_{pa} = 90\%$ dos atributos de cada conjunto de dados com maiores valores de χ^2, pois com $P_{pa} = 90\%$ obteve-se ganho no resultado da acurácia em Sentiment140 e este valor foi utilizado em todos os outros conjuntos de dados, pois foi proposto que todos passem pelas mesmas etapas. A Tabela 5.13 dispõe a redução da dimensionalidade da matriz de atributos para cada conjunto de dados após os estágios de preprocessamento e seleção dos atributos. A quantidade de tuítes que cada conjunto possui representa o número de linhas da matriz, enquanto os atributos representam as colunas da matriz.

Esta etapa é importante porque a quantidade de atributos impacta no desempenho computacional para seleção e otimização hiperparamétrica dos modelos de classificação. Por exemplo, o conjunto de dados SS-Twitter em que os dados de treinamento

do classificador possuem dimensão de 2.289 × 38.727 após a extração dos dados sem nenhum preprocessamento, contra 2.289 × 21.474 após a seleção dos atributos.

Tabela 5.13 - Demonstrativo da redução de dimensionalidade da matriz de atributos com os estágios de preprocessamento de texto e seleção de atributos.

Conjunto de Dados	**Tuítes**	**Atributos**				
		Estágio 1	**Estágio 2**	**Estágio 3**	**Estágio 4**	χ^2
Sentiment140	359	6594	5437	3908	3859	3472
STS	183	3278	2724	1924	1902	1711
HCR	490	9906	7757	5993	5823	5240
SS-Twitter	2289	38727	31617	24613	23861	21474
Sanders	1091	16810	13495	10290	9892	8902

5.13 Resultados da otimização dos hiperparâmetros e seleção de modelo

A Tabela 5.14 dispõe os hiperparâmetros que são considerados para a otimização e os valores utilizados para avaliar o modelo ótimo dentro do espaço de busca definido. A importância da redução de dimensionalidade se torna mais evidente ao observar que serão testadas 114 variações na busca em grade, o que resulta em 1140 ciclos de treinamento e testes. É utilizada, assim como no treinamento isolado dos classificadores, validação cruzada com k=10.

Tabela 5.14 - Grade de hiperparâmetros para busca de modelo otimizado.

Classificador	**Grade de hiperparâmetros**			
	Hiperparâmetro	**Padrão**	**Testados**	**Quantidade**
MultinomialNB	alpha	1.0	[1.0, 1e-1, 1e-2, ..., 1e-10]	11
	fit_prior	[True]	[True, False]	2
	Modelos candidatos			**22**
SVM	C	1.0	[1, 10, 100, 1000]	4
	gamma	1.0	[1, 0.1, 0.001, 0.0001]	4
	kernel	'linear'	'linear', 'rbf'	2
	Modelos candidatos			**32**
KNeighborsClassifier	n_neighbors	7	[3, 7, 11]	3
	weights	'uniform'	'uniform', 'distance'	2
	algorithm	'auto'	'auto', 'ball_tree','kd_tree','brute'	4
	Modelos candidatos			**24**
DecisionTreeClassifier	criterion	entropy'	'gini', 'entropy'	2
	max_depth	None	2,4,6,8,10,12	6
	max_features	None	'auto', 'sqrt', 'log2'	3
	Modelos candidatos			**36**
TOTAL DE MODELOS CANDIDATOS				**114**

A otimização e seleção de modelo é realizada para todos os conjuntos de dados, utilizando para extração de atributos as técnicas de saco-de-palavras e TF-IDF. Tal qual realizado no desenvolvimento dos modelos de teste inicial foi utilizada repre-

sentação dos dados em 2-grama. A Tabela 5.15 dispõe o resultado para otimização dos modelos para cada conjunto de dados utilizando saco-de-palavras. Em todas as situações o modelo Naive Bayes obteve a maior acurácia entre todos os modelos testados, resultando em variações únicas dos parâmetros de ajuste para os conjuntos de dados.

Para todos conjuntos de dados os resultados obtidos com a otimização superaram os valores de referência. Para Sentiment140 a melhor acurácia havia sido 82,45% com classificador Naive Bayes utilizando saco-de-palavras para extração de atributos e dados sem preprocessamento, enquanto modelo otimizado com saco-de-palavras obteve 88,02% e com TF-IDF 84,70%. Os modelos otimizados para STS obtiveram 83,06% e 84,70% com saco-de-palavras e TF-IDF respectivamente. Para esse conjunto de dados a melhor acurácia sem otimização é de 76,50% com classificador SVM e saco-de-palavras com os dados sem preprocessamento.

Tabela 5.15 - Modelos e hiperparâmetros selecionados com extração de atributos utilizando saco-de-palavras.

Conjunto de Dados	**Classificador**	**Hiperparâmetros**	**Acurácia**
Sentiment140	MultinomialNB	alpha=1e-3, fit_prior=True	88,02%
STS	MultinomialNB	alpha=1e-2, fit_prior=True	83,06%
HCR	MultinomialNB	alpha=1e-4, fit_prior=False	86,94%
SS-Twitter	MultinomialNB	alpha=1e-3, fit_prior=True	83,57%
Sanders	MultinomialNB	alpha=1e-9, fit_prior=False	89,64%

Para HCR utilizando saco-de-palavras obteve-se acurácia de 86,95% contra 73,47% obtido com classificador SVM utilizando TF-IDF para extração dos atributos e com **Estágio 2** de preprocessamento. Em SS-Twitter o modelo otimizado resultou em 83,57% de acurácia enquanto o maior resultado dos valores de referência foi 75,32% com classificador SVM, **Estágio 1** de preprocessamento e saco-de-palavras para extração dos atributos. Para o conjunto de dados Sanders a melhor acurácia de referência foi de 77,82% com classificador Naive Bayes e saco-de-palavras, em contraste o modelo otimizado obteve 89,94% de acurácia.

A Tabela 5.16 dispõe o resultado da otimização dos modelos de classificação de aprendizado de máquina utilizando TF-IDF para extração de atributos para os conjuntos de dados utilizados no estudo. A extração de atributos com TF-IDF superou a técnica saco-de-palavras na maioria das situação. A exceção foi para o conjunto de dados HCR que obteve 86,12% de acurácia enquanto o modelo com saco-de-palavras

resultou em 86,94%.

Tabela 5.16 - Modelos e hiperparâmetros selecionados com extração de atributos utilizando TF-IDF.

Conjunto de Dados	Classificador	Hiperparâmetros	Acurácia
Sentiment140	MultinomialNB	alpha=1e-10, fit_prior=False	89,97%
STS	MultinomialNB	alpha=1e-9, fit_prior=True	84,70%
HCR	MultinomialNB	alpha=1e-7, fit_prior=False	86,12%
SS-Twitter	MultinomialNB	alpha=1e-1, fit_prior=True	84,71%
Sanders	MultinomialNB	alpha=1e-10, fit_prior=True	90,01%

Diversos estudos desenvolvem modelos de aprendizado de máquina para classificar opiniões em tuítes utilizando os mesmos conjuntos de dados deste estudo. Os trabalhos de Saif et al. (2013), Saif et al. (2014), Lima et al. (2015), Deshmukh e Pawar (2015), Lima (2016) e Ahuja et al. (2019) estão dispostos na Tabela 5.17 e são alguns exemplos. Dos estudos relacionados apenas Saif et al. (2013) utilizou validação cruzada para treinamento do modelo.

Lima (2016) utiliza o conjunto de dados de Sentiment140 e a melhor acurácia alcançada para este conjunto é 77,36% utilizando árvore de decisão J48, algoritmo presente no *software* Weka. Entre os estudos que desenvolveram modelo para conjunto de dados STS Saif et al. (2014) se destaca por obter 81,06% de acurácia utilizando classificador Naive Bayes. O estudo de Saif et al. (2013) alcança acurácia de 78,68% com o algoritmo máxima entropia, sendo a maior entre os trabalhos que desenvolveram modelo para HCR. Para SS-Twitter a maior acurácia é obtida por Lima et al. (2015) utilizando classificador Naive Bayes. E por último, Deshmukh e Pawar (2015) obtêm acurácia de 88,65% utilizando SVM e floresta aleatória para o conjunto SS-Twitter.

Tabela 5.17 - Resultados de acurácia de outros estudos similares.

Estudo	Classificador	Atributos	Conjunto de Dados				
			Sentiment140	STS	HCR	SS-Twiiter	Sanders
Lima (2016)	AdaBoost	Saco-de-palavras					69,92%
	Bagging	LIWC				71,91%	67,16%
	J48	LIWC	77,36%			65,54%	67,39%
Saif et al. (2013)	MaxEnt	Saco-de-palavras		80,17%	78,68%	73,40%	83,84%
Deshmukh e Pawar (2015)	SVM e Random Forest	Saco-de-palavras					88,65%
Ahuja et al. (2019)	Logistic Regression	TF-IDF				57,00%	
Lima et al. (2015)	SVM	TF-IDF					64,25%
	Naive Bayes	TF-IDF				77,10%	
Saif et al. (2014)	MaxEnt	SS-Pattern		77,82%	77,02%	72,84%	83,62%
	Naive Bayes	SS-Pattern		81,06%	74,27%	72,36%	83,62%
Trabalho proposto	Naive Bayes	Saco-de-palavras	88,02%	83,06%	86,94%	83,57%	89,64%
	Naive Bayes	TF-IDF	89,97%	84,70%	86,12%	84,71%	90,01%

5.14 Discussão

A partir da análise inicial dos conjuntos de dados selecionados para o estudo foi definido que seria utilizada a técnica k-fold validação cruzada para treinamento e teste dos modelos. A opção por esta técnica se deu devido aos conjuntos Sentiment140, STS e HCR terem quantidade reduzida de tuítes e consequentemente menos atributos a serem extraídos. Dividir o conjunto em partes de treinamento e teste poderia causar o sobreajuste dos modelos de classificação de opiniões. Os primeiros modelos foram desenvolvidos para avaliar o ganho de desempenho de acurácia para diferentes valores de k, sendo 10-$folds$, 15-$folds$ e 20-$folds$. Como não houveram melhorias significativas no desempenho da classificação optou-se por utilizar $k = 10$, como comumente é utilizado na literatura.

Os resultados com dados sem nenhum preprocessamento e abordagem 1-grama apresentam-se competitivos do ponto de vista de desempenho de classificação dos modelos. Contudo, isto exige demasiado custo computacional para modelagem de classificadores devido o aumento na dimensionalidade dos atributos. Por isto, a utilização das técnicas de preprocessamento são essenciais para preparar os dados para modelagem dos classificadores.

A utilização de técnica de seleção de atributos com algoritmo χ^2 permite ganhos de desempenho da acurácia de todos classificadores utilizados no estudo. Porém, não é avaliado o ganho de desempenho dos modelos com a seleção de quantidade diferente de atributos à 90% para os conjuntos STS, HCR, SS-Twitter e Sanders pois o valor foi definido a partir de testes em Sentiment140. Quantidades menores de atributos selecionados com χ^2 pode ter impacto positivo de desempenho dos modelos para os conjuntos SS-Twitter e Sanders pois, são conjuntos de dados extensos.

Para a otimização dos hiperparâmetros e seleção do classificador com melhor desempenho, a opção por 10-$folds$ para validação cruzada torna o processo lento para o conjunto SS-Twitter. Pois este, mesmo depois da seleção, possui elevada quantidade de atributos e nesta etapa são necessários 1140 ciclos de treinamento e teste para a obtenção do modelo otimizado. O método proposto para otimização hiperparamétrica produz resultados desejados. Porém, a busca do modelo ótimo via busca em grade exige que sejam realizados testes para todas as possibilidades do espaço definido. Esta situação faz que o aumento da quantidade de valores para as possíveis configurações dos hiperparâmetros eleve também o custo computacional.

Por fim, a ferramenta desenvolvida através do método proposto pode ser empregada

para a mineração de opinião de dados provenientes de outras mídias sociais além do Twitter como em resenhas de filmes, avaliações de produtos de *marketplaces*[5], resenhas de hotéis em sites de avaliação, dados de pesquisas de satisfação de clientes e entre outros.

[5] *E-commerce*, mediado por empresa, em que vários lojistas se inscrevem e vendem seus produtos.

CAPÍTULO 6

CONCLUSÃO

Este trabalho propôs método para construção de modelos de aprendizado de máquina otimizados para classificar opiniões de tuítes. Cinco conjuntos de dados foram extraídos, preparados e preprocessados para o treinamento dos classificadores Naive Bayes, máquinas de vetores de suporte, K-vizinhos mais próximos e árvores de decisão. Os resultados para estes classificadores demonstraram que em vários casos o treinamento de modelos para classificação com tuítes sem preprocessamento de texto obteve maiores acurácias. Isso pode ser devido a limitação de caracteres das mensagens dos tuítes que produz pouca informação relevante para a extração dos atributos que são utilizados no treinamento dos modelos.

O treinamento utilizando dados sem preprocessamento pode ser útil para avaliar e comparar o desempenho de diversos algoritmos em aplicações práticas e então filtrar para a etapa de seleção e otimização hiperparamétrica os que obtiverem melhor desempenho para o contexto ao qual será aplicado. Apesar das técnicas de aprendizado supervisionados apresentarem dependência de dados rotulados para treinamento, o método proposto produz desempenhos semelhantes para conjuntos de dados de tamanhos variados, o que possibilita o desenvolvimento de modelos de classificação otimizados com quantidade reduzida de dados rotulados.

Esta característica do estudo permite o desenvolvimento de ferramentas de classificação de polaridade em opiniões de mídias sociais em tempo e possivelmente custos reduzidos. Contudo, é preciso avaliar a capacidade de generalização do modelo elaborado. Por fim, a hipótese elaborada no início do projeto foi satisfeita pois, de posse dos conjuntos de dados, foi realizada análise exploratória do desempenho de classificadores e então foi possível desenvolver modelo otimizado para classificar opinião de usuários de mídias sociais.

6.1 Contribuições do Trabalho

As contribuições do trabalho podem ser assim descritas:

- Comprovação que os classificadores K-vizinhos mais próximos e árvores de decisão são pouco competitivos para a classificação de polaridade de opiniões em tuítes;
- Os resultados obtidos na etapa de treinamento e avaliação dos modelos

podem ser utilizados como comparativo para seleção dos algoritmos de aprendizado de máquina;

- A inclusão de etapas de preprocessamento geralmente não implica a melhoria de acurácia em conjuntos de dados de tuítes. Isto torna possível avaliar o desempenho geral de classificadores com dados sem processamento para que sejam utilizados na etapa de otimização e seleção de modelo somente os algoritmos quem obtiveram melhores acurácias nos testes iniciais;

- Desenvolvimento de método para construção de modelo de mineração de opiniões otimizado com aprendizado supervisionado de máquina utilizando para treinamento quantidade reduzida de dados.

Artigos em congresso:

- BRANDÃO, J. G.; CALIXTO, W. P. **N-gram and tf-idf for feature extraction on opinion mining of tweets with svm classifier**. In: 2019 International Artificial Intelligence and Data Processing Symposium (IDAP). [S.l.: s.n.], 2019. p. 1-5. 48

6.2 Sugestões para trabalhos futuros

- Incluir novos algoritmos de aprendizado supervisionado de máquina para análise de desempenho e posterior otimização hiperparamétrica e seleção de modelo;

- Utilizar outros métodos de otimização hiperparamétrica como Busca Randomizada e Algoritmos Evolutivos;

- Utilizar métodos diferentes de seleção de atributos para reduzir dimensionalidade de conjuntos de dados extensos e melhorar acurácia do modelo;

- Avaliar desempenho do método proposto para conjuntos de dados em língua portuguesa;

- Analisar custos computacionais para o treinamento dos modelos de classificação;

- Analisar a capacidade de generalização dos modelos construído a partir do métodos proposto.

REFERÊNCIAS BIBLIOGRÁFICAS

AHUJA, R.; CHUG, A.; KOHLI, S.; GUPTA, S.; AHUJA, P. The impact of features extraction on the sentiment analysis. **Procedia Computer Science**, Elsevier, v. 152, p. 341–348, 2019. 21, 59

AIZAWA, A. An information–theoretic perspective of tf–idf measures. **Information Processing & Management**, Information Processing & Management, v. 39, n. 1, p. 45–65, 2003. ISSN 0306–4573. 32

ALVARENGA JÚNIOR, W. J. **Métodos de otimização hiperparamétrica: um estudo comparativo utilizando árvores de decisão e florestas aleatórias na classificação binária**. Dissertação (Mestrado) — Universidade Federal de Minas Gerais, 2018. 38, 39

ARAÚJO, M.; GONÇALVES, P.; BENEVENUTO, F.; CHA, M. Métodos para análise de sentimentos no twitter. In: **Proceedings of the 19th Brazilian symposium on Multimedia and the Web (WebMedia'13)**. [S.l.: s.n.], 2013. 27

BARBOSA, L.; FENG, J. Robust sentiment detection on twitter from biased and noisy data. In: **COLING**. [S.l.: s.n.], 2010. 21

BARNARD, G. A. Introduction to pearson (1900) on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In: **Breakthroughs in statistics**. [S.l.]: Springer, 1992. p. 1–10. 33

BECKER, L.; ERHART, G.; SKIBA, D.; MATULA, V. Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. In: **Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)**. [S.l.: s.n.], 2013. v. 2, p. 333–340. 27

BOLLEN, J.; MAO, H.; ZENG, X. Twitter mood predicts the stock market. **Journal of computational science**, Elsevier, v. 2, n. 1, p. 1–8, 2011. 27

BORDIN JUNIOR, A. **Aplicação de programação genética na análise de sentimentos**. Dissertação (Mestrado) — Universidade Federal de Goiás, 2018. 27, 29

BRANDÃO, J. G.; CALIXTO, W. P. N-gram and tf-idf for feature extraction on opinion mining of tweets with svm classifier. p. 1–5, 2019. 48, 51

BRITO, E. M. N. D. Mineração de textos: detecção automática de sentimentos em co-mentários nas mídias sociais. **Projetos e Dissertações em Sistemas de Informação e Gestão do Conhecimento**, v. 6, n. 1, 2017. 28, 32, 37

BUGEJA, R. **Twitter Sentiment Analysis for Marketing Research**. Tese (Doutorado) — Tesis, University of Malta), 2014. 27

CAMILO, C. O.; SILVA, J. C. d. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, p. 1–29, 2009. 31

CHEN, B.; ZHU, L.; KIFER, D.; LEE, D. What is an opinion about? exploring political standpoints using opinion scoring model. In: CITESEER. [S.l.], 2010. 27

DAVE, K.; LAWRENCE, S.; PENNOCK, D. M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: **Proceedings of the 12th International Conference on World Wide Web**. [S.l.]: ACM, 2003. 19

DESHMUKH, R.; PAWAR, K. Twitter sentiment classification on sanders data using hybrid approach. **IOSR Journal of Computer Engineering (IOSR-JCE)**, v. 17, p. 118–123, 07 2015. 59

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. d. L. et al. Inteligência artificial: Uma abordagem de aprendizado de máquina. 2011. 34

FELDMAN, R.; ROSENFELD, B.; BAR-HAIM, R.; FRESKO, M. The stock sonar-sentiment analysis of stocks based on a hybrid approach. In: **Twenty-Third IAAI Conference**. [S.l.: s.n.], 2011. 27

GO, A.; BHAYANI, R.; HUANG, L. Twitter sentiment classification using distant supervision. **CS224N project report, Stanford**, v. 1, n. 12, p. 2009, 2009. 29, 30, 47

HAN, J.; KAMBER, M.; PEI, J. **Data mining concepts and techniques, third edition**. Waltham, Mass.: Morgan Kaufmann Publishers, 2012. 34, 35, 36

HEMMATIAN, F.; SOHRABI, M. K. A survey on classification techniques for opinion mining and sentiment analysis. **Artificial Intelligence Review**, Springer, p. 1–51, 2017. 19, 22, 25

HONG, Y.; SKIENA, S. The wisdom of bookies? sentiment analysis vs. the nfl point spread. In: **Fourth International AAAI Conference on Weblogs and Social Media**. [S.l.: s.n.], 2010. 27

HUAN LIU; SETIONO, R. Chi2: feature selection and discretization of numeric attributes. In: **Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence**. [S.l.: s.n.], 1995. p. 388–391. 33

JOACHIMS, T. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In: . [S.l.: s.n.], 1997. p. 143–151. 32

LIMA, A. C. E.; CASTRO, L. N. de; CORCHADO, J. M. A polarity analysis framework for twitter messages. **Applied Mathematics and Computation**, Elsevier, v. 270, p. 756–767, 2015. 59

LIMA, A. C. E. S. **Mineração de mídias sociais como ferramenta para a análise de tríade da persona virtual**. Tese (Doutorado) — Universidade Presbiteriana Mackenzie, 2016. 23, 27, 28, 29, 30, 31, 32, 35, 37, 47, 59

LIU, B. Sentiment analysis and opinion mining. **Synthesis lectures on human language technologies**, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012. 19, 25, 26, 27, 29, 37

LIU, J.; CAO, Y.; LIN, C.-Y.; HUANG, Y.; ZHOU, M. Low-quality product review detection in opinion summarization. In: **Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)**. Prague, Czech Republic: Association for Computational Linguistics, 2007. p. 334–342. 27

MCGLOHON, M.; GLANCE, N.; REITER, Z. Star quality: Aggregating reviews to rank products and merchants. In: **Fourth International AAAI Conference on Weblogs and Social Media**. [S.l.: s.n.], 2010. 27

MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. **Ain Shams engineering journal**, Elsevier, v. 5, n. 4, p. 1093–1113, 2014. 28, 29

MICHAELIS, D. B. d. L. P. **Editora Melhoramentos Ltda. 2019**. 2019. 25

MITCHELL, T. M. **Machine learning**. [S.l.]: McGraw hill, 1997. 28

MITTAL, A.; GOEL, A. Stock prediction using twitter sentiment analysis. **Standford University, CS229**, v. 15, 2012. 27

NASUKAWA, T.; YI, J. Sentiment analysis: Capturing favorability using natural language processing. In: **Proceedings of the 2Nd International Conference on Knowledge Capture**. New York, NY, USA: ACM, 2003. (K-CAP '03), p. 70–77. ISBN 1-58113-583-1. 20

NHACUONGUE, J. A. O campo da ciência da informação: contribuições, desafios e perspectivas da mineração de dados para o conhecimento pós-moderno. Universidade Estadual Paulista (UNESP), 2015. 27

O'CONNOR, B.; BALASUBRAMANYAN, R.; ROUTLEDGE, B. R.; SMITH, N. A. From tweets to polls: Linking text sentiment to public opinion time series. In: **Fourth International AAAI Conference on Weblogs and Social Media**. [S.l.: s.n.], 2010. 27

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up? sentiment classification using machine learning techniques. In: **Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (EMNLP '02), p. 79–86. 20

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. 47, 51

POZZI, F. A.; FERSINI, E.; MESSINA, E.; LIU, B. **Sentiment analysis in social networks**. [S.l.]: Morgan Kaufmann, 2016. 25, 26

PROVINCE, B. N. The effects of parameter tuning on machine learning performance in a software defect prediction context. 2015. 38, 39

RAMBOCAS, M.; PACHECO, B. G. Online sentiment analysis in marketing research: a review. **Journal of Research in Interactive Marketing**, Emerald Publishing Limited, v. 12, n. 2, p. 146–163, 2018. 27

RAVI, K.; RAVI, V. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. **Knowledge-Based Systems**, Elsevier, v. 89, p. 14–46, 2015. 19, 22, 28

ROBERTSON, S. Understanding inverse document frequency: on theoretical arguments for idf. **Journal of Documentation**, Journal of Documentation, v. 60, n. 5, p. 503–520, 2004. ISSN 0022–0418. 32

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**. [S.l.]: Malaysia; Pearson Education Limited,, 2016. 34

SAIF, H.; FERNANDEZ, M.; HE, Y.; ALANI, H. Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold. 2013. 59

SAIF, H.; HE, Y.; FERNANDEZ, M.; ALANI, H. Semantic patterns for sentiment analysis of twitter. In: SPRINGER. **International Semantic Web Conference**. [S.l.], 2014. p. 324–340. 59

SALUNKHE, P.; SURNAR, A.; SONAWANE, S. A review: Prediction of election using twitter sentiment analysis. **International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)**, v. 6, n. 5, 2017. 27

SCHUTZE, C. D. M. P. R. H. **Introduction to information retrieval**. [S.l.]: Cambridge University Press, 2008. 32

SINGH, J.; SINGH, G.; SINGH, R. Optimization of sentiment analysis using machine learning classifiers. **Human-centric Computing and Information Sciences**, SpringerOpen, v. 7, n. 1, p. 32, 2017. 21

SOUZA, E.; SANTOS, D.; OLIVEIRA, G.; SILVA, A.; OLIVEIRA, A. L. Swarm optimization clustering methods for opinion mining. **Natural Computing**, Springer, p. 1–29, 2018. 21

TORRES, G. M.; ZAINA, L. A.; ALMEIDA, T. A. de. Aprendizagem em redes socias: uma análise de dados do twitter. In: SBC. **Anais Estendidos do XVIII Simpósio Brasileiro de Sistemas Multimídia e Web**. [S.l.], 2012. p. 87–90. 31

TUMASJAN, A.; SPRENGER, T. O.; SANDNER, P. G.; WELPE, I. M. Predicting elections with twitter: What 140 characters reveal about political sentiment. In: **Fourth international AAAI conference on weblogs and social media**. [S.l.: s.n.], 2010. 27

TURNEY, P. D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: **ACL**. [S.l.: s.n.], 2002. 20

WANI, G. P.; ALONE, N. V. Analysis of indian election using twitter. **International Journal of Computer Applications**, Citeseer, v. 121, n. 22, 2015. 27

YANO, T.; SMITH, N. A. What's worthy of comment? content and comment volume in political blogs. In: **Fourth International AAAI Conference on Weblogs and Social Media**. [S.l.: s.n.], 2010. 27

ZAFARANI, R.; ABBASI, M. A.; LIU, H. **Social Media Mining**. 1. ed. New York: Cambride University Press, 2014. 27, 28, 33