UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Lucas Lima de Oliveira

UTILIZAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREVER A POPULARIDADE DE TUÍTES

Santa Maria, RS
2018

Lucas Lima de Oliveira

UTILIZAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREVER A POPULARIDADE DE TUÍTES

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Bacharel em Sistemas de Informação**.

ORIENTADOR: Prof. Sérgio Luís Sardi Mergen

Santa Maria, RS
2018

Ficha catalográfica elaborada através do Programa de Geração Automática
da Biblioteca Central da UFSM, com os dados fornecidos pelo(a) autor(a).

de Tal, Fulano
TÍTULO DO TRABALHO / Fulano de Tal.-2015.
50 f.; 30cm

Orientador: João da Silva
Coorientadora: Maria da Costa
Tese (doutorado) - Universidade Federal de Santa Maria, Centro de Ciências Naturais e Exatas, Programa de Pós-Graduação em Meteorologia, RS, 2015

1. Teste 1 2. Teste 2 3. Teste 3 I. da Silva, João II. da Costa, Maria III. Título.

End. Eletr.: loliveira@inf.ufsm.com.br

Lucas Lima de Oliveira

UTILIZAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREVER A POPULARIDADE DE TUÍTES

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Bacharel em Sistemas de Informação**.

Aprovado em 29 de novembro de 2018:

_____

Sérgio Luís Sardi Mergen, Dr. (UFSM)
(Presidente/Orientador)

_____

João Carlos Damasceno Lima, Dr. (UFSM)

_____

Joaquim Assunção, Dr. (UFSM)

Santa Maria, RS
2018

DEDICATÓRIA

Dedico este trabalho à minha mãe, Siminea Lima, mulher guerreira e batalhadora que me ensinou a sempre correr atrás dos sonhos e superar os obstáculos. Cada ensinamento passado me permitiu trilhar esse caminho.

AGRADECIMENTOS

Agradeço primeiramente a minha mãe, Siminea Lima e namorada, Amanda Rodrigues, por todo apoio e suporte prestado durante toda a jornada da graduação. Sem vocês ao meu lado, nada disse seria possível.

Agradeço também meu orientador, professor Sérgio Mergen, que além de ter participado da realização deste trabalho, também me deu suporte durante a graduação na realização de outras pesquisas. Agradeço imensamente por ter me acolhido nessa jornada, por todos os conselhos, amizade e por ter acreditado na minha capacidade.

Agradeço minha família e entes queridos, que sempre estiveram presentes e me apoiaram a cada passo dado. O apoio de todas essas pessoas foi fundamental nesta caminhada.

A todos os demais professores do Curso de Sistemas de Informação que sempre mostram dedicação e comprometimento em transmitir seus conhecimentos. Cada ensinamento será levado para toda a vida.

À Universidade Federal de Santa Maria, ao Centro de Tecnologia e à Coordenação do Curso de Sistemas de Informação por fornecer toda a estrutura necessária e proporcionar uma educação de qualidade.

Ao Programa de Educação Tutorial do curso de Sistemas de Informação e todos colegas que fizeram parte do grupo junto comigo. Todas as atividades realizadas pelo grupo me permitiram evoluir como profissional e como ser humano. Ter participado deste grupo foi uma honra e me proporcionou muitas oportunidades de aprendizado.

Agradeço também todos os colegas e amigos feitos durante essa trajetória que permitiram trocas de conhecimentos e experiências inesquecíveis. Espero poder levar essas amizades por toda a vida.

Que todos os nossos esforços estejam sempre focados no desafio à impossibilidade. Todas as grandes conquistas humanas vieram daquilo que parecia impossível.

(Charles Chaplin)

RESUMO

UTILIZAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREVER A POPULARIDADE DE TUÍTES

AUTOR: Lucas Lima de Oliveira
ORIENTADOR: Sérgio Luís Sardi Mergen

É conhecida a popularidade do Twitter e o poder que um único tuíte pode ter nos dias de hoje, servindo, inclusive, como fonte para portais de notícias renomados. A simplicidade e o volume de dados trafegados pela plataforma diariamente, fazem dela uma fonte de dados poderosa. Focando na análise das mensagens veiculadas e no interesse dos usuários em aumentar o número de seguidores, o objetivo deste trabalho é a elaboração de modelos, utilizando algoritmos de aprendizado de máquina, para realizar a predição e classificação da popularidade de tuítes com base em atributos extraídos do corpo das mensagens. Para alcançar esse objetivo, a metodologia adotada envolve a definição dos atributos de interesse, extração e processamento dos dados, além do estudo e aplicação de algoritmos de aprendizado de máquina para realizar a classificação dos tuítes.

Palavras-chave: Aprendizado de Máquina Supervisionado. Classificação de Dados. Coleta de Dados.

ABSTRACT

USE OF MACHINE LEARNING ALGORITHMS TO PREDICT TWEETS POPULARITY

AUTHOR: Lucas Lima de Oliveira
ADVISOR: Sérgio Luís Sardi Mergen

The popularity of Twitter and the power that a single tweet can have today are well-known, even serving as a source for renowned news portals. The simplicity and volume of data trafficked by the platform daily make it a powerful data source. Focusing on the analysis of the messages and in the interest of the users to increase the number of followers, the objective of this work is the elaboration of models, using algorithms of machine learning, to make predictions and classifications of the tweets' popularity based on extracted attributes of the message body. In order to reach this objective, the methodology adopted involves the definition of interest attributes, extraction and processing of data, as well as the study and application of machine learning algorithms to perform the classification of tweets.

Keywords: Supervised Machine Learning. Data Classification. Data collect.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface*
$CART$	*Classification and Regression Trees*
IA	Inteligência Artificial
$LSTM$	*Long Short-Term Memory*
PCA	*Principal Component Analysis*
RNN	*Recurrent Neural Network*
SVM	*Support Vector Machine*

SUMÁRIO

1 INTRODUÇÃO

Com a grande popularização dos chamados influenciadores digitais, é notável o crescimento das mídias sociais como meios de comunicação e divulgação de conteúdos. Neste cenário, onde o número de seguidores determina a sua influência, torna-se muito importante que essas personalidades compreendam seu público, pois conteúdos direcionados refletem diretamente no alcance de suas publicações. Dentre as redes sociais mais utilizadas atualmente, o Twitter é um meio de veiculação de mensagens que se destaca, por sua simplicidade e objetividade. Embora não tenha o mesmo destaque que outras plataformas, como o Facebook ou o Instagram, o Twitter conta com cerca de 335 milhões de usuários ativos, segundo Statista [1], e em média 500 milhões de tuítes que são publicados diariamente, segundo *Internet Live Stats* [2], o que faz dessa rede uma fonte de dados muito poderosa.

Uma das preocupações de usuários do Twitter é alavancar sua popularidade, através do aumento no número de seguidores. Essa preocupação é fundamental para empresas e personalidades públicas que utilizam suas imagens para fins monetários. Nesses casos, o uso das redes sociais deve ser planejado e monitorado. Quando isso é realizado da maneira correta, a marca e/ou a pessoa ficam muito mais próximos de seus fãs e seguidores, o que consequentemente, faz sua popularidade e influência aumentar. Um dos indicadores capaz de medir a influência de um usuário em redes sociais é a taxa de engajamento, que leva em consideração as interação dos usuários com as publicações de uma página, dentre essas interações, podem ser considerados os retuítes, curtidas e comentários. Considerando esse fator, pode-se afirmar empiricamente que o aumento na quantidade de retuítes leva a um aumento na quantidade de seguidores, devido a propagação exponencial daquele conteúdo.

Tendo em vista o interesse dos usuários em aumentar o alcance de suas postagens, poder identificar os fatores que têm maior influência sobre a popularidade de suas mensagens pode ser uma grande vantagem ao tentar aumentar o engajamento por parte de seus seguidores. Ser capaz de prever/estimar a popularidade que um tuíte poderá obter, baseando-se nas características presentes no corpo de sua mensagem, permite a realização de diferentes análises a cerca o conteúdo disseminado por aquela conta, o que pode trazer muitos benefícios aos usuários com relativa influência nessa rede social.

Como afirma (SUH et al., 2010), a propagação de um tuíte está diretamente ligada ao conteúdo e valor informativo contido nele. Nesse sentido, os autores avaliaram um conjunto de características extraídas das mensagens. Os resultados mostra-

[1] *Statista:* `https://www.statista.com/topics/737/twitter/`

[2] *Internet Live Stats:* `http://www.internetlivestats.com/twitter-statistics/`

ram que a utilização de *hashtags* e URLs são fatores muito significativos e que ajudam a impulsionar uma publicação. Apesar de ser um resultado muito relevante, o trabalho não realizou uma análise exaustiva das características que podem ser extraídas do corpo das mensagens de cada tuíte.

É conhecido que hoje existem inúmeras pesquisas sendo realizadas envolvendo dados coletados do Twitter. Além de (SUH et al., 2010), outros trabalhos relacionados que podem ser citados aqui, como (DUAN et al., 2010), (BENEVENUTO et al., 2010), (NAVEED et al., 2011), (KHARDE; SONAWANE, 2016), (HONG; DAN; DAVISON, 2011) e (XU; YANG, 2012), tem como parte de seus objetivos, a identificação de fatores impactantes no conteúdo das mensagens propagadas no Twitter, além disso, utilizam também técnicas de aprendizado de máquina como forma de realizar suas pesquisas.

Dentre os trabalhos citados logo acima, (DUAN et al., 2010), (BENEVENUTO et al., 2010), (KHARDE; SONAWANE, 2016) e (XU; YANG, 2012) utilizam técnicas de Máquina de Vetores de Suporte (SVM, do inglês: Support Vector Machine) na realização de seus experimentos. O trabalho (BENEVENUTO et al., 2010) também utilizou métodos da ferramenta Weka para identificar a importância e selecionar os atributos. O trabalho de (KHARDE; SONAWANE, 2016) também utilizou o algoritmo Naive Bayes em sua pesquisa, além de outras técnicas também apresentadas. O trabalho de (XU; YANG, 2012) também utilizou árvores de decisão para realizar a classificação dos dados.

Tratando-se da proposta dos trabalhos, (SUH et al., 2010), (NAVEED et al., 2011) e (HONG; DAN; DAVISON, 2011) realizam análises sobre fatores que influenciam na popularidade de um tuíte. Além disso, os trabalhos (SUH et al., 2010), (DUAN et al., 2010) e (BENEVENUTO et al., 2010) identificaram a utilização de URLs, e o alcance do autor do tuíte (podendo ser medido por seus seguidores ou métricas mais complexas) como fatores de grande relevância durante seus experimentos.

Dentro deste contexto, o objetivo deste trabalho é monitorar e extrair tuítes de determinadas contas do Twitter, a fim de elaborar modelos, utilizando algoritmos de aprendizado de máquina, para realizar a predição e classificação da popularidade de tuítes com base em suas características. Serão estudados e testados algoritmos já consolidados, como Naive Bayes e árvores de decisão. Como entrada para estes algoritmos, serão utilizados dados provenientes do pré-processamento dos tuítes coletados, sendo consideradas três características que não foram contempladas pelo estudo de (SUH et al., 2010): o tamanho em caracteres, o sentimento (que mede a emoção transmitida) e a banalidade (que mede a relevância da mensagem). Para fins de comparação, a presença de *hashtags* e URLs também foi avaliada.

A escolha dos algoritmos de classificação foi feita com base em sua popularidade e destaque para solucionar o problema de classificação de dados. Além disso,

são frequentemente analisados e aplicados por pesquisadores em artigos que utilizam o aprendizado de máquina como foco.

Este trabalho está estruturado nas seguintes seções. O capítulo 2 apresenta a fundamentação teórica, abordando conceitos e algoritmos de aprendizado de máquina. O capítulo 3 apresenta a definição dos atributo e a arquitetura de extração de tuítes usada, que realiza desde a coleta até a preparação dos dados para análise. O capítulo de Experimentos apresentará as análises realizadas a partir dos dados coletados juntamente com a aplicação dos algoritmos de aprendizado de máquina estudados. O capítulo de Conclusões apresenta as considerações finais a cerca do trabalho realizado.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados os conceitos relacionados ao aprendizado de máquina, na seção 2.1, definindo as diferenças entre a aprendizagem supervisionada e a não supervisionada. Em seguida, na seção 2.3, são apresentados alguns dos principais algoritmos do segmento supervisionado, os quais também foram utilizados na realização de experimentos no decorrer deste trabalho, sendo estes o Naive Bayes, Árvores de Decisão e Redes Neurais.

2.1 APRENDIZADO DE MÁQUINA

Entende-se como sistemas inteligentes, aqueles que são capazes de processar dados de entrada e ajustar padrões internos a fim de otimizar seus resultados de saída, de acordo com os objetivos esperados para aquele algoritmo. Dentro deste contexto, o aprendizado de máquina foca no treinamento desses algoritmos para melhorar seu desempenho. Esse processo está ligado com a redução de dimensionalidade, classificação e associação dos dados e previsão de comportamentos.

Algoritmos de aprendizado de máquina (ou *machine learning* em inglês) dividem-se em dois segmentos, aqueles que necessitam de uma supervisão para melhorar seus resultados e aqueles fazem esse processo de maneira independente. Nesta seção serão apresentados esses dois tipos de algoritmos, especificando suas características e diferenças.

2.1.1 Aprendizado de Máquina Supervisionado

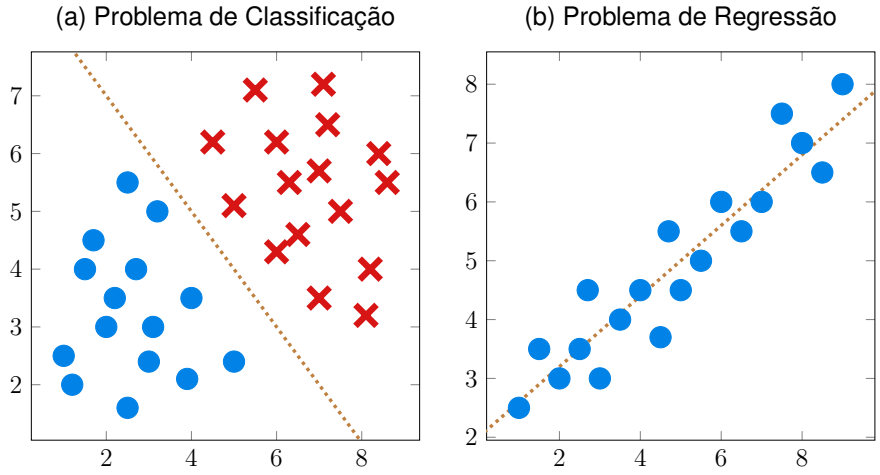
A aprendizagem supervisionada realiza o treinamento dos algoritmos com dados para os quais suas respostas já sejam conhecidas. Ou seja, dependem sempre da entrada de um padrão de valores e da comparação das respostas do sistema com aquelas consideradas corretas. Conforme o algoritmo é treinado seus padrões vão sendo ajustados a fim de diminuir o erro e otimizar as respostas.

Os problemas solucionados através da aprendizagem supervisionada são divididos em problemas de regressão e classificação de dados, como ilustra a Figura 2.1. Segundo (RUSSELL; NORVIG, 2014), quando o resultado esperado pelo algoritmo for um conjunto finito de valores, (como fraco, mediano ou forte), trata-se de problema de classificação, pois os dados de entrada devem ser categorizados dentro daquele

grupo. No caso do resultado esperado ser numérico, trata-se de um problema de regressão, na qual tenta-se identificar uma tendência nos valores com base nos dados de entrada.

Nesse tipo de aprendizagem o algoritmo recebe as entradas já categorizadas para realizar o treinamento e, a cada iteração, ajusta seus parâmetros para de obter a melhor saída, podendo ser, por exemplo, minimizar o erro, maximizar a precisão ou a acurácia. Frequentemente, após a etapa de treinamento, é realizada uma etapa de validação, passando ao algoritmo entradas sem classificação, dessa forma seu desempenho pode ser realmente avaliado e, se necessário, o treinamento pode ser realizado novamente com novos ajustes em seus parâmetros.

Figura 2.1 – Exemplo de Problemas de Classificação e Regressão.



Fonte: Produção do próprio autor.

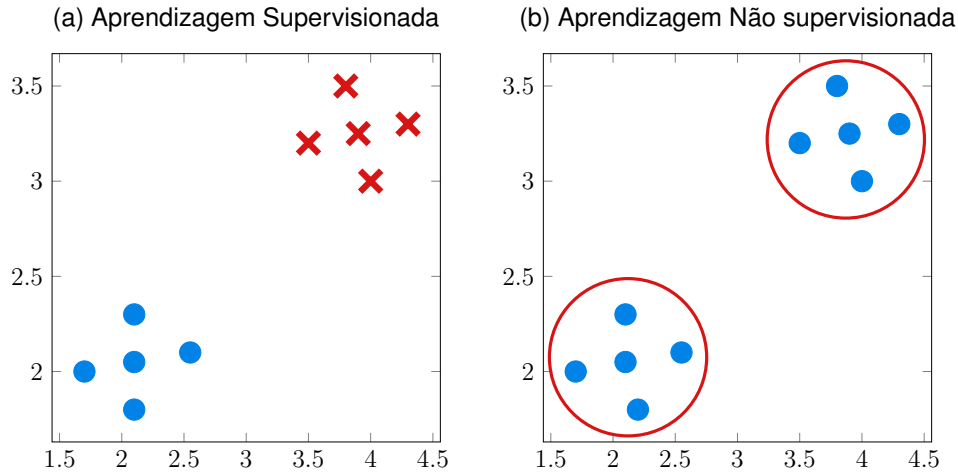
2.1.2 Aprendizado de Máquina Não Supervisionado

No caso dos algoritmos de aprendizagem não supervisionada, ao contrário do segmento apresentado na subseção anterior, estes recebem os dados sem nenhum classificação prévia, impossibilitando o aferimento das classes de cada entrada. Consequentemente, conforme os dados vão sendo recebidos, o próprio algoritmo é responsável por identificar as relações e padrões presentes nos dados, o que por si só pode ser considerado um objetivo a ser alcançado. A aprendizagem não supervisionada não prevê soluções específicas para realizar o treinamento e validação dos resultados, ou seja, não há um *feedback* explícito sobre os resultados previstos.

Como explica (RUSSELL; NORVIG, 2014), o exemplo mais comum de aprendizagem não supervisionada, é o de agrupamento, onde o objetivo é detectar grupos potencialmente úteis dentro dos valores de entrada, que podem ser semelhantes ou

estar relacionados por diferentes variáveis. A Figura 2.2 exemplifica as diferenças entre esses dois tipos de abordagem.

Figura 2.2 – Exemplo da diferença entre as diferentes abordagens.



Fonte: Produção do próprio autor.

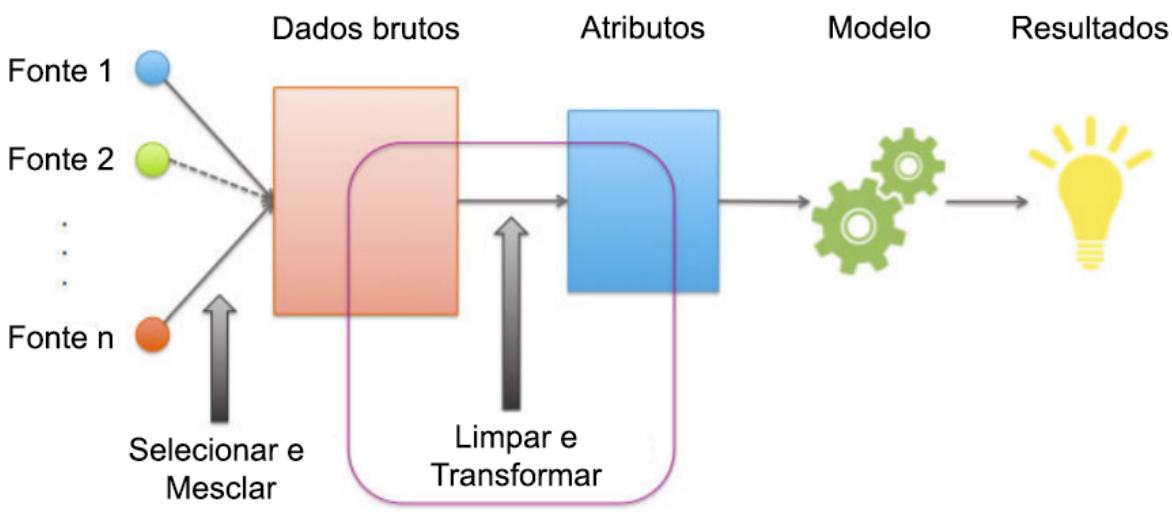
2.2 ENGENHARIA DE ATRIBUTOS

Engenharia de atributos (ou *Feature Engineering*) é um conjunto de técnicas muito utilizadas no processo de aprendizado de máquina. Essas técnicas tem o objetivo de agregar valor mais significativo aos dados coletados e assim, melhorar os modelos de IA (ZHENG; CASARI, 2018). Muitas vezes, o processo de engenharia de atributos por si só pode trazer melhores resultados independente dos algoritmos utilizados. Como já disse Peter Norvig, escritor de renomados livros na área de inteligência artificial e atualmente diretor de pesquisa no Google, "mais dados batem algoritmos inteligentes, mas dados melhores batem mais dados", ou seja mesmo utilizando algoritmos poderosos de aprendizagem de máquina, os resultados não serão tão bons se os dados não forem significativos.

A escolha de quais técnicas utilizar está ligada, na maioria dos casos, tanto aos dados quanto ao modelo, pois alguns deles podem estar mais adaptados a um determinado tipo de atributo (ZHENG; CASARI, 2018). A Figura 2.3 mostra onde o a aplicação de engenharia de atributos se localiza no processo de aprendizagem de máquina e torna claro que as escolhas tomadas em qualquer uma das etapas tem influência direta nas suas subsequentes.

Como mencionado, existem inúmeras técnicas de engenharia de atributos dependendo do tipo de dados, e para cada uma delas, existem diferentes estratégias para sua aplicação. Para dados numéricos, podem ser aplicadas técnicas como: pre-

Figura 2.3 – Engenharia de Atributos no processo de Aprendizado de Máquina



Fonte: Traduzido de (ZHENG; CASARI, 2018)

enchimento de valores faltantes, que tem por objetivo evitar a perda de informações; arredondamento de valores, pois muitas casas decimais podem representar ruídos; transformação logarítmica, para reduzir a diferença na escala entre valores muito grandes e outros muito pequenos; normalização, para padronizar a escala dos valores; dentre outras.

No caso de dados textuais, existem técnicas para realizar a preparação do texto, ou pré-processamento, que podem envolve alguns procedimentos como transformação em letras minúsculas, lematização das palavras, remoção de acentuações, caracteres não textuais e palavras comuns (ou *stopwords*). Além disso, outros métodos de engenharia de atributos podem ser aplicados, como vetorização do texto com técnicas como *bag-of-words* e *word embeddings* (respectivamente bolsa de palavras e incorporação de palavras, em tradução livre), que buscam determinar a frequência e similaridade das as palavras presentes nos textos.

Em consequência de todos os procedimentos, muitos atributos novos podem ser gerados e o processamento de todos eles pode se tornar muito custoso. Nestes casos, existem abordagens com o intuito de selecionar os atributos mais representativos dentre todos disponíveis. Dentre estas abordagens de seleção, pode-se citar Análise de Componentes Principais(PCA, do inglês: *Principal Component Analysis*); método de filtragem, que busca identificar correlações entre os dados; método *Wrapper* (ou de embrulhar, em tradução livre), que é baseado na tentativa e erro para encontrar a melhor combinação de atributos; e o método *Embedded* (ou incorporado, em tradução livre), casos em que a seleção destes atributos já faz parte do modelo.

2.3 ALGORITMOS DE APRENDIZADO DE MÁQUINA SUPERVISIONADO

Dentro do escopo deste trabalho, que tem como um dos objetivos realizar a predição da popularidade de tuítes, requere-se a utilização da aprendizagem de máquina supervisionada, pois os resultados esperados estão diretamente ligados a classificação dos dados. Como mencionado, nesta seção serão abordados alguns dos principais algoritmos que se encaixam neste segmento e que serão utilizados no decorrer deste trabalho, apresentado suas características, funcionamento, vantagens e desvantagens na sua utilização.

2.3.1 Naive Bayes

A técnica Naive Bayes pode ser considerada como uma das mais populares para classificação de dados utilizando aprendizado de máquina. O algoritmo utiliza de métodos probabilísticos, baseados na Teoria Bayesiana, criada por Thomas Bayes no século XVIII. Para compreender melhor o funcionamento dessa técnica, é importante entender também um pouco sobre o teorema do qual ela teve origem.

Como mostra (RUSSELL; NORVIG, 2014), o teorema, ou regra de Bayes é uma formula simples, definida pela Equação 2.1, que vem da regra do produto de probabilidades, assumindo que *prob(D|H) = prob(H|D)*, sendo H a hipótese a ser validada e D os dados observados, podendo ser tratados também como *causa* e *efeito*. Apesar de simples, essa regra é a base de grande parte dos sistemas de IA (Inteligência Artificial) que utilizam inferência probabilística.

$$prob(H|D) = \frac{prob(D|H)prob(H)}{prob(D)} \tag{2.1}$$

Dividindo as partes do teorema, do lado esquerdo, *prob(H|D)* é chamada de probabilidade posterior da hipótese após a realização do experimento; do lado direito, *prob(D|H)* chamada função de verossimilhança, é a distribuição de probabilidade dos dados, a qual multiplica-se por *prob(H)*, denominada *Prior*, que é a probabilidade da hipótese ser verdadeira; por fim, o denominador *prob(D)*, é a probabilidade total.

Ainda que possa parecer um teorema simples, seu alcance está na sua capacidade de interpretação. No caso do modelo Naive Bayes, ou Bayes Ingênuo, assume-se que os atributos *efeito* são condicionalmente independentes entre si, dada a *causa* – daí a denominação de "ingênuo". A distribuição probabilística deste modelo pode ser descrita conforme indica a Equação 2.2, sendo *C* a classe, ou causa, que deve ser prevista, enquanto que o conjunto $\{x_1, ..., x_n\}$ são os atributos, ou efeitos.

$$P(C|x_1, ..., x_n) = \alpha P(C) \prod_i P(x_i|C) \tag{2.2}$$

Este modelo de aprendizagem é facilmente escalável para problemas maiores, funcionando muito bem com uma ampla variedade de aplicações, apesar de se destacar e ser comumente utilizado em uma série algoritmos para classificação de textos. Além disso, este modelo não apresenta grandes complicações com dados ruidosos ou faltantes, podendo inclusive realizar previsões adequadas nestes casos. Esses fatores fazem o Naive Bayes ser (provavelmente) o modelo de rede Bayesiana mais comumente utilizado em algoritmos de aprendizado de máquina.

Tomando como exemplo a clássica classificação de sentimentos em textos, como mencionado, o algoritmo irá assumir que as palavras de uma determinada mensagem não possuem uma relação entre si. Sendo assim, o classificador poderá presumir que uma frase seja positiva, caso a maioria das palavras presentes nela tenham maior probabilidade de ter este mesmo sentimento, independentemente do contexto em que foram utilizadas.

Para classificar uma determinada frase, inicialmente é preciso montar uma base de treinamento, contendo a classificação dos dados de entrada, que no caso da análise de sentimentos, será positivo ou negativo. A partir destes dados, é criada uma tabela para guardar a frequência de cada uma das entradas com suas classes e a probabilidades de cada entrada. Para testar uma nova entrada, é calculada sua probabilidade para cada uma das possíveis classificações com base nas ocorrências anteriores. Para os casos em que o dado de teste não está presente na base de treinamento ou não foi classificado para uma das classes, existem algumas técnicas capazes de corrigir esse problema. Uma técnica muito comum aplicada para estes casos é a suavização de Laplace, a qual soma o valor 1 para todos os valores, desta forma, nenhuma operação é realizada utilizando o valor 0.

Utilizando como exemplo a frase "*With great power comes great responsibility*", e considerando a Tabela 2.1 (fictícia) apresentada logo abaixo, na qual consta a frequência das palavras para cada classe e a probabilidade de cada uma, para classificar a palavra "*great*" como popular ou impopular, considerando também que essa possa ser uma frase extraída do Twitter, seriam realizadas as seguintes operações listadas a seguir.

Tabela 2.1 – Exemplo de tabela com frequências de palavras e suas classes.

Palavras	Popular	Impopular	Probabilidade
responsibility	1	2	3/14 = 0,21
power	2	1	3/14 = 0,21
great	3	1	4/14 = 0,28
bad	0	2	2/14 = 0,14
good	2	0	2/14 = 0,14
Total	8	6	
		Positivo	8/14 = 0,57
		Negativo	6/14 = 0,42

Fonte: Produção do próprio Autor.

$$P(great|popular) = 3/8 = 0.37 \tag{2.3}$$
$$P(popular) = 8/14 = 0.57 \tag{2.4}$$
$$P(great) = 4/14 = 0.28 \tag{2.5}$$
$$P(great|unpopular) = 1/6 = 0.16 \tag{2.6}$$
$$P(unpopular) = 6/14 = 0.42 \tag{2.7}$$

$$P(popular|great) = 0.37 * 0.57/0.28 = 0.75 \tag{2.8}$$
$$P(unpopular|great) = 0.16 * 0.42/0.28 = 0.24 \tag{2.9}$$

A partir dos cálculos realizados, com base na Tabela 2.1 apresentada, obtém-se como resultado uma probabilidade maior para a palavra '*great*' ser popular. Para realizar a classificação considerando toda a frase, essa operação é aplicada para cada palavra, as probabilidades resultantes para cada classe são multiplicadas e os resultados são aplicados na regra de Bayes, conforme a Equação 2.1, para cada uma das possíveis classes. Mesmo sendo um exemplo simples da aplicação da técnica Naive Bayes, é possível observar a facilidade da aplicação deste algoritmo para a classificação de dados utilizando um método probabilístico.

2.3.2 Árvores de Decisão

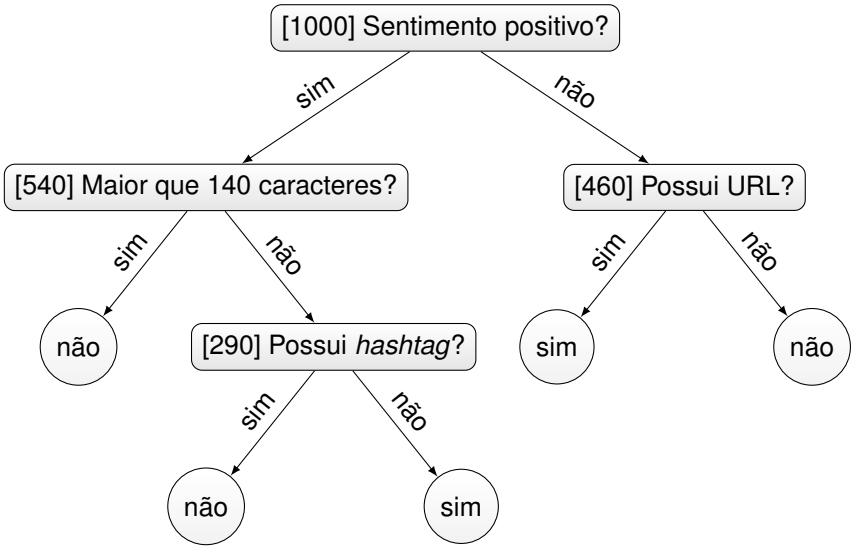
Abstraindo o conceito computacional, uma árvore de decisão pode ser definida por um fluxograma, no qual cada nó, com exceção do último nível, representa um teste

sobre as informações disponíveis. O ponto de partida é denominado nó raiz. A medida que os nós filhos vão sendo explorados, as informações são divididas com o objetivo de agrupá-las por similaridade e buscando o balanceamento entre os subgrupos. Ao percorrer toda a árvore, os últimos elementos, denominados nós folha, representam a decisão a ser tomada. Apesar de ser um conceito simples, a complexidade computacional desta técnica está no processo de indução da estrutura da árvore, feita de maneira automática e não-paramétrica, podendo lidar com dados multidimensionais.

Assim como outras técnicas dentro do escopo de aprendizagem supervisionada, árvores de decisão também são muito populares para a resolução de problemas de classificação de dados e regressão linear. Segundo (HAN; PEI; KAMBER, 2011), a popularização deste tipo de algoritmo na aprendizagem de máquina está diretamente ligada à sua característica não-paramétrica, que permite a indução de árvores sem a o total domínio ou configuração prévia dos dados, o que torna-se muito interessante no âmbito deste trabalho no que se refere à descoberta de maneira exploratória.

Ainda conforme (HAN; PEI; KAMBER, 2011), para realizar a classificação de dados, cada registro percorre um determinado caminho dentro da estrutura da árvore, partindo do nó raiz até o nó folha, o qual determina a classe para aquela entrada de dados. Para exemplificar, foi criada a árvore de decisão fictícia apresentada na Figura 2.4, a qual considera 4 atributos de uma mensagem de texto para considerá-la como popular ou não. Na figura, os nós com o formato retangular representam os testes feitos com cada registro de entrada, sendo que antes de cada teste é indicado o número de instâncias que estão naquele nó, partindo da raiz com 1000, já os nós com o formato circular são os nós folha, que representam a classificação final para indicar se a mensagem seria considerada como popular ou não.

Figura 2.4 – Exemplo de árvore para classificação de um tuíte como popular.



Fonte: Produção do próprio Autor.

Utilizando novamente como exemplo a frase "*With great power comes great responsibility*", ao aplicá-la na árvore de decisão apresentada, ela seria classificada como popular, pois percorreria o seguinte caminho: Sentimento positivo; Menor que 140 caracteres; e não possui *hashtag*. Uma diferença em relação a técnica *Naive Bayes* que já é possível notar através deste exemplo, é que árvores de decisão são capazes de lidar com a correlação entre os atributos.

Mesmo existindo vários vários algoritmos com diferentes propostas para realizar a indução de árvores de decisão, duas etapas estão presentes na grande maioria deles durante o construção da árvore, a seleção das medidas de atributos e a "poda da árvore" que, respectivamente, são responsáveis por definir quais a melhores partições dos dados; e por remover, ou reduzir, ruídos nas ramificações gerados durante o treinamento. Apesar de serem etapas comuns na implementação deste tipo de algoritmo, também são as etapas que os diferenciam uns dos outros dependendo da abordagem adotada para realiza-las.

Um dos algoritmos que merece destaque por ser referência neste âmbito, é o CART (*Classification and Regression Trees*), criado em 1984 por um grupo de estatísticos (L. Breiman, J. Friedman, R. Olshen, and C. Stone). Ele realiza uma abordagem de construção recursiva de divisão e conquista partindo de cima para baixo, segundo (HAN; PEI; KAMBER, 2011).

3 PROPOSTA

A proposta deste trabalho é a elaboração de modelos, utilizando algoritmos de aprendizado de máquina supervisionada, capaz de classificar o nível de popularidade de tuítes com base na correlação entre a popularidade dos mesmos em função de um conjunto de características presentes no corpo das mensagens. Para atingir esse objetivo, é necessário coletar os tuítes, extrair suas características e aplicar o algoritmos já mencionados para realizar o treinamento e classificação dos dados. A partir disso, esta seção apresenta a definição dos atributos, a definição de popularidade e os classificadores utilizados para a realização do trabalho.

3.1 DEFINIÇÃO DOS ATRIBUTOS

Como parte do objetivo deste trabalho é a correlação entre a popularidade e as características do texto de cada tuíte, é de fundamental importância a definição e extração de características relevantes que possam influenciar no interesse dos usuários sobre uma determinada mensagem. Esta etapa corresponde a definição dos atributos que serão extraídos de cada um dos tuítes coletados. Os itens abaixo definem cada um destes atributos e a razão de terem sido escolhidos:

Presença de URLs: O uso de URLs em um tuíte pode indicar uma informação proveniente de outros meios, podendo ser sites de notícias ou outras mídias sociais, o que pode despertar, ou não, o interesse de usuários por um determinado tipo de informação. Esse atributo é representado pelo tipo de dados booleano, podendo ser verdadeiro ou falso.

Presença de *hashtags*: De maneira geral, as *hashtags* são palavras-chave ou termos utilizados para indicar que uma determinada mensagem está diretamente ligada a um tópico ou discussão em específico. O que, de maneira semelhante ao uso de URLs, pode atrair o interesse de usuários por tópicos específicos. Este atributo também é do tipo booleano.

Tamanho da mensagem: Essa característica é basicamente a contagem da quantidade de caracteres usados no corpo do tuíte, que pode fazer com que os usuários percam o interesse em ler seu conteúdo, por ser muito curto ou muito extenso. Por tratar-se de um valor contínuo, este atributo é representado por um valor inteiro.

Sentimento da mensagem: O sentimento é um valor que classifica o teor do texto como positivo ou negativo. Fator que pode estar diretamente ligado a intenção de cada usuário em propagar mensagens com um determinado humor. Este atributo tam-

bém pode chamado de polaridade da mensagem e trata-se de um valor decimal, que pode variar entre -1 e 1, onde -1 corresponde a uma mensagem totalmente negativa, 0 corresponde a neutra e 1 corresponde a totalmente positiva.

Banalidade da mensagem: No contexto deste trabalho, a banalidade corresponde à importância do que foi escrito no corpo do tuíte, levando em consideração a presença de palavras que são frequentemente usadas em textos escritos. Sendo assim, quanto maior o número de palavras frequentes, mais banal é a mensagem. Este atributo é representado por um valor decimal, que varia entre 0 e 1, sendo que quanto mais próximo de 1, mais banal é a mensagem. O cálculo desta métrica utiliza a Equação 3.1, apresentada logo abaixo.

$$\frac{\sum_{i=1}^{n}(freq(P_i))}{n} \tag{3.1}$$

onde o conjunto $\{P_1, ..., P_n\}$ são as palavras da mensagem após a remoção de *stopwords* (preposições e artigos que normalmente são descartados durante o processamento de um texto). Já a função $freq(P)$ retorna 1 caso a palavra P seja frequente e zero caso não seja.

3.2 DEFINIÇÃO DE POPULARIDADE

De maneira geral, em mídias sociais, a popularidade de uma conta pode ser medida através da quantidade de seguidores que ela detém, quanto maior o número de seguidores, mais influente, ou popular, a conta é considerada. Porém, este é um indicador simples que não determina o alcance real das publicações. Para isso, existem várias métricas que permitem uma medição mais precisa sobre o impacto causado pelas ações realizadas por uma determinada página ou usuário. Uma métrica muito conhecida e utilizada para medir o alcance real de uma página sobre seus seguidores é a taxa de engajamento. Esse índice considera as interações dos fãs com os conteúdos publicados, de forma que quanto maior é essa interação, maior é o nível de engajamento.

Como exposto em (PILLAT; PILLAT, 2017), para calcular a taxa de engajamento de uma determinada publicação, por convenção, é realizada a fórmula apresentada na Equação 3.2. Cada elemento da equação refere-se estritamente ao valor, em quantidade, obtido por cada publicação. Trazendo para a realidade do Twitter, os compartilhamentos são substituídos pelos retuítes e os comentários pelas respostas a um determinado tuíte.

$$E(x) = \frac{curtidas + compartilhamentos + comentários}{seguidores} * 100 \tag{3.2}$$

Apesar de existir esta convenção para o cálculo do engajamento, a fórmula pode variar, dependendo das informações fornecidas por cada rede social. Como por exemplo, no caso do Facebook, o total de seguidores pode ser substituído pelo total de visualizações obtidas por cada publicação, ou então, como também apresentado em (PILLAT; PILLAT, 2017), substituído pelo seguidores da própria página mais os seguidores dos próprios fãs.

3.3 CLASSIFICADORES

Para a realização da predição da popularidade de cada tuíte tendo como base a taxa de engajamento e considerando os atributos já mencionados, serão utilizados os algoritmos de classificação Naive Bayes e Árvores de decisão. Os experimentos com cada algoritmo serão realizados através da implementação de código utilizando a linguagem Python, devido a sua popularidade e recursos para lidar com aprendizado de máquina e manipulação de dados, ou com o auxilio da ferramenta Weka, que já oferece uma coleção de algoritmos capazes de realizar a classificação de dados.

4 METODOLOGIA

Para alcançar os objetivos elencados, uma série de etapas foi elaborada para a realização deste projeto, nas quais o sucesso de uma está diretamente ligado ao da outra. A fim de proporcionar maior compreensão destas etapas, este capítulo foi destinado à descrição de cada uma destas etapas juntamente com a apresentação do cronograma do projeto, no qual consta a previsão de realização de cada uma delas.

Análise de trabalhos similares: Nesta etapa será realizada uma pesquisa sobre trabalhos similares já realizados na área envolvendo aprendizado de máquina e extração de dados extraídos do Twitter, com o objetivo de que os mesmos possam vir a auxiliar e contribuir com a proposta do trabalho.

Definição dos atributos: Esta etapa é designada para analisar e definir quais serão os atributos de interesse que devem ser extraídos de cada tuíte, considerando a relevância que cada um deles pode inferir sobre a popularidade daquele tuíte.

Extração e pré-processamento dos dados: Dispondo dos atributos de interesse, os tuítes devem ser coletados e seus conteúdos pré-processados. Para isso, será realizado um estudo acerca dos melhores métodos e ferramentas que permitem a extração de dados do Twitter e posteriormente de seus atributos de interesse.

Estudo de algoritmos de aprendizado de máquina: Estando em posse dos dados já processados, será realizado um estudo acerca dos algoritmos de aprendizado de máquina para classificação de dados e quais são as melhores opções considerando os dados de entrada e os atributos de interesse.

Implementação dos algoritmos: Dispondo do conhecimento necessário acerca algoritmos de aprendizado de máquina e os atributos disponíveis, serão implementados tais algoritmos utilizando tecnologias adequadas para tal.

Experimentos com os algoritmos selecionados: Esta etapa corresponde a realização de testes com os algoritmos estudados e implementados nas etapas anteriore. O experimentos serão realizados com o objetivo de classificar a popularidade dos tuítes com base nos atributos extraídos.

Análise dos resultados obtidos: Esta etapa do trabalho destina-se a realização de uma análise sobre os algoritmos testados e os resultados obtidos com cada um deles ao tentar realizar a predição da popularidade dos tuítes.

Documentação: Por fim, é nesta etapa que os resultados obtidos com o decorrer do projeto são analisados e documentados, havendo a chance de originar um artigo, considerando que houve certa contribuição científica com a área, além do próprio relatório do Trabalho de Conclusão de Curso.

4.1 CRONOGRAMA

Com base nas etapas descritas, a Tabela 4.1 a seguir resume o cronograma das atividades previstas. Dividido entre quinzenas, o cronograma tem início a partir do mês de Agosto (08). As atividades já realizadas até o momento compreendem os itens marcados com bola (•) e fundo em cinza, enquanto que os marcadores em xis ($\times$) com fundo branco representam as atividades que ainda devem ser realizadas.

Tabela 4.1 – Cronograma do projeto

Atividade	**8/2**	**9/1**	**9/2**	**10/1**	**10/2**	**11/1**	**11/2**	**12/1**
Análise de trabalhos similares	•							
Definição dos Atributos	•	•	•					
Extração e pré-processamento		•	•	•	•			
Estudo de algoritmos de IA				•	•	•		
Implementação dos algoritmos				•	•	$\times$	$\times$	$\times$
Experimentos com algoritmos				•	•	$\times$	$\times$	$\times$
Análise dos resultados obtidos						$\times$	$\times$	$\times$
Documentação	•	•	•	•	•	$\times$	$\times$	$\times$

REFERÊNCIAS BIBLIOGRÁFICAS

BENEVENUTO, F. et al. Detecting spammers on twitter. In: **Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)**. [S.l.: s.n.], 2010. v. 6, n. 2010, p. 12.

DUAN, Y. et al. An empirical study on learning to rank of tweets. In: **Proceedings of the 23rd International Conference on Computational Linguistics**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. (COLING '10), p. 295–303. Disponível em: <http://dl.acm.org/citation.cfm?id=1873781.1873815>.

HAN, J.; PEI, J.; KAMBER, M. **Data Mining: Concepts and Techniques**. Elsevier Science, 2011. (The Morgan Kaufmann Series in Data Management Systems). ISBN 9780123814807. Disponível em: <https://books.google.com.br/books?id=pQws07tdpjoC>.

HONG, L.; DAN, O.; DAVISON, B. D. Predicting popular messages in twitter. In: **Proceedings of the 20th International Conference Companion on World Wide Web**. New York, NY, USA: ACM, 2011. (WWW '11), p. 57–58. ISBN 978-1-4503-0637-9. Disponível em: <http://doi.acm.org/10.1145/1963192.1963222>.

KHARDE, V. A.; SONAWANE, S. Sentiment analysis of twitter data : A survey of techniques. **CoRR**, abs/1601.06971, 2016. Disponível em: <http://arxiv.org/abs/1601.06971>.

NAVEED, N. et al. Bad news travel fast: A content-based analysis of interestingness on twitter. In: **Proceedings of the 3rd International Web Science Conference**. New York, NY, USA: ACM, 2011. (WebSci '11), p. 8:1–8:7. ISBN 978-1-4503-0855-7. Disponível em: <http://doi.acm.org/10.1145/2527031.2527052>.

PILLAT, V. G.; PILLAT, V. G. Comparação entre duas fórmulas utilizadas para o cálculo da taxa de engajamento utilizando como base a porcentagem de visualizações e o total de fãs. **Revista Brasileira de Pesquisas de Marketing**, 2017. ISSN 2317-0123. Disponível em: <http://www.revistapmkt.com.br/pt-br/anteriores/anteriores.aspx?udt_863_param_detail=8650>.

RUSSELL, S.; NORVIG, P. **Inteligência artificial: Tradução da 3a Edição**. Elsevier Editora Ltda., 2014. 1056 p. ISBN 9788535251418. Disponível em: <https://books.google.com.br/books?id=BsNeAwAAQBAJ>.

SUH, B. et al. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: IEEE. **2010 IEEE Second International Conference on Social Computing**. 2010. p. 177–184. Disponível em: <http://ieeexplore.ieee.org/document/5590452/>.

XU, Z.; YANG, Q. Analyzing user retweet behavior on twitter. In: **2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining**. [S.l.: s.n.], 2012. p. 46–50.

ZHENG, A.; CASARI, A. **Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists**. [S.l.]: "O'Reilly Media, Inc.", 2018.