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 25 de dezembro 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

À família!

AGRADECIMENTOS

Aos professores e colegas que, sem eles, nada disso seria possível!

O livro é uma criatura frágil, ele sofre o desgaste do tempo, ele teme os roedores, os elementos e mãos desajeitadas. Então o livreiro proteje os livros não apenas da humanidade, mas também da natureza e devota sua vida a uma guerra contra as forças do esquecimento.

(Umberto Eco)

RESUMO

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

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

Resumo aqui.

Palavras-chave: Palavra Chave 1. Palavra 2. Palavra 3. (...)

ABSTRACT

USE OF MACHINE LEARNING ALGORITHMS TO PREDICT TWEETS POPULARITY

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

Abstract here.

Keywords: Keyword 1. Keyword 2. Keyword 3. (...)

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface*
IA	Inteligência Artificial
$CART$	*Classification and Regression Trees*

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. Como mencionado em (OLIVEIRA; MERGEN, 2018), um dos indicadores capaz de medir a influência de um usuário no Twitter, é a quantidade de retuítes que suas mensagens recebem, presente inclusive na fórmula para o cálculo de engajamento, o qual será explanado posteriormente. 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 relevante, o trabalho não realizou uma análise exaustiva das características que podem ser extraídas das mensagens.

Dentro deste contexto, o objetivo deste trabalho é monitorar e extrair tuítes de determinadas contas do Twitter, a fim de elaborar um modelo, utilizando algoritmos de aprendizado de máquina, para realizar a predição e classificação da popularidade de tuítes com base em suas características. Devido ao grande volume de dados, faz-se necessário automatizar o processo de análise e classificação dos dados, para isso, serão estudados e testados algoritmos já consolidados, como Naive Bayes, J48 e LTM, contando inclusive com o auxílio de ferramentas já reconhecidas como Weka, para a aplicação e análise dos resultados destes algoritmos. 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.

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 5 apresenta as análises realizadas a partir dos dados coletados juntamente com os algoritmos de aprendizado de máquina estudados. O capítulo 6 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

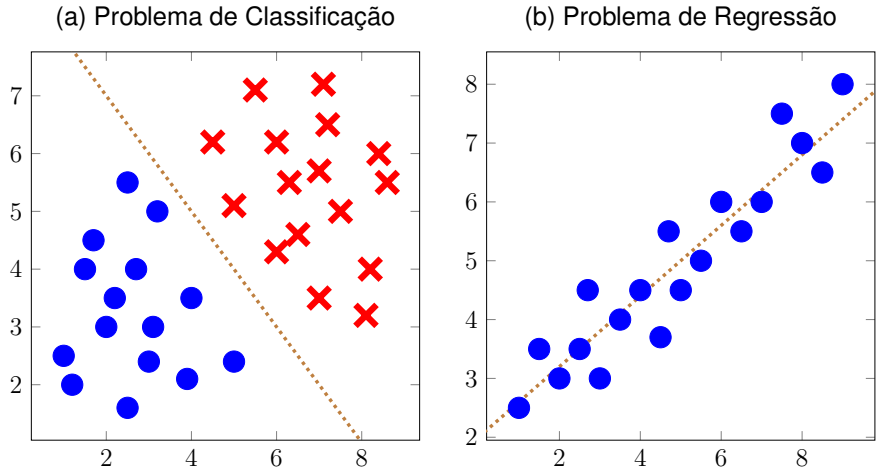
A aprendizagem supervisionada realiza o treinamento dos algoritmos com dados para os quais suas respostas já sejam conhecidas. Ou seja, dependem sempre da entrada de um padrão de valores e da comparação das respostas do sistema com aquelas consideradas corretas. Conforme o algoritmo é treinado seus padrões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

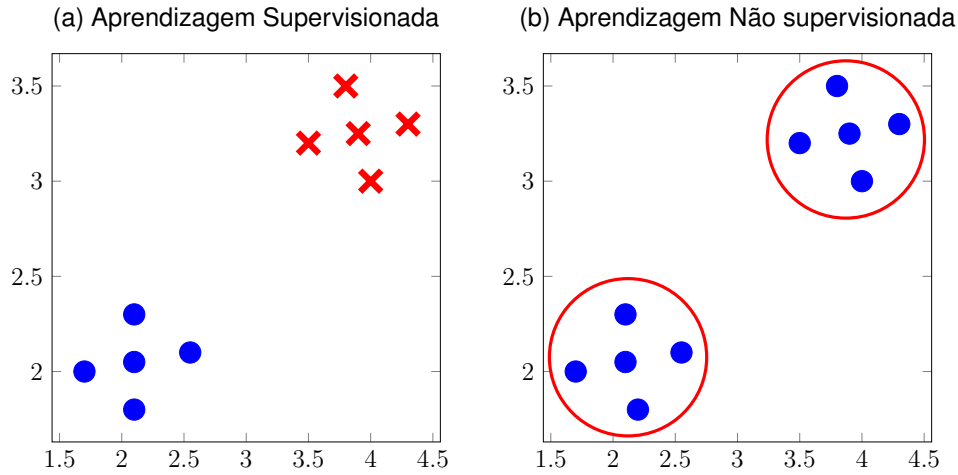
2.1.2 Aprendizado de Máquina Não Supervisionado

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.

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.

$$P(great|popular) = 3/8 = 0.37 \tag{2.3}$$

$$P(popular) = 8/14 = 0.57 \tag{2.4}$$

$$P(great) = 4/14 = 0.28 \tag{2.5}$$

$$P(great|unpopular) = 1/6 = 0.16 \tag{2.6}$$

$$P(unpopular) = 6/14 = 0.42 \tag{2.7}$$

$$P(popular|great) = 0.37 * 0.57/0.28 = 0.75 \quad (2.8)$$

$$P(unpopular|great) = 0.16 * 0.42/0.28 = 0.24 \quad (2.9)$$

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

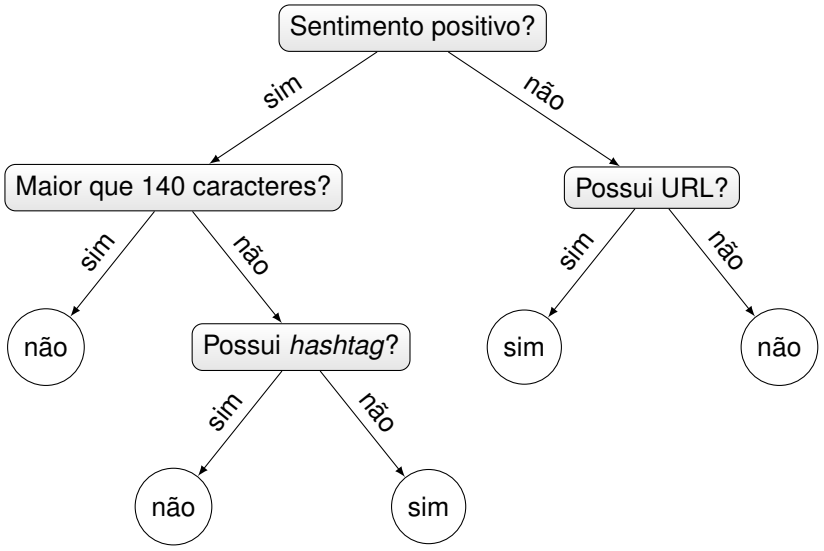
Abstraindo o conceito computacional, uma árvore de decisão pode ser definida por um fluxograma, no qual cada nó, com exceção do último nível, representa um teste sobre as informações disponíveis. O ponto de partida é denominado nó raiz. A medida que os nós filhos vão sendo explorados, as informações são divididas com o objetivo de agrupá-las por similaridade. 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, que é 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 2.3, 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 cada registro de entrada, 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 2.3 – 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). Outros algoritmos serão apresentados e descritos conforme forem utilizados e aplicados na Seção 5, correspondente aos experimentos.

2.3.3 Redes Neurais Recorrentes

Redes Neurais.

3 PROPOSTA

A proposta deste trabalho é a elaboração de um modelo, 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. Esta seção apresenta a arquitetura de coleta e extração utilizada neste 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 também 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

Explicar que os classificadores utilizados para realização do trabalho foram Naive Bayes, árvores de decisão (quais algoritmos) e redes neurais recorrentes, com *word embeddings (word2vec)*.

4 METODOLOGIA

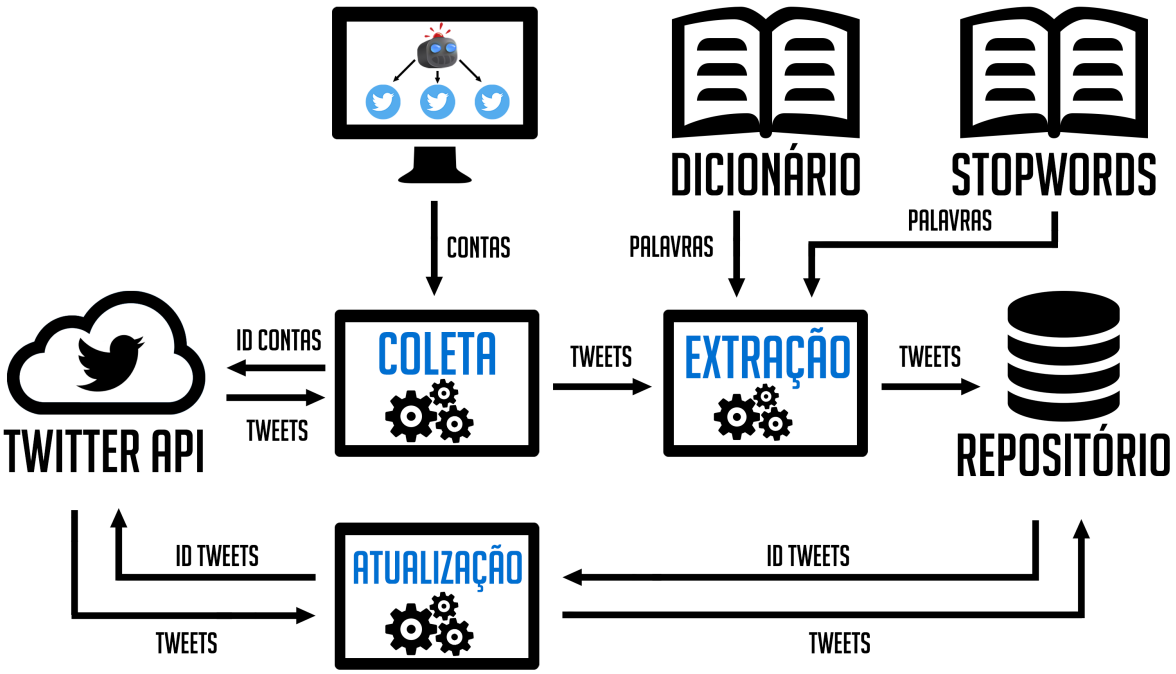
5 EXPERIMENTOS

Com base nos fundamentos e proposta apresentados, respectivamente nas seções 2 e 3, neste capítulo serão apresentados os passos para construção da base de dados, experimentos realizados e resultados obtidos com a aplicação dos algoritmos sugeridos sobre os dados coletados. De maneira geral, os experimentos tem por objetivo realizar análises sobre a aplicação dos algoritmos de aprendizado de máquina correlacionando as características extraídas de tuítes com a sua respectiva taxa de engajamento.

5.1 EXTRAÇÃO DE TUÍTES

Esta seção é responsável por descrever todos os processos envolvidos na coleta dos tuítes que formam a base de dados utilizada nos experimentos. Para realizar todo o processo foi adotada a arquitetura, exibida na Figura 5.1, contendo os seguintes módulos principais: (a) **Coleta** dos tuítes publicados por cada uma das contas acompanhadas; (b) **Extração** das características de cada tuíte; e (c) **Atualização** periódica dos dados coletados.

Figura 5.1 – Arquitetura adotada para extração de tuítes



Fonte: Produção do próprio autor.

A coleta dos tuítes foi realizada tendo como base contas de personalidades influentes que utilizam o Twitter periodicamente. Ao todo, foram consideradas 40 contas de diversas áreas de atuação, como por exemplo Donald J. Trump (atual presidente dos Estados Unidos), Jimmy Fallon (famoso apresentador de TV americano) e Katy Perry (cantora detentora da conta com o maior número de seguidores no Twitter).

A escolha deve-se ao fato de que a análise do impacto de publicações em redes sociais é mais relevante para esse tipo de usuário, uma vez o cálculo da taxa de engajamento para contas com poucos seguidores resultaria sempre em um valor próximo a zero. O que também é reforçado por (SUH et al., 2010), quanto maior a audiência, maiores são as chances de um tuíte ser retuitado. A coleta dos tuítes foi realizada durante o ano de 2018, totalizando cerca de 9500 registros distribuídos entre as contas de interesse que foram escritos na língua inglesa.

5.1.1 Coleta de tuítes

O módulo de coleta é responsável por extrair tuítes de usuários específicos. A extração ocorre de forma contínua, usando recursos de *streaming* disponibilizados pela API do Twitter[1]. São coletados todos tuítes publicados a partir do momento que o *streaming* entra em execução.

A especificação das contas a serem seguidas é feita a partir de uma conta raiz, a partir da qual são extraídos os tuítes publicados em inglês por todos usuários seguidos por essa conta. Essa estratégia permite que novas contas sejam adicionadas à lista sem que haja interrupções na execução do algoritmo. O módulo também conta com tratamento de exceções para que a coleta não seja interrompida devido à problemas temporários de acesso aos dados, como indisponibilidade do serviço ou extrapolação do limite de requisições permitido por instante de tempo.

Na Tabela 5.1 podem ser visualizadas as informações extraídas de cada tuíte através da API. O campo “mensagem” é usado para a extração das características. Já os campos “seguidores”, “retuítes” e “curtidas” são utilizados no cálculo para medir a taxa de engajamento de cada tuíte. Por sua vez, os campos “identificação” e “data/hora” são usados pelo módulo de atualização.

Infelizmente a API do Twitter não permite a extração da quantidade de respostas na versão gratuita, apenas na versão para assinantes, impossibilitando a contabilização desse valor na fórmula de engajamento. Desta forma, a Equação 3.2, para o cálculo da taxa de engajamento, apresentada na seção 3.2, foi adaptada para considerar apenas as informações disponíveis, resultando na Equação 5.1.

[1] *Twitter Developer Platform:* `https://dev.twitter.com`

Tabela 5.1 – Dados coletados para cada tuíte

Informação	**Conteúdo**
autor	código e nome da conta que originou o tuíte
seguidores	quantidade de seguidores da conta que originou o tuíte
identificação	código do tuíte (permite a consulta posterior)
mensagem	texto de no máximo 280 caracteres
data e hora	data e hora da publicação do tuíte em seu país de origem
retuítes	quantidade de retuíte que a mensagem recebeu
curtidas	quantidade de vezes que o tuíte foi favoritado

$$E(x) = \frac{curtidas + retuítes}{seguidores} * 100 \tag{5.1}$$

5.1.2 Extração das Características

Esta etapa corresponde a extração das características de cada um dos tuítes coletados. A extração ocorre imediatamente após a coleta. Os itens abaixo mostram como cada característica foi extraída:

Presença de URLs e *hashtags*: O uso desses recursos na mensagem é facilmente detectado pela presença de prefixos específicos no corpo da mensagem. Por exemplo, o prefixo “http” indica que URLs foram usadas, enquanto que o prefixo “#” denota o uso de *hashtags*.

Tamanho da mensagem: O tamanho é extraído através da contagem da quantidade de caracteres presentes no texto. A contagem desconsidera caracteres usados em URLs, assumindo que *hiperlinks* não transmitam nenhuma mensagem. A remoção de URLs foi realizada a partir da aplicação de uma expressão regular.

Extração do sentimento: Para realizar a extração do sentimento, foi utilizada a biblioteca TextBlob da linguagem Python (LORIA et al., 2014). Essa biblioteca permite a obtenção da polaridade e subjetividade de conteúdos textuais na língua inglesa. A API também fornece a possibilidade de tradução do conteúdo de textos escritos em outras linguagens. A extração do sentimento realizada pela biblioteca se baseia em Árvores de Decisão e no modelo de classificação *Naive Bayes* – ambos já apresentados na seção 2 –, o que elimina a necessidade de elaborar no novo algoritmo para realizar essa função.

Extração da banalidade: A verificação da frequência utiliza um dicionário contendo 3000 palavras comuns da língua inglesa[2]. Também são removidas as *hashtags*

[2] *3000 most common words in English:* `https://www.ef.com/english-resources/english-vocabulary/top-3000-words/`

e menções a outros usuários, por entender que não se tratam de palavras que podem ser caracterizadas como banais ou não.

Na Tabela 5.2 pode ser visto um exemplo geral de todas as caraterísticas extraídas nesta etapa.

Tabela 5.2 – Dados obtidos na etapa de Extração

Informação	**Conteúdo**
sentimento	valor entre -1 e 1 correspondente a polaridade do texto
URL	valor 1 se houver URL no texto e 0 se não houver
hashtag	valor 1 se houver *hashtag* no texto e 0 se não houver
tamanho	quantidade de caracteres utilizados na mensagem
banalidade	somatório baseado na no uso de palavras frequentes

5.1.3 Atualização dos dados de retuítes e curtidas

Como o módulo de coleta funciona por meio de *streaming*, os tuítes são coletados no instante de sua criação. Nesse momento, a quantidade de retuítes e curtidas recebidos têm o valor zero. Dessa forma, é necessária uma conferência periódica para a obtenção dos dados atualizados.

A atualização é realizada através de um recurso da API do Twitter que obtém informações de um tuíte a partir do seu código de identificação. Para evitar sobrecarga de processamento, apenas os tuítes publicados no intervalo de 15 dias são atualizados. Como os dados de tuítes mais antigos raramente são modificados, a busca para a atualização de cada um deles seria ao mesmo tempo custosa e improdutiva.

5.2 CLASSIFICAÇÃO DA POPULARIDADE DE TUÍTES

Dois parágrafos falando sobre a ideia da classificação no trabalho, para não deixar em branco a de classificação dos experimentos até o seminário de andamento.

6 CONCLUSÕES

Considerações finais e trabalhos futuros.

REFERÊNCIAS BIBLIOGRÁFICAS

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>.

LORIA, S. et al. Textblob: simplified text processing. **Secondary TextBlob: Simplified Text Processing**, 2014.

OLIVEIRA, L. L. de; MERGEN, S. L. S. Análise da popularidade de tuítes com base em características extraídas de seu conteúdo. **Escola Regional de Banco de Dados (ERBD)**, v. 14, n. 1/2018, 2018. ISSN 2595-413X. Disponível em: <http://portaldeconteudo.sbc.org.br/index.php/erbd/article/view/2834>.

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/>.