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

Lucas Lima de Oliveira

PREDIÇÃO DA POPULARIDADE DE TUÍTES UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Santa Maria, RS
2018

Lucas Lima de Oliveira

PREDIÇÃO DA POPULARIDADE DE TUÍTES UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

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

Lucas Lima de Oliveira

PREDIÇÃO DA POPULARIDADE DE TUÍTES UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

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 6 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

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 disso 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 cada uma dessas pessoas foi fundamental nesta caminhada. A família é para sempre e devemos prezar por ela.

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.

Acrescento agradecimentos especiais também aos meus amigos, já chamados de irmãos, Mateus Berndt e Gregory Fontoura, por toda parceria de sempre, suporte prestado e conselhos dados. A amizade de vocês sempre significou muito e têm sido essencial em todo meu caminho.

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.

Também estendo agradecimentos especiais aos grandes amigos que hoje fazem parte da Confraria Socão na Cara. Essas amizades, criadas durante a graduação, foram fundamentais, proporcionando momentos de descontração e alívio da tensão quando houve sobrecarga de trabalhos.

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

PREDIÇÃO DA POPULARIDADE DE TUÍTES UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

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. Muitas vezes, por parte de empresas e personalidades públicas que utilizam suas imagens para fins monetários, há uma grande preocupação com sua popularidade e o alcance das mensagens. Além de sua relevância, a simplicidade e o volume de dados trafegados pela plataforma diariamente, fazem dela uma fonte de dados muito poderosa. Focando na análise das mensagens veiculadas e no interesse dos usuários em aumentar seu 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 e o próprio texto. Agrega-se também à proposta deste trabalho a realização de análise variando a taxa de engajamento e considerando dados de contas individualizadas. Para alcançar os objetivos destacados, 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

PREDICTION OF TWEETS' POPULARITY USING MACHINE LEARNING ALGORITHMS

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. Many times, in cases of companies and public personalities whose use their images for business purposes, there are a huge concern about popularity and the reach of messages. In addition to its relevance, the simplicity and volume of data sent 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 purpose of this work is the elaboration of models, using algorithms of machine learning, to make predictions and classifications of the tweets' popularity based attributes extracted from the body and the own text. It also adds to the purpose, analysis modifying the engagement rate and considering data from specific users' account. In order to reach this goal, the methodology adopted involves the definition of 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 collection.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface*
$ARFF$	*Attribute-Relation File Format*
$CART$	*Classification and Regression Trees*
IA	Inteligência Artificial
LMT	*Logistic Model Trees*
$LSTM$	*Long Short-Term Memory*
PCA	*Principal Component Analysis*
RNN	*Recurrent Neural Network*
SVM	*Support Vector Machine*
$WEKA$	*Waikato Environment for Knowledge Analysis*

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 aumentarem. Um dos indicadores capazes 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 do 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) e (XU; YANG, 2012), tem como parte de seus objetivos, a análise e identificação de fatores impactantes no conteúdo das mensagens propagadas no Twitter. Outro fator em comum é que esses trabalhos também utilizam técnicas de aprendizado de máquina na realização de suas pesquisas.

Dentre as principais técnicas utilizadas nestes trabalhos estão: Máquina de Vetores de Suporte (SVM, do inglês: *Support Vector Machine*); árvores de decisão; Naive Bayes; e Regressão logística. No caso dos atributos utilizados nestas pesquisas, foram considerados como fatores de relevância: utilização de URLs e *hashtags*; o alcance do autor do tuíte (podendo ser medido por seus seguidores ou métricas mais complexas); sentimento da mensagem; além dos fatores de popularidade relacionados a cada tuíte, como as curtidas e os retuítes.

Ainda que cada um destes trabalhos apresentem contribuições muito significativas no contexto da descoberta de conhecimento através de dados coletados do Twitter com aprendizagem de máquina, as análises e experimentos são realizados de maneira genérica, sendo aplicadas as mesmas regras para todos os tipos de usuários. Além disso, as análises sobre os modelos não medem a qualidade das previsões com base em fatores variáveis de engajamento.

Dentro deste contexto, o objetivo deste trabalho é elaborar modelos, utilizando algoritmos de aprendizado de máquina, para realizar a predição e classificação da popularidade de tuítes. Como entrada para os modelos, são utilizadas características extraídas de seu conteúdo e são consideradas taxas de engajamento variáveis. Para formar a base de dados, também é estabelecido como parte do objetivo monitorar e extrair tuítes de determinadas contas do Twitter que possuam certo grau de influência.

No que se refere a aprendizagem de máquina, são testados os algoritmos, já consolidados, Naive Bayes e árvores de decisão. Escolha que foi tomada com base na popularidade e destaque dos algoritmos na solução de problemas de classificação de dados. Como entrada para estes modelos, são utilizados dados provenientes do pré-processamento dos tuítes coletados, sendo consideradas as seguintes características: tamanho em caracteres; sentimento (que mede a emoção transmitida); banalidade (que mede a relevância da mensagem); presença de *hashtags* e URLs, que também foram utilizados nos trabalhos relacionados, além do próprio texto do tuíte.

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 4 apresentará as análises realizadas a partir dos dados coletados juntamente com a aplicação dos algoritmos de aprendizado de máquina estudados. O capítulo de 5 apresenta as considerações finais acerca 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 2.2, são abordados conceitos e técnicas para realização de engenharia de atributos. Já 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.

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

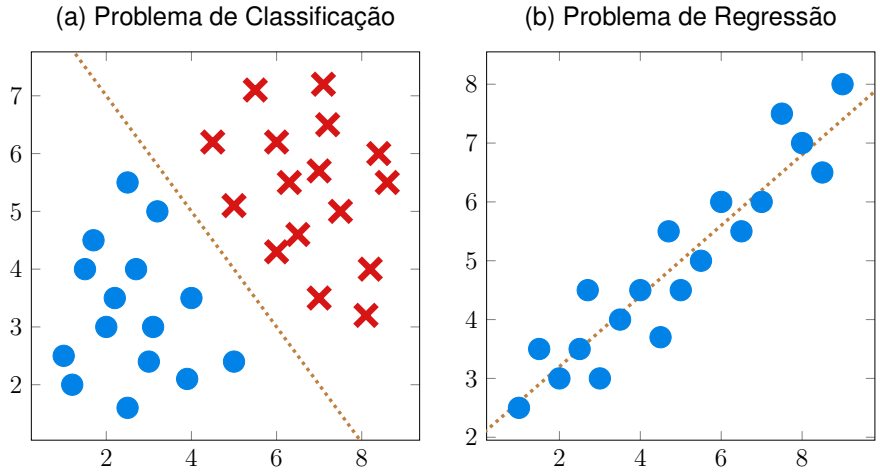
A aprendizagem supervisionada realiza o treinamento dos algoritmos com dados para os quais suas respostas já sejam conhecidas. Ou seja, dependem sempre da entrada de um padrão de valores e da comparação das respostas do sistema com aquelas consideradas corretas. Conforme o algoritmo é treinado seus padrões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

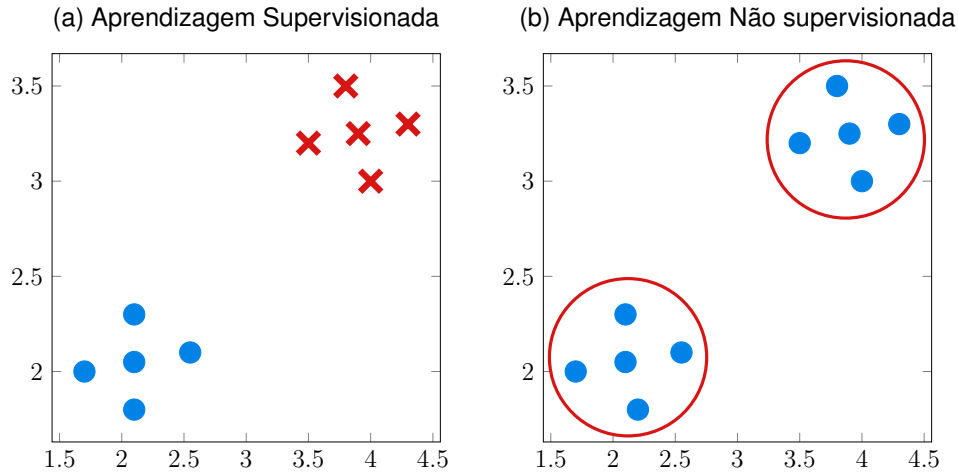
2.1.2 Aprendizado de Máquina Não Supervisionado

No caso dos algoritmos de aprendizagem não supervisionada, ao contrário do segmento apresentado na subseção anterior, eles recebem os dados sem nenhuma 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.

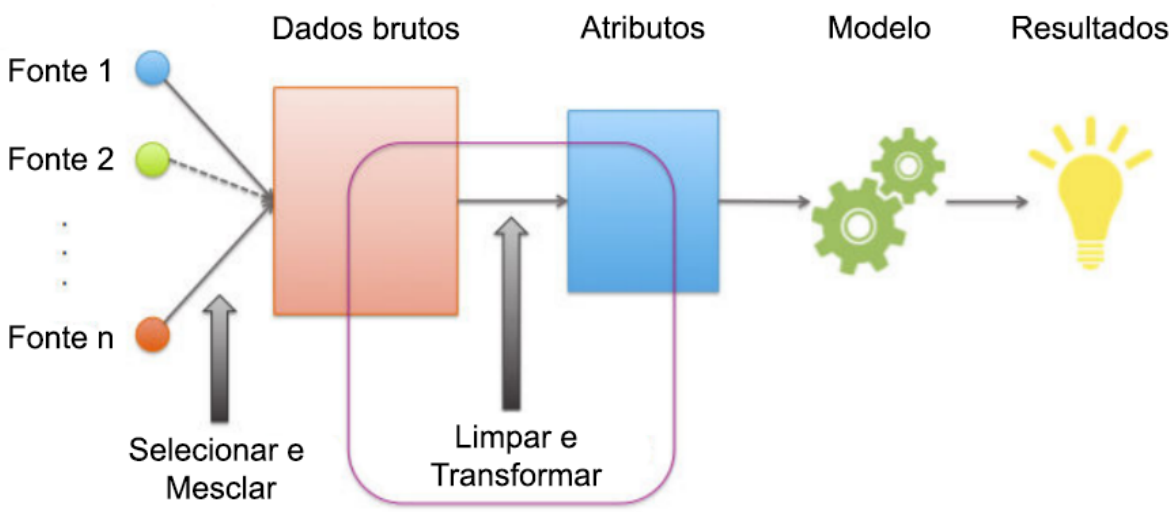
2.2 ENGENHARIA DE ATRIBUTOS

Engenharia de atributos (ou *Feature Engineering*) é um conjunto de técnicas muito utilizadas no processo de aprendizado de máquina. Essas técnicas têm 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 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 têm influência direta nas etapas 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 envolver 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* (MIKOLOV et al., 2013a) e *word-embeddings* (MIKOLOV et al., 2013b) (respectivamente bolsa de palavras e incorporação de palavras), que buscam determinar a frequência e similaridade das palavras presentes nos textos.

Como 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*, 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, utiliza-se aprendizagem de máquina supervisionada, pois os resultados esperados estão diretamente ligados à classificação dos dados. Neste tipo de modelo, para cada instância de treino, é conhecida sua classe, enquanto que na validação, o objetivo é descobrir a classe das instâncias. A justificativa na utilização deste tipo de algoritmo está ligada à tentativa de realizar a predição, uma vez que o alcance de tuítes futuros é desconhecido.

Como mencionado, nesta seção serão abordados alguns dos principais algoritmos que se encaixam no segmento de aprendizagem supervisionada e que serão utilizados no decorrer deste trabalho, sendo eles: Naive Bayes e árvores de decisão. Serão apresentadas suas características, funcionamento, vantagens e desvantagens em suas utilizações. Para exemplificação da aplicação de cada modelo, será utilizada a frase em comum: "*With great power comes great responsibility*".

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 capaci-

dade 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 de 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 tradicional 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 esteja presente na base de treinamento ou não foi classificado para uma das classes, técnicas adicionais devem ser usadas. Uma técnica muito comum aplicada para estes casos é a suavização de Laplace, que soma o valor 1 para todos os valores, desta forma, nenhuma operação é realizada utilizando o valor 0.

Para realizar uma exemplificação, será considerada a frase de exemplo apresentada no início da seção e considerando a Tabela 2.1 – fictícia – apresentada logo abaixo. Na tabela consta a frequência das palavras para cada classe e a probabilidade de cada uma. O objetivo neste exemplo é classificar a palavra "*great*" como popular ou impopular, considerando também que essa possa ser uma frase extraída do Twitter. Neste caso, seriam realizadas as seguintes operações listadas na sequência, após a tabela.

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

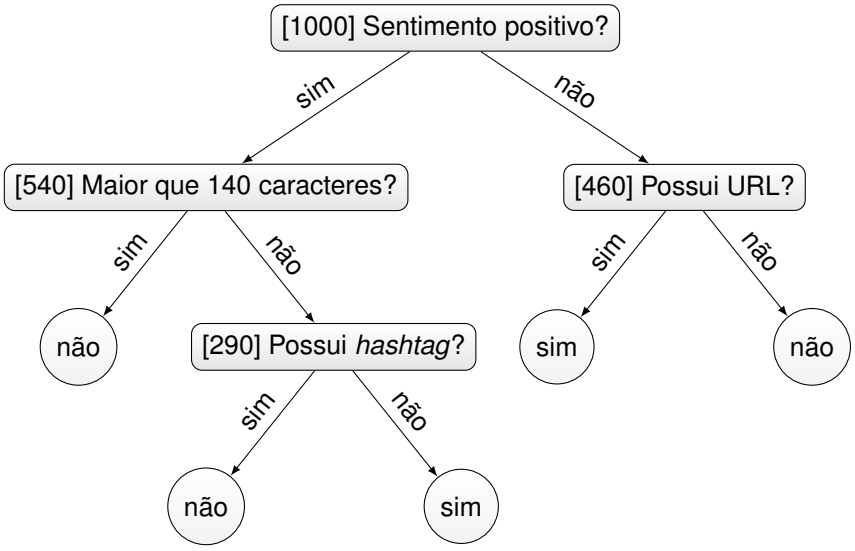
Abstraindo o conceito computacional, uma árvore de decisão pode ser definida por um fluxograma, no qual cada nó, com exceção do último nível, representa um teste

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 buscar 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 destes algoritmo na aprendizagem de máquina está diretamente ligada à sua característica não-paramétrica. Este fator 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, que determina a classe para aquela entrada de dados. Para exemplificar, foi criada a árvore de decisão fictícia apresentada na Figura 2.4, que considera quatro 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 indicado também o número de instâncias que estão naquele nó, partindo da raiz com 1000. Por sua vez, 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 a frase de exemplo apresentada no início da seção, 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*. Mesmo sendo um exemplo fictício e bastante simplificado, é possível perceber que pode haver uma correlação entre os atributos.

Mesmo existindo 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.

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).

Com relação ao à técnica *Naive Bayes*, apresentada na subseção anterior, a maior diferença e, talvez a mais significativa, dentre esses dois tipos de algoritmos, é que árvores de decisão são capazes de lidar com a correlação entre os diferentes atributos utilizados. Enquanto que o algoritmo ingênuo de bayes considera que os atributos são independentes entre si.

3 PROPOSTA

A proposta deste trabalho é a elaboração de modelos, utilizando algoritmos de aprendizado de máquina supervisionado, capazes de classificar o nível de popularidade de tuítes com base na correlação entre a taxa de engajamento dos mesmos em função de um conjunto de características presentes nas mensagens, incluindo o próprio texto. Para atingir esse objetivo, é necessário coletar os tuítes, extrair suas características e aplicar os algoritmos já mencionados para realizar o treinamento e classificação dos dados. A partir disso, esta seção apresenta a definição dos atributos, definição de popularidade, arquitetura de processamento dos tuítes e os classificadores utilizados na realização do trabalho.

3.1 DEFINIÇÃO DOS ATRIBUTOS DE INTERESSE

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 à 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 específica. De maneira semelhante ao uso de URLs, pode atrair o interesse de usuários por determinados tópicos. 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. Esse fator pode estar diretamente ligado à intenção

de cada usuário em propagar mensagens com um determinado humor. Este atributo também pode ser 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, como também em (OLIVEIRA; MERGEN, 2018), 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 um determinado usuário. Uma métrica muito conhecida e utilizada para medir o alcance real de um usuário 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.

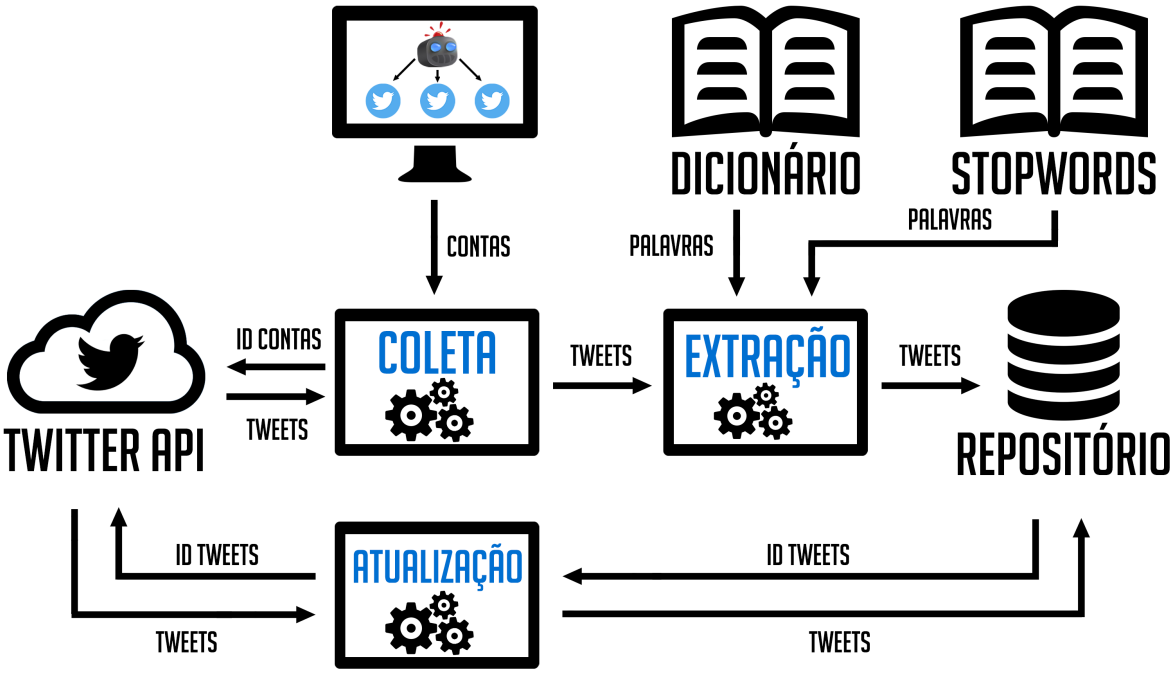
$$E(x) = \frac{curtidas + compartilhamentos + comentários}{seguidores} * 100 \quad (3.2)$$

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 do próprio usuário mais os seguidores dos próprios fãs.

3.3 PROCESSAMENTO DOS TUÍTES

Esta seção engloba a descrição e detalhamento dos processos envolvidos na arquitetura adotada, apresentada na Figura 3.1, para realizar o processamento dos tuítes. Esta arquitetura contém 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 3.1 – Arquitetura adotada para extração de tuítes



Fonte: Produção do próprio autor.

3.3.1 Coleta dos 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 (Twitter, 2018). São coletados todos tuítes publicados a partir do momento que o *streaming* entra em execução.

A especificação das contas a serem seguidas é feita através de uma conta raiz, a partir da qual são extraídos os tuítes publicados por todos usuários seguidos por esta conta. Esta 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 3.1 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.

Tabela 3.1 – 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

Infelizmente a API do Twitter não permite a extração da quantidade de respostas à cada tuíte 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 3.3. Tanto na versão original da fórmula, quanto a adaptação, o valor resultante é um percentual, podendo ser superior a 100%. Esta particularidade é devida ao fato de que o número de iterações que um tuíte obtém pode ser maior que o total de seguidores da conta.

$$E(x) = \frac{curtidas + retuítes}{seguidores} * 100 \tag{3.3}$$

3.3.2 Extração dos Atributos

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[1]. Também são removidas as *hashtags* e menções a outros usuários, por entender que não se tratam de palavras que podem ser caracterizadas como banais ou não.

Na Tabela 3.2 pode ser visto um exemplo geral de todas as caraterísticas extraídas nesta etapa.

Tabela 3.2 – 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

[1] *3000 most common words in English:* `https://www.ef.com/english-resources/english-vocabulary/top-3000-words/`

3.3.3 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.

3.4 CLASSIFICAÇÃO DOS TUÍTES

Esta etapa descreve os processos e ferramentas utilizados na realização da classificação dos tuítes como populares ou não populares. Para a realização da predição dos tuíte tendo como base a taxa de engajamento e considerando os atributos já mencionados, serão utilizados os algoritmos de classificação já apresentados Naive Bayes e Árvores de decisão. Os experimentos com cada algoritmo são realizados através da implementação de código utilizando a linguagem de programação Python (Python Software Foundation, 2018) ou com o auxilio da ferramenta Weka (The University of Waikato, 2018).

A utilização da linguagem Python é justificada pela grande quantidade de bibliotecas que proporcionam maior facilidade em lidar com a manipulação de dados e aprendizado de máquina. Além disso, a linguagem conquistou grande popularidade dentre a comunidade que trabalha com inteligência artificial. Dentre estas bibliotecas, uma delas merece ser destacada aqui, que é a SciKit-Learn (Scikit-Learn Community, 2018), que consiste em um conjunto de funcionalidades específicas para trabalhar com diferentes modelos de aprendizado de máquina, dentre elas classificação, regressão, agrupamento, redução de dimensionalidade e pré-processamento.s

O Weka (sigla em inglês para *Waikato Environment for Knowledge Analysis* e também nome de uma ave da Nova Zelândia) é uma ferramenta desenvolvida na linguagem Java que oferece um ambiente preparado para auxiliar no processo de análise e mineração de dados. Esse ambiente provê uma coleção de algoritmos capazes para realizar tarefas como preparação de dados, classificação, regressão, agrupamento, mineração de regras de associação e visualização (The University of Waikato, 2018).

3.4.1 Balanceamento das instâncias

Como apresentado em (HAN; PEI; KAMBER, 2011), a maioria dos modelos tradicionais de aprendizagem de máquina consideram que os dados de entrada já estão bem distribuídos dentre as classes, o que geralmente não acontece em bases de dados reais. Para resolver este problema, existem diversas técnicas para aperfeiçoar a classificação com dados desbalanceados. Duas destas técnicas são: *oversampling*, que consiste no preenchimento dos dados da classe com menor número até que ambas sejam equivalentes; e o *undersampling*, que consiste na redução dos dados da classe com o maior número (KOTSIANTIS et al., 2006).

Desta forma, para realizar o treinamento e a validação dos dados, independente do modelo de aprendizagem aplicado, é importante que as entradas estejam bem distribuídas entre as classes. O desbalanceamento pode resultar em um modelo tendencioso, havendo a probabilidade de classificar os dados como uma determinada classe devido à predominância da mesma no momento do treinamento.

Além do balanceamento, é preciso preparar a base de dados para cada teste, isto é, configurar os dados de entrada para cada modelo considerando a taxa de engajamento e o usuário autor dos tuítes. Para realizar este processo, foi elaborado uma algoritmo que divide os registros igualmente entre as duas classes considerando uma determinada taxa de engajamento e usuário passados por parâmetro. Este processo também foi escrito na linguagem Python e uma versão simplificada, em pseudo código, pode ser vista logo abaixo no Algoritmo 1.

O algoritmo tem como entrada a taxa de engajamento e o código do usuário, que pode não ser informado, para o caso de análises gerais. Como primeira instrução, é feita uma busca pelos tuítes considerando a taxa e o usuário. A própria função de busca faz o tratamento da necessidade de distinção dos dados por usuário. Tento a lista completa, é identificada a classe que possui o menor número de registros, sendo esta a variável responsável pelo balanceamento. São inicializadas as variáveis de controle (*nPop* e *nNaoPop*), *dados* e *classes*. Em seguida é feito um laço de repetição para cada tuíte. Para cada iteração do laço é feita uma verificação da quantidade de entradas em cada classe através das variáveis de controle e, se o valor for menor que o máximo, os atributos e classe são guardados. Por fim, as variáveis de controle são incrementadas. O laço termina quando a quantidade de dados alcança o valor máximo, retornando então as listas balanceados. Para ilustrar o funcionamento desse algoritmo, foi elaborada a figura 3.2, apresentada logo abaixo.

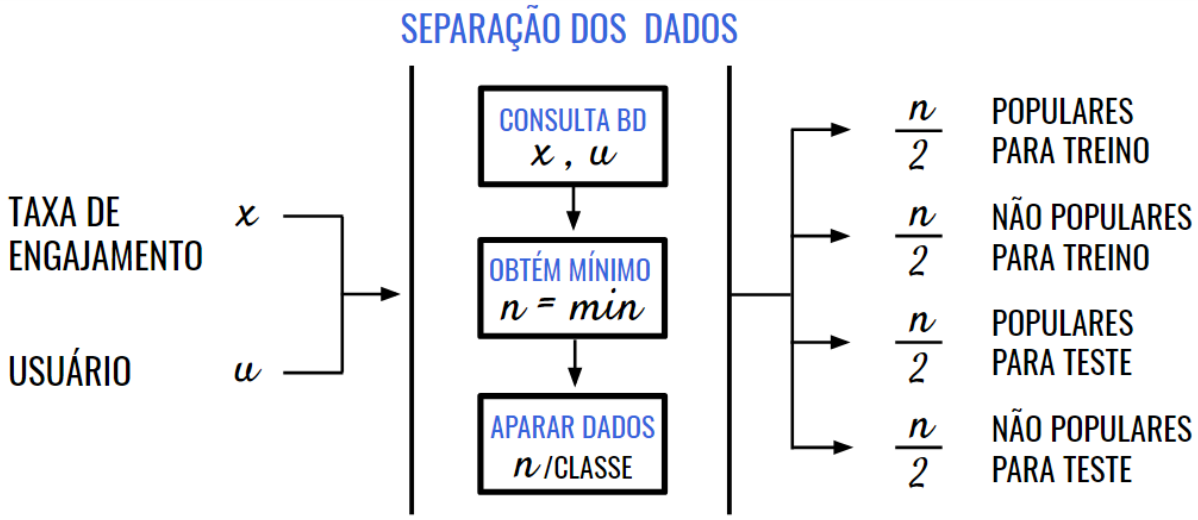
A figura apresentada a entrada de dois parâmetros, *x* e *u*, sendo respectivamente a taxa de engajamento e o usuário. Então o módulo de preparação dos dados faz a consulta, a contagem da quantidade de dados por classe, sendo *n* a quantidade menor e então faz a separação dos dados, devolvendo *2n* registros, distinguindo atri-

Algoritmo 1 Algoritmo para preparação dos dados

```
1:  function PREPARADADOS(taxa, autor = 0)
2:      tuites ← embaralha(buscaListaDeTuites(taxa, autor))
3:
4:      max ← maximoPorClasse(tuites)
5:      nPop, nNaoPop ← 0, 0
6:      dados, classes ← [ ], [ ]
7:
8:      for each tw in tuites do
9:          if (nPop == max E tw['pop']) OU (nNaoPop == max E !tw['pop']) then
10:             pular
11:         dados.novo(tw['atributos'])
12:         classes.novo(tw['pop'])
13:         if tw['popular'] then
14:             nPop ← nPop + 1
15:         else
16:             nNaoPop ← nNaoPop + 1
17:         if quantidade(dados) == max * 2 then
18:             parar
19:     return dados, classes
```

butos de classes. Esta representação, assim como o código descrito no Algoritmo 1, é uma versão simplificada processamento real.

Figura 3.2 – Processo de balanceamento



Fonte: Produção do próprio autor.

3.4.2 Métodos de Classificação Utilizados

A os processos de classificação dos dados foram divididos em dois tipos: utilizando somente o texto pré-processado dos tuítes e utilizando os atributos obtidos de cada tuíte na etapa de extração, descrita na subseção 3.3.2. Os experimentos considerando o texto foram realizados via código em Python, enquanto que os experimentos considerando os atributos foram realizados utilizando a ferramenta Weka. Apesar de realizados em duas vertentes, independente do algoritmo, é realizado o processo de balanceamento descrito anteriormente.

Para a estratégia utilizando o texto pré-processado, é utilizado o algoritmo Naive Bayes, por sua grande popularidade no processamento de conteúdo escrito. O algoritmo para este modelo foi implementado utilizando a linguagem Python, com o auxílio da biblioteca para aprendizado de máquina Scikit-Learn. O algoritmo é baseado em cálculos probabilísticos conforme apresentado na seção 2.3.1 e, assim como código de balanceamento exposto, também recebe como parâmetros de entrada a taxa de engajamento e o usuário autor. Ao final da execução, como retorno são apresentadas as métricas de avaliação do modelo, juntamente com a matriz de confusão. Esse formato de saída foi adotado com o intuito de manter a consistência com os resultados obtidos a partir dos testes utilizando o Weka.

No caso da estratégia utilizando utilizando os atributos coletados, também é utilizado o Naive Bayes, além árvores de decisão, focando nos modelos J48 e LTM (sigla para *Logistic Model Trees*). Para todos estes casos, optou-se pela realização dos testes utilizando o Weka, devido sua praticidade em realizar o treinamento e validação dos dados a partir de arquivos.

A geração destes arquivos de entrada para o Weka é feita de maneira automatizada, a partir de um módulo agregado ao algoritmo de balanceamento, que têm como saída um arquivo do tipo ARFF. Como conteúdo do mesmo, além dos próprios dados, são descritos os atributos utilizados e seus tipos. Cada arquivo, que tem como variante a taxa de engajamento e o usuário, é carregado na ferramenta e aplicado para cada um dos três algoritmos. Com os resultados de cada execução, são gerados gráficos para uma análise visual do desempenho de cada modelo. As análises mais significativas são mostradas na seção 4 a seguir.

4 EXPERIMENTOS

Este capítulo apresenta os experimentos realizados e resultados obtidos com a aplicação dos algoritmos Naive Bayes e árvores de decisão sobre os dados coletados a partir da plataforma do Twitter. De maneira geral, os experimentos têm por objetivo realizar análises, variando o usuário e a taxa de engajamento, para classificar os tuítes como populares ou não. O intuito é identificar uma possível correlação entre estas variantes e as características extraídas de cada tuítes.

4.1 DADOS COLETADOS

Com relação aos processos envolvendo a coleta dos dados, explanados na seção 3.3, eles foram realizados a partir do monitoramento das contas de personalidades influentes que utilizam o Twitter periodicamente. Ao todo, foram consideradas 30 contas de diversas áreas de atuação, como por exemplo Donald J. Trump (atual presidente dos Estados Unidos), Jimmy Fallon (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.

4.2 MÉTRICAS DE AVALIAÇÃO

A fim de realizar a avaliação dos modelos de avaliação, é importante que sejam definidas as métricas utilizadas na comparação entre os modelos. Para isso, serão consideradas cinco métricas: acurácia; sensibilidade; especificidade; valor preditivo positivo; e valor preditivo negativo. Estas medidas, apresentadas respectivamente nas equações 4.1, 4.2, 4.3 e 4.4, retiradas e traduzidas de (HAN; PEI; KAMBER, 2011), são aplicadas após a etapa de validação. Elas consideram os acertos e erros do algoritmo sobre cada uma das classes em relação ao total de cada classe.

A **acurácia** é uma medida de análise geral dos acertos, sem diferenciação entre

as classes. Já a **sensibilidade** e a **especificidade** são as medidas que indicam a capacidade do modelo em realizar a predição das entradas como a classe positiva e negativa, respectivamente. Por sua vez, os valores **preditivos**, expressam a relação do total de predições corretas de uma das classes com o total de predições realizadas para essa mesma classe. As formulas de valor preditivo negativo e positivo, assim como a sensibilidade e especificidade, são equivalentes, variando apenas a classe em questão.

$$Acurácia = \frac{AcertosPositivos + AcertosNegativos}{TotalPositivos + TotalNegativos} \tag{4.1}$$

$$Sensibilidade = \frac{AcertosPositivos}{TotalPositivos} \tag{4.2}$$

$$Especificidade = \frac{AcertosNegativos}{TotalNegativos} \tag{4.3}$$

$$PreditivoPositivo = \frac{AcertosPositivos}{AcertosPositivos + FalsosPositivos} \tag{4.4}$$

$$PreditivoNegativo = \frac{AcertosNegativos}{AcertosNegativos + FalsosNegativos} \tag{4.5}$$

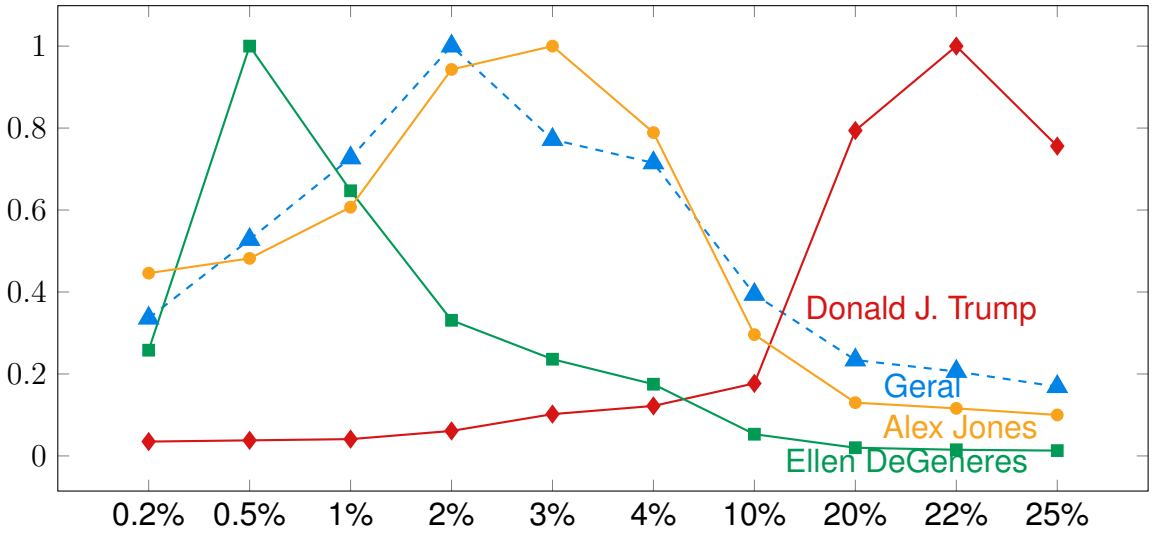
Apesar da importância da utilização de todas as métricas apresentadas, destacam-se duas delas: a acurácia e a sensibilidade. A acurácia por expressar o acerto do modelo de maneira geral, e a sensibilidade por expressar o acerto do modelo sobre a classe de interesse, que é o caso de um tuíte ser positivo. Desta forma, pode-se estabelecer como prioridade a otimização dos resultados obtidos nestas medidas.

4.3 RELAÇÃO ENTRE ENGAJAMENTO E BALANCEAMENTO

No intuito de analisar a relação entre a quantidade de instâncias balanceadas para cada classe e a variação da taxa de engajamento, foi elaborado o gráfico da Figura 4.1. Para melhorar a visualização das demarcações entre as contas, as quantidades foram normalizadas, assumindo como 100% os valores que maximizam a distribuição de dados por classes. No gráfico constam quatro curvas, sendo: todos os registros, em azul, identificada por ”Geral“ e com os marcadores no formato de triângulo; apenas tuítes de ”Donald J. Trump“, em vermelho, com os marcadores em formato de losango; os tuítes de ”Ellen DeGeneres“, em verde, identificada pelos marcadores no formato quadrado; e os tuítes de Alex Jones, em amarelo, identificada pelos marcadores no formato de bola. Estas contas foram utilizadas por estarem entre

aquelas com o maior número de mensagens publicadas.

Figura 4.1 – Número de instâncias balanceadas por taxa de engajamento.



Fonte: Produção do próprio autor.

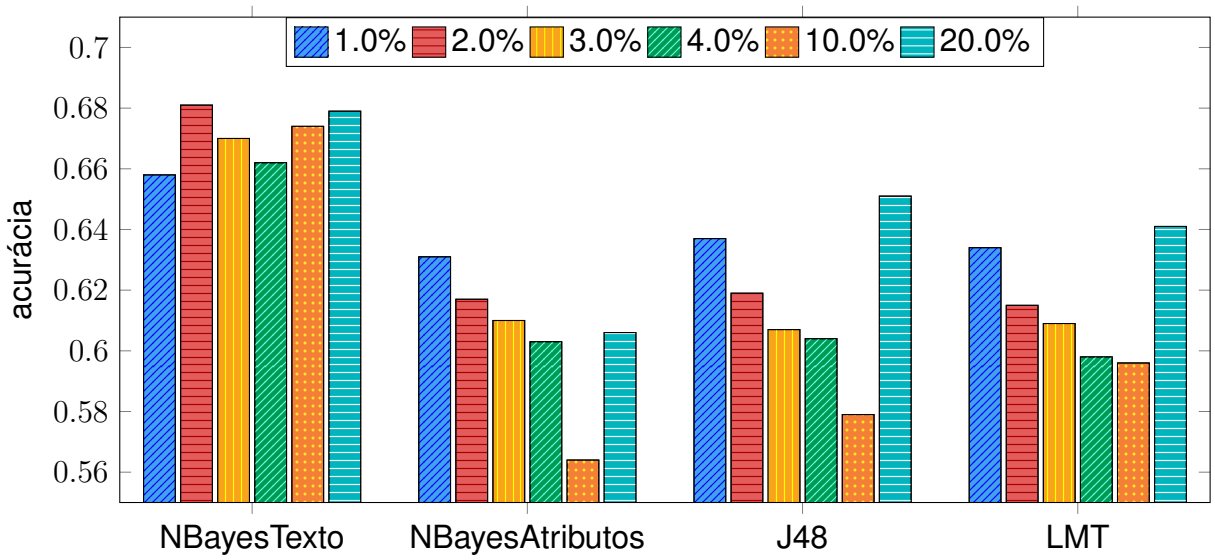
O foco da análise neste gráfico, é a observação dos pontos em que a taxa de engajamento maximiza no número de instâncias balanceadas, isto é, o maior número possível de dados distribuídas entre as classes. No gráfico, isso é expressado pelo eixo *y*, de forma que quanto mais próximo de um, maior é o aproveitamento dos dados de cada conta.

A primeira coisa que pode ser observada, no caso da curva ”Geral“, é que o valor ótimo para maximizar o número de dados balanceados é de 2%, sendo que a variação desta taxa, para qualquer direção, resulta em uma perda de dados considerável, padrão que também se aplica aos demais casos. Pode-se observar também que para cada uma das curvas, o pico é determinado por taxas de engajamento diferentes. O que reforça o propósito da aplicação dos modelos considerando contas individualizadas, já que a taxa de engajamento ideal para um usuário, pode não ser a mesma para os demais.

4.4 ANÁLISE GENERALIZADA DAS CONTAS

Nesta seção, são considerados os tuítes publicados por qualquer uma das contas contidas na base de dados. Os testes foram realizados para medir do desempenho dos diferentes modelos de aprendizagem de máquina supervisionada em função das métricas. O primeiro teste realizado, exposto no gráfico da Figura 4.2, apresenta a variação da acurácia de cada um dos quatro modelos conforme a taxa de engajamento é modificada.

Figura 4.2 – Acurácia de cada algoritmo aplicado sobre toda a base com diferentes taxas de engajamento.



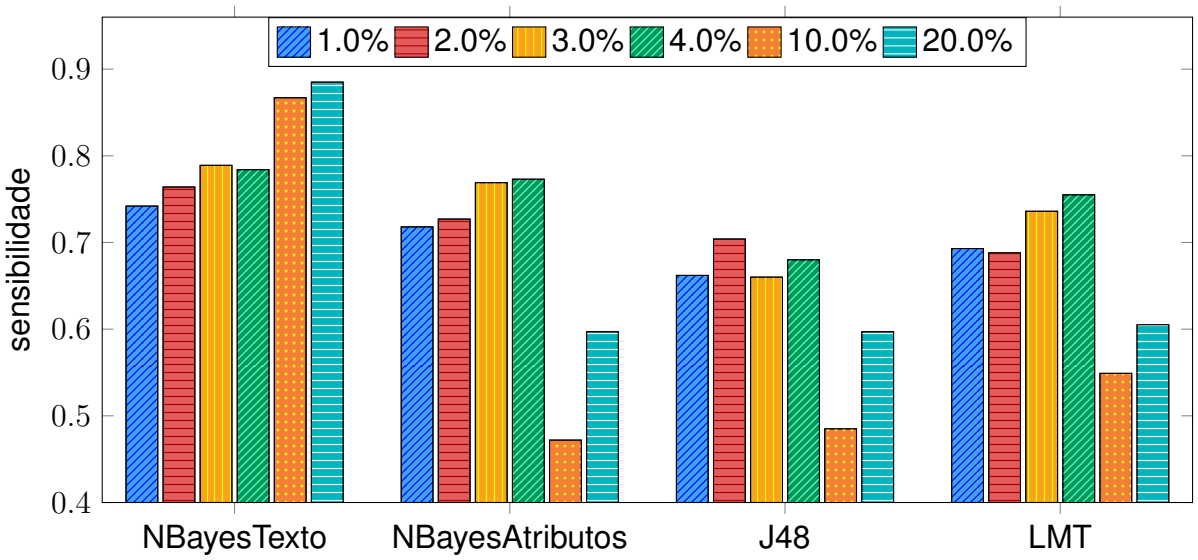
Fonte: Produção do próprio autor.

É possível identificar que o algoritmo Naive Bayes aplicado ao texto apresentou resultados melhores em relação aos demais, que foram aplicados utilizando os atributos extraídos de cada mensagem. O melhor resultado obtido neste modelo foi com a taxa de engajamento de 2%, que alcançou 68.1% de acurácia. Apesar de apresentar um melhor resultado neste experimento, todos os algoritmos obtiveram valores muito próximos, sendo 56.4% o menor deles – caso do Naive Bayes aplicado aos atributos utilizando a taxa de engajamento de 10%.

Ainda realizando a análise geral sobre os dados de todas as contas, foi elaborado o gráfico da Figura 4.3. Neste caso, é exposta a variação no valor da sensibilidade para cada algoritmo, também considerando a modificação na taxa de engajamento. Novamente, apesar de intervalos muito próximos, nota-se que o modelo Naive Bayes aplicado ao texto obteve resultados ligeiramente melhores que os demais modelos, que foram aplicados sobre os atributos. Também é possível observar que o algoritmo aplicado ao texto apresenta resultados que melhoram – dentro do espectro testado – conforme aumenta a taxa de engajamento. Isso mostra que esse modelo pode ter um comportamento mais previsível que os demais.

Com o objetivo de tentar identificar uma possível relação entre a variação no número de instâncias e as métricas de acurácia, sensibilidade e especificidade, foi elaborado o gráfico da Figura 4.4, exibido logo abaixo. Neste caso, foram delineadas as curvas de cada medida apenas para o modelo Naive Bayes aplicado ao texto, além da própria curva do balanceamento das instâncias – também normalizado, assim como no gráfico 4.1.

Figura 4.3 – Sensibilidade de cada algoritmo aplicado sobre toda a base com diferentes taxas de engajamento.



Fonte: Produção do próprio autor.

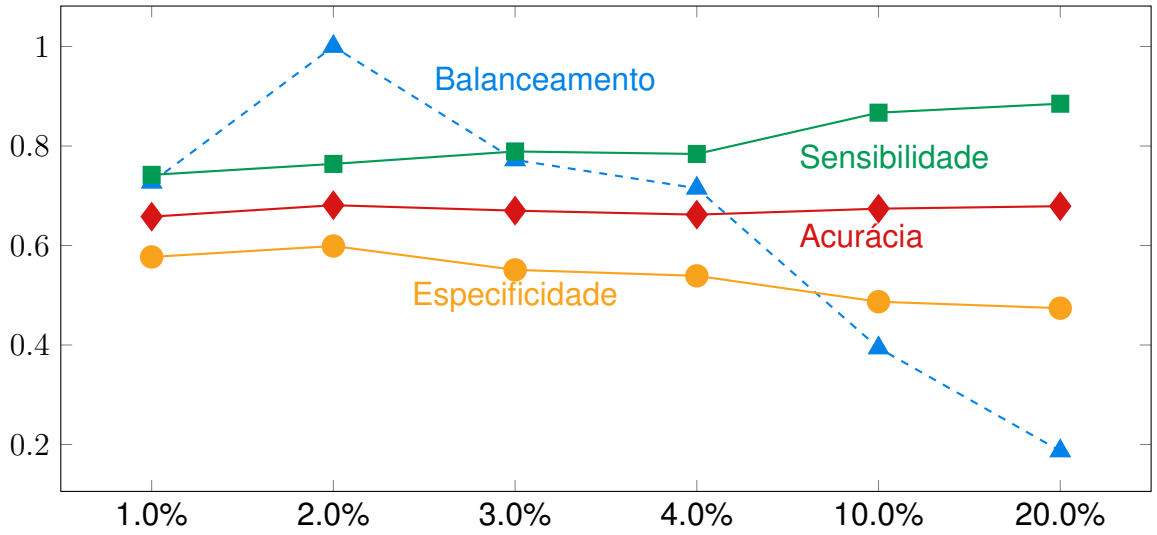
Observa-se inicialmente, que não há uma relação nítida entre as curvas, porém, isso pode ser atribuído ao fato de que a variação no intervalo das métricas é muito pequeno. Apesar disso, considerando também o gráfico 4.2, pode-se notar que a melhor acurácia, também é aquela que maximiza o balanceamento entre as classes. O mesmo padrão não se aplica ao caso da sensibilidade, que aparentemente tem a tendência de aumentar, conforme o número de instâncias é reduzido. Em contrapartida, a curva de especificidade aparentemente tem a melhor relação com o balanceamento das instâncias, já que também tem como pico a taxa de 2% e tende a diminuir na mesma proporção ao variar a taxa para qualquer direção.

4.5 ANÁLISE INDIVIDUALIZADA DAS CONTAS

Para a realização dos processos de treinamento e validação de dados utilizando o aprendizado de máquina, torna-se extremamente importante a utilização de uma quantidade de registros consideravelmente grande. Isso deve-se a dificuldade em fazer estimativas quando se tem poucos dados disponíveis, pois nestes casos, a simples ordenação dos dados pode influenciar na predição final dos algoritmos, já que cada entrada representa uma porcentagem relativamente grande de toda a base. A utilização de poucos dados de entrada faz com que os valores das métricas seja variável para cada execução dos algoritmos.

Devido esse fator, diferentemente dos experimentos realizados na seção anterior, os testes individualizados serão realizados considerando apenas a taxa de enga-

Figura 4.4 – Variação no número de instâncias em relação às métricas para o algoritmo Naive Bayes utilizando o texto.



Fonte: Produção do próprio autor.

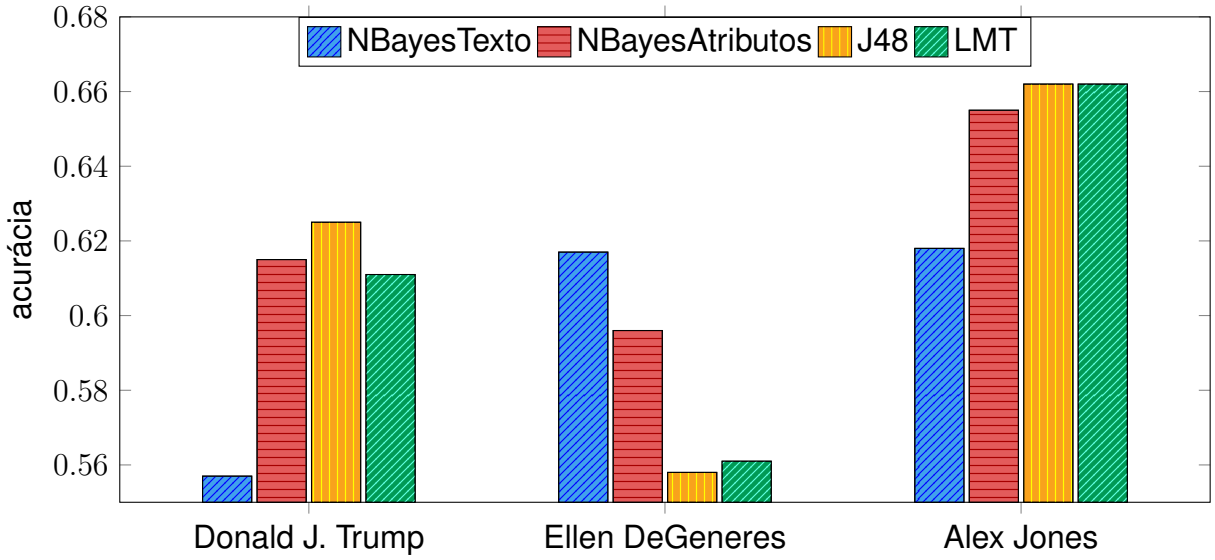
jamento que maximiza a quantidade de registros distribuídos dentre as classes. Caso fossem utilizadas outras taxas de engajamento, ocorria uma grande perda dos registros, devido a necessidade de manter o balanceamento entre as classes, fato que é reforçado pelo gráfico 4.1, o qual apresenta a análise da relação entre a taxa de engamento e o balanceamento das instâncias.

Ainda em função da quantidade de dados, a escolha das contas para realização dos testes nesta etapa foi feita considerando o número de mensagens publicadas por cada uma delas e, assim como na seção 4.3, foram escolhidas as três contas de usuários com o maior número de tuítes capturados, sendo eles: Donald J. Trump, Ellen DeGeneres e Alex Jones. Na sequência são apresentados os gráficos 4.5 e 4.6, que contém os valores obtidos, respectivamente, de acurácia e sensibilidade com cada um dos algoritmos aplicadas sobre os tuítes destas três contas.

Os resultados encontrados nestes casos não foram realmente satisfatórios, conforme o esperado da análise de contas individualmente. De maneira geral, os dados demonstram uma acurácia inferior em relação aos modelos utilizando todas as contas da base. Esse fator pode ser atribuído à quantidade de dados disponíveis para cada conta de usuário. Análises mais completas devem ser realizadas, utilizando uma quantidade consideravelmente maior de tuítes por conta, para assegurar se o mesmo padrão se mantém ou se a acurácia dos modelos é, de fato, inferior quando aplicados sobre contas individuais.

Apesar de insipientes e um pouco inferiores aos resultados gerais, os gráficos mostram que há relevância na realização das análises individualizadas. Pode-se observar que para cada uma das contas comparadas, não há um padrão entre valores obtidos com cada algoritmo e ainda, para cada caso, um modelo diferente obteve a

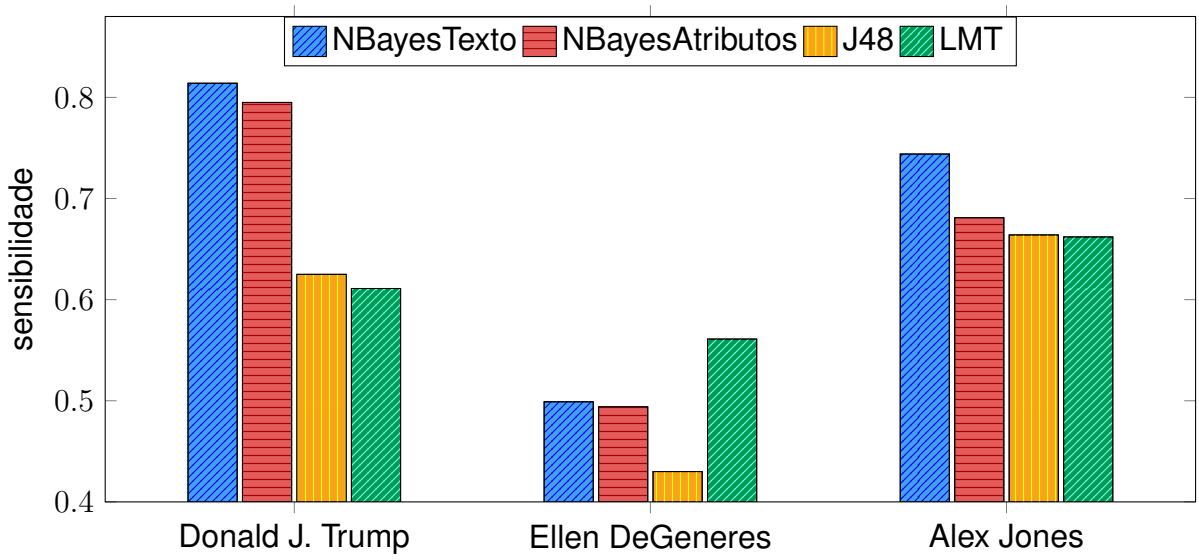
Figura 4.5 – Acurácia dos algoritmos utilizando contas individualizadas.



Fonte: Produção do próprio autor.

melhor acurácia. Isso mostra que cada usuário pode ter suas especificidades e dependendo disso, um determinado algoritmo pode se sobressair em relação aos demais. Enquanto que para a conta de Donald Trump o Naive Bayes aplicado ao texto obteve a pior acurácia, o mesmo modelo aplicado sobre a conta de Ellen DeGeneres, obteve o melhor valor referente à mesma métrica.

Figura 4.6 – Sensibilidade dos algoritmos utilizando contas individualizadas.



Fonte: Produção do próprio autor.

Dentre as contas testas, aquela que obteve resultados mais relevantes foi a de Alex Jones, sendo este, inclusive, o usuário detentor do maior número de tuítes presentes na base. Além da melhor acurácia dentre os demais casos, os valores obtidos

com essa conta tornam-se interessantes pela consistência entre as métricas para os modelos aplicados sobre os atributos coletados. Se comparada a sensibilidade entre esta conta e a de Donald Trump, nitidamente os resultados são melhores para esta segunda. Porém, ao analisar os valores de especificidade, os resultados para Donald Trump se comportam de maneira completamente inversa, sendo este o pior dos casos, enquanto que para Alex Jones os valores se mantém. Isso mostra, no caso da conta do atual presidente dos Estados Unidos, que aqueles modelos que obtêm as métricas de sensibilidade e especificidade de maneira completamente opostas, têm uma tendência maior em classificar os tuítes para uma determinada classe.

Assim como pode ser visto no gráfico 4.1, a conta de Ellen DeGeneres, dentre as três observadas nestes experimentos, obteve a menor taxa de engajamento para poder maximizar a distribuição dos dados entre as classes. Curiosamente, esta também foi a conta com os menores resultados, tanto de acurácia quanto sensibilidade. Este fator leva a crer que ao considerar usuários com uma taxa de engajamento muito baixa, as análises gerais também podem ter seus resultados afetados negativamente. Para identificar o que pode levar a este cenário, seriam necessárias análises mais aprofundadas sobre a relação dos atributos com a popularidade dos tuítes desta conta.

5 CONCLUSÕES

Medir a variação do índice de popularidade pode ser do interesse de administradores de grandes contas do Twitter, pois pode indicar o sucesso ou fracasso de uma determinada campanha realizada ou a dimensão de um escândalo e, tendo consciência disso, ações preventivas ou corretivas podem ser tomadas e sua repercussão pode ser monitorada. Tratando-se de personalidades públicas, é importante identificar quais assuntos e/ou abordagens agradam mais o público-alvo para que assim possam ser mantidas ou evitadas. Assim, pode ser relevante poder prever se uma mensagem tem maior probabilidade de se tornar popular antes mesmo de publicá-la.

Apesar de incipientes e não tão satisfatórios quanto esperado, os resultados alcançados neste trabalho mostram que modelos de aprendizagem de máquina supervisionada podem sim realizar a classificação de tuítes como populares e impopulares. Além disso, pode-se concluir que a variação na taxa de engajamento e análises de contas de usuários específicos são fatores que determinantes, devendo ser levados em consideração na hora de realizar o treinamento e validação dos modelos.

Juntamente com os resultados observados em (OLIVEIRA; MERGEN, 2018), pode-se concluir que as características avaliadas podem sim exercer influência sobre a popularidade de um tuíte em um nível de interesse. Podendo também podem ser levadas em consideração na elaboração de estratégias que visem impulsionar publicações, no caso de empresas e personalidades públicas, como já mencionado.

Como trabalhos futuros, pretende-se realizar análises mais completas sobre as contas individualizadas, incluindo a obtenção de uma quantidade realmente significativa de tuítes para cada um das contas acompanhadas. Desta forma, será possível concluir se aplicação dos modelos sobre cada conta é realmente vantajosa ou não, no que se refere a prever a popularidade dos tuítes. Juntamente isso, pretende-se realizar análises sobre a diferença na influência que cada atributo exerce sobre os modelos quando aplicados à diferentes contas.

Além disso, podem ser elencados como trabalhos futuros a realização de experimentos utilizando outros modelos de aprendizagem de máquina, considerando algoritmos genéticos, redes neurais e aprendizagem profunda. O uso de algoritmos deste tipo é cada vez mais difundido e, se aliado a outras técnicas como *word embedding*, pode vir a trazer resultados muito interessantes. Destacam-se neste contexto as redes neurais convolucionais e as de Memória de Longo Prazo (LSTM, do inglês: *Long Short-Term Memory*).

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>.

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>.

KOTSIANTIS, S. et al. Handling imbalanced datasets: A review. **GESTS International Transactions on Computer Science and Engineering**, v. 30, n. 1, p. 25–36, 2006.

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

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. **CoRR**, abs/1301.3781, 2013. Disponível em: <http://arxiv.org/abs/1301.3781>.

____. Distributed representations of words and phrases and their compositionality. In: BURGES, C. J. C. et al. (Ed.). **Advances in Neural Information Processing Systems 26**. Curran Associates, Inc., 2013. p. 3111–3119. Disponível em: <http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>.

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>.

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>.

Python Software Foundation. **Python 3.7.1 documentation**. Python Software Foundation, 2018. Acessado em nov 2018. Disponível em: <https://docs.python.org/3/>.

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>.

Scikit-Learn Community. **Scikit-Learn: machine learning in Python**. Scikit-Learn Community, 2018. Acessado em nov 2018. Disponível em: <https://scikit-learn.org/stable/>.

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/>.

The University of Waikato. **Weka 3 - Data Mining with Open Source Machine Learning Software in Java**. The University of Waikato, 2018. Acessado em nov 2018. Disponível em: <https://www.cs.waikato.ac.nz/ml/weka/>.

Twitter. **Twitter Developer Platform - Twitter Developers**. Twitter, 2018. Acessado em nov 2018. Disponível em: <https://developer.twitter.com/>.

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.