UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Uma abordagem bilingue na identificação de sentimentos em tweets: Classificação de tweets em sentimentos positivos e negativos

Anna Gabriella Breganholi de Almeida

Trabalho de Conclusão de Curso - MBA em Ciência de Dados (CEMEAI)

ICMC USP SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Uma abordagem bilingue na identificação de sentimentos em tweets

Anna Gabriella Breganholi de Almeida

USP - São Carlos
2020

ANNA GABRIELLA BREGANHOLI DE ALMEIDA

Uma abordagem bilingue na identificação de sentimentos em tweets: Classificação de tweets em sentimentos positivos e negativos

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciência de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Ronaldo Dias

USP - São Carlos

2020

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

B833a

Breganholi de Almeida, Anna Gabriella
Uma abordagem bilingue na identificação de sentimentos em tweets: Classificação de tweets em sentimentos positivos e negativos / Anna Gabriella Breganholi de Almeida; orientador Ronaldo Dias; coorientador Francisco Louzada Neto. -- São Carlos, 2021.
41 p.

Tese (Doutorado - MBA em Ciência de Dados) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2021.

1. Redes sociais. 2. Aprendizado de máquina. 3. Processamento de Linguagem Natural. 4. Algoritmos de Classificação . I. Dias, Ronaldo, orient. II. Louzada Neto, Francisco, coorient. III. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

RESUMO

ALMEIDA, A. G. B. **Uma abordagem bilingue na identificação de sentimentos em tweets**: Classificação de *tweets* em sentimentos positivos e negativos. 2021. 52 f. Trabalho de conclusão de curso (MBA em Ciência de Dados) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2020.

Em 2015, cerca de 300 milhões de pessoas em todo o mundo foram diagnosticadas com depressão e 260 milhões com transtornos de ansiedade. Menos da metade dessas pessoas procuram ajuda médica e apenas cerca de um quarto recebe tratamento adequado. Diante dessa situação, estudos com algoritmos adaptativos começaram a ser realizados buscando identificar padrões psicológicos nos conteúdos publicados nas redes sociais. Com isso, descobriu-se que essas plataformas possuem uma grande quantidade de insumos para a identificação de sentimentos em textos disponibilizados com a escrita coloquial.

Neste estudo, é avaliado como diferentes metodologias de classificação de dados se comportam quando textos (*tweets*) são apresentados para classificar sentimentos em positivos e negativos. Para o estudo foi utilizada uma base em inglês, já classificada, encontrada na plataforma Kaggle, e uma base em português, já classificada, do estudo de "Um dataset para análise de sentimentos na língua portuguesa". Tratando principalmente, mas não exclusivamente, de uma abordagem bilingue (português e inglês). Concluindo que é possível que um algoritmo de aprendizagem seja treinado com uma base de dados que contenha dados em diferentes idiomas, apresentando resultados muito semelhantes a quando treinado com bases contendo textos em um único idioma.

Palavras-chave: Redes sociais. Depressão. Aprendizado de máquina. Linguagem natural, Bilingue.

ABSTRACT

ALMEIDA, A. G. B. **A bilingual approach to identifying feelings in tweets**: Classifying tweets as positive and negative sentiment. 2021. 52 f. Trabalho de conclusão de curso (MBA em Ciência de Dados) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2020.

In 2015, about 300 million people worldwide were diagnosed with depression and 260 million with anxiety disorders. Less than half of people seek help and only about a quarter receive adequate treatment. In view of this situation, studies with adaptive algorithms began to be carried out seeking to identify psychological patterns in the content of social networks. With that, it was discovered that these platforms have many inputs for the identification of sentiment in non-formal texts.

In this study, it is evaluated how different data classification methodologies behave when texts (tweets) are presented to be classified as having a positive or negative sentiment. A English dataset acquired from Kaggle website, and a Portuguese from a previous study called "A dataset for the analysis of feelings in Portuguese" was used for the classification tests. Mainly, but not exclusively, dealing with a bilingual approach (Portuguese and English). Concluding, from several tests, that it is possible for a learning algorithm to be trained with a database that contains data in different languages, presenting results remarkably like when trained with databases containing texts in a single language.

Keywords: Social networks. Depression. Machine learning. Natural language. Bilingual.

SUMÁRIO

1 INTRODUÇÃO

Em 2017 estimou-se que cerca de 264 milhões de pessoas no mundo são diagnosticadas com depressão (transtorno caracterizado por tristeza, perda de interesse ou prazer, sono ou apetite perturbado, cansaço, e dificuldade de concentração). De acordo com a OMS (Organização Mundial de Saúde) a depressão é principal fator contribuinte para as mortes por suicídio, sendo estas cerca de 800.000 a cada ano ("Mental Health - Our World in Data", [s.d.]).

Diversos estudos têm indicado que os médicos inicialmente acionados possuem dificuldades em identificar sintomas relacionados a transtornos psicológicos. Um dos maiores problemas dos métodos existentes é o fato de serem baseados em entrevistas presenciais, tornando-os demorados, e do paciente nem sempre se sentir confortável em falar sobre seus problemas, o que faz com que nem sempre seja possível o diagnóstico correto (HAMILTON, 1960).

No ramo da computação existem algoritmos de inteligência artificial que tem a capacidade de analisar uma grande quantidade de dados, e trazer resultados baseados em uma entrada de treinamento (uma série de dados já classificada, utilizada para a etapa de aprendizado do algoritmo). Utilizando desses algoritmos, alguns estudos foram realizados, identificando que as redes sociais possuem uma grande quantidade de dados que podem ser utilizados como indicadores da saúde mental de seus usuários (CALVO et al., 2017; COPPERSMITH et al., 2015; GO et al., 2009; MOWERY et al., 2017). Isso ocorre pois nestas plataformas, os usuários buscam se comunicar uns com os outros, usando a escrita para demonstrar seus sentimentos, estado mental, esperanças e desejos (MOWERY et al., 2017).

O uso desses algoritmos pode prover tanto análises individuais quanto análises a nível populacional, de forma muito mais rápida que por métodos tradicionais. Além disso, a colaboração do processamento de linguagem natural com a saúde mental possibilita, de maneira escalável, o monitoramento de doenças mentais, possibilitando ainda uma melhor compreensão e acompanhamento dos casos de transtornos psicológicos (MOWERY et al., 2017).

1.1 Objetivo do estudo

O estudo tem como objetivo comparar metodologias de classificação na tarefa de classificar textos em positivos e negativos, com foco na comparação dos resultados da

classificação de bases inteiramente compostas por um único idioma, com bases bilíngues (contendo os dois idiomas estudados – Português e Inglês).

2 REVISÃO BIBLIOGRÁFICA

2.1 Psicologia

2.1.1 Emoções

O significado da palavra emoção é bem discutido entre psicólogos, partindo inicialmente de sua definição no dicionário, a emoção é uma reação, física ou psicológica, que é causada por uma confusão de sentimentos que faz com que o corpo se comporte de alguma maneira (“Emoção - Dicio, Dicionário Online de Português”, [s.d.]).

Para Sartre, uma emoção remete ao que ela significa, que é a totalidade das relações da realidade humana com o mundo. Em sua obra, define a emoção como toda forma de interação e expressão natural de cada ser humano, sendo algo presente em características individuais: expressões faciais, movimentos do corpo (SARTRE, 1939).

Ainda em seus estudos, Sartre define a consciência emocional, que é a consciência do mundo. Para que exista uma emoção, é necessário um evento que desencadeie nela, quando se tem medo, existe um objeto que é o motivo do medo (medo *de* alguma coisa). Sartre nos mostra a emoção como uma forma de entender o mundo (SARTRE, 1939).

De uma perspectiva diferente, a exploração das emoções é feita como no filme Divertidamente (*“Inside out”*). Por meio de personagens que vivem na cabeça da protagonista, este longa metragem nos traz a representação das emoções. A animação conta com uma abordagem extremamente colorida e expressiva, que faz com que identifiquemos as emoções por meio de expressões individuais e cores de cada um dos personagens, remetendo ao sentimento que cada um representa (RODRIGUES; NASCIMENTO, 2019). É possível entender essa representação pois ela está relacionada ao que entendemos por sentimento em um contexto social, o que nos foi ensinado perante ao que é julgado pela sociedade como “correto” (JUNG, 2014).

2.1.2 Sentimentos

Assim como o significado da palavra sentimento é tão complicado de explicar quanto “emoção”. Não é possível encontrar muitas definições sobre o que é o sentimento em sua essência, mas pelo dicionário, é a ação de sentir, de ser sensível (“Sentimento - Dicio, Dicionário Online de Português”, [s.d.]).

Seguindo por uma lógica junguiana, os sentimentos raramente são entendidos de uma maneira correta. Frequentemente, em homens, isso ocorre por uma necessidade se não parecer inferior. Quando somos socialmente conscientizados, perdemos a noção intrínseca do que é entendido por sentimentos, e passa-se a se acreditar em um contexto universal (JUNG, 2014).

2.1.3 Transtornos psicológicos: Depressão

O desenvolvimento de transtornos mentais cresce continuamente com o passar dos anos ("Mental disorders", [s.d.]). Dentre esses, a depressão (transtorno caracterizado por tristeza, perda de interesse ou prazer, sono ou apetite perturbado, cansaço, e dificuldade de concentração) e transtornos relacionados à ansiedade (grupo de transtornos mentais caracterizados por sentimentos de ansiedade e medo) são os mais encontrados, tendo sido estimado, em 2017, respectivamente, 264 milhões e 284 milhões casos no mundo ("Mental Health - Our World in Data", [s.d.]).

A depressão, quando não corretamente tratada, pode levar ao suicídio, aproximadamente 800.000 pessoas tiram suas vidas anualmente, sendo essa a segunda maior causa de mortes de pessoas entre 15 e 29 anos, e é possível dizer que transtornos depressivos estão entre os diagnósticos mais comuns entre as pessoas que cometem suicídio (KEITH, 2013). Ainda que esses números sejam alarmantes, existem barreiras para que pessoas sejam tratadas de forma efetiva, sendo algumas delas: falta de recursos, falta de profissionais qualificados, a vergonha que é associada à saúde mental, e principalmente a não acurácia dos diagnósticos, o que frequentemente leva a pessoas receberem prescrição de antidepressivos sem que realmente tenham depressão ("Depression", [s.d.]).

Aproximadamente 15% das crianças e adolescentes apresentam alguns sintomas que podem ser relacionados à depressão, sendo que 7% das crianças entre 9 e 17 anos apresentam depressão grave, e, 3% apresenta quadro de distimia (transtorno depressivo persistente – caracterizado pela perda de interesse em atividades diárias, desesperança, falta de produtividade, baixa autoestima e um grande sentimento de inadequação) (KEITH, 2013; "Persistent depressive disorder (dysthymia) - Symptoms and causes - Mayo Clinic", [s.d.]).

Além disso, a depressão aguda é a maior causa para comportamentos suicidas em jovens. Mais de 70% das crianças e adolescentes com transtornos depressivos, e outros transtornos que afetam o humor, não recebem o diagnóstico correto, levando-as a não serem submetidas a um tratamento adequado (KEITH, 2013).

2.2 Redes sociais

Uma rede social é, por definição, um *website* ou programa de computador que permite a interação e troca de informações entre pessoas por meio de conexão com a internet, usando um computador ou celular ("SOCIAL NETWORK | meaning in the Cambridge English Dictionary", [s.d.]). Com base na definição, entram na categoria de redes sociais não apenas os aplicativos mais conhecidos, como Facebook, Twitter, Instagram, Reddit, mas também outros que são primariamente utilizados para troca de mensagens, por exemplo, Whatsapp, Facebook Messenger. Da mesma forma a plataforma do YouTube também é considerada como uma rede social.

Hoje temos 3,8 bilhões de pessoas no mundo que utilizam ativamente alguma rede social disponível. Dentre as disponíveis temos na Tabela 1, em ordem decrescente de quantidade de usuários (limitando a tabela aos que serão citados no estudo):

Tabela 1 – Plataformas sociais mais utilizadas no mundo (Janeiro 2020)

Rede social	Quantidade de usuários (em milhões)
FACEBOOK	2.449
INSTAGRAM	1.000
REDDIT	430
TWITTER	340

Fonte: ("Global social media research summary 2020 | Smart Insights", [s.d.]).

Com a ampliação do uso de aparelhos celulares, cada vez mais o acesso a redes sociais é facilitado, chegando a uma quantidade de aproximadamente 99% daqueles que possuem *smartphones* o utilizam para acessar suas redes sociais. Quando estudamos crianças e adolescentes na faixa dos 12 aos 15 anos, essa quantidade é próxima aos 83, e ainda 69% delas possui conta em pelo menos uma rede social ("Global social media research summary 2020 | Smart Insights", [s.d.]).

2.2.1 Padrões em redes sociais

Um dos maiores usos da escrita é a comunicação entre pessoas. Redes sociais são utilizadas tanto para citar fatos, notícias ou acontecimentos, quanto para falar sobre atividades do dia a dia, transmitir sentimentos, estado mental, esperanças e desejos pessoais (CALVO et al., 2017).

Estudos realizados em *tweets* (publicações de até 240 caracteres na plataforma *Twitter*) de pessoas que se identificaram como diagnosticadas com depressão, mostram que existe um grande uso de palavras negativas em seus textos. Frequentemente usuários com esse perfil utilizam um vocabulário mais focado em si, e falam sobre o próprio tratamento psicológico no qual estão submetidos (MOWERY et al., 2017).

2.3 Processamento de linguagem natural

Técnicas de Processamento de Linguagem Natural (PNL, ou *Natural Language Processing – NPL* em inglês) fazem inferências sobre o que as pessoas expressam, o que pode ser interpretada e trazer uma resposta. A análise de textos representa um estudo de dados não-estruturados, que não apresentam uma organização clara, o que torna o estudo mais complexo ("Diferença entre Dados Estruturados e Não Estruturados - Cultura Analítica", [s.d.]). Um dos usos mais comuns de PNL é no marketing, onde diversas empresas analisam e-mails, e publicações em redes sociais, e com base nessas informações passam a oferecer seus produtos de uma forma personalizada (CALVO et al., 2017).

Seu uso nas áreas de vendas já é amplamente conhecido e utilizado, porém esse tipo de análise tem a capacidade de lidar com diversos outros aspectos da sociedade. Estudos indicam a possibilidade de utilizar as técnicas PNL de forma a identificar o sentimento que as mensagens transmitem (CALVO et al., 2017; COPPERSMITH et al., 2015; LIMA et al., 2015; MOWERY et al., 2017).

Frequentemente o estudo de PNL utiliza dicionários léxicos, que consiste em uma lista de palavras que são categorizadas de acordo com alguma metodologia, definida a partir de estudos e análises específicas para cada uso. É possível encontrar diversos dicionários léxicos disponíveis para uso, no estudo foi utilizado o VADER, um dos dicionários disponibilizado gratuitamente em na biblioteca "*nltk.sentiment.vader*" do python. No próximo trecho, está presente uma breve introdução sobre esse dicionário léxico, além de uma adaptação para a língua portuguesa.

2.3.1 *VADER*

O dicionário VADER (*Valence Aware Dictionary for sEntiment Reasoning*) foi criado por uma combinação de métodos quantitativos e qualitativos, com um maior foco para análise de textos em contextos de mídias sociais.

Por ter sido baseado em outros dicionários léxicos já existentes, traz os benefícios de diversos deles, e incrementando-os, considerando acrônimos, gírias, emoticons, intensidade do sentimento, e outros aspectos que também são importantes para análise de sentimentos, o que faz com que funcione de forma mais eficiente em um contexto de mídias sociais. O dicionário contém mais de 11.000 palavras que podem ser encaixadas em uma ou mais das 183 categorias existentes, sendo ele composto de 1.915 palavras categorizadas como positivas e 2.291 negativas (HUTTO et al., 2014).

Foi desenvolvido em python de forma a ser compatível com os modelos de aprendizado de máquina. Tem foco em aumentar as vantagens dos modelos baseados em regras de parcimônia, de forma a construir um sistema de análise de sentimentos que funcione bem em mídias sociais. Além disso, não utiliza dados de treinamento, e sim é construído através de um padrão criado com uma curadoria humana, não sofre perda de velocidade, mesmo com alta performance (HUTTO; GILBERT, 2014).

2.3.2 *LeIA*

LeIA (Léxico para Inferência Adaptada) é uma adaptação do dicionário VADER para a língua portuguesa, com suporte a emojis e focado para análise de sentimentos em mídias sociais. O dicionário está disponibilizado no github (https://github.com/rafjaa/LeIA), mas poucos estudos utilizando a adaptação foram encontrados (ALMEIDA, 2018).

2.4 Estudos semelhantes

Uma pesquisa realizada por Calvo R. *et al.*, tinha como objetivo identificar quais tecnologias de linguagem natural têm sido utilizadas em textos criados por usuários no contexto de saúde mental. Com esse intuito foi identificado que na maioria dos estudos realizado em nível populacional analisa os sentimentos em postagens no Facebook e Twitter, classificando-os em positivos ou negativos. Além disso o estudo apresenta que o humor das pessoas decai no decorrer do dia (CALVO et al., 2017).

Nesse mesmo estudo foi identificado que usuários que sofrem de depressão apresentam alguns sinais como: tendência a publicar tweets tarde da noite, utilização frequente de pronomes na primeira pessoa, raramente seguem outras pessoas e não por novos seguidores (CALVO et al., 2017).

Um outro estudo realizado por Coppersmith G. *et al.* utilizando dicionários léxicos calculava a proporção de *tweets* relacionados a sintomas específicos de depressão, e publicações realizadas no período de meia noite às 4 da manhã. Com isso foi possível demonstrar que existem diferenças entre a linguagem utilizada por pessoas com diagnóstico de depressão, e a utilizada por pessoas que não o possuíam (COPPERSMITH et al., 2015).

No Brasil, é possível encontrar estudos com intuito de identificação de sentimentos em *tweets*, em 2013 alunos da pós-graduação da Universidade Federal do Rio de Janeiro (UFRJ), fizeram uma análise para identificar sentimentos relacionados aos protestos ocorridos entre Junho e Agosto daquele ano, porém uma das conclusões é a falta de bases já classificadas para treino e teste, e isso acarretou, após a classificação utilizada por eles, em uma base desbalanceada, com predominância da classe positiva (FRANÇA; OLIVEIRA, 2014).

2.4.1 Dados para treinamento

Diversas abordagens foram realizadas para o treinamento dos programas, um dos estudos utilizou dados de treinamento provenientes de *tweets* de usuários que haviam declarado que receberam diagnóstico de depressão, distúrbio bipolar, entre outras, em algum momento nos seus *tweets*. Esse estudo realizou um filtro que removia contas com menos de 25 postagens e que não possuíam pelo menos 75% delas em inglês. E para os dados que entrariam na categoria de "não diagnosticado", foi realizada uma busca de forma aleatória, selecionando 10.000 usuários, buscando por postagens das últimas duas semanas, seguindo a mesma regra de exclusão (COPPERSMITH et al., 2015).

Outra abordagem realizada foi a captura de um conjunto menor de postagens, e cada um deles foi categorizado manualmente com base em uma análise realizada por diversos psicólogos e psiquiatras envolvidos no estudo, neste caso específico os textos foram categorizados como: feliz, sem angústia, pouco angustiado, e muito angustiado (CALVO et al., 2017).

Um método utilizado de uma forma mais simplista, utilizou *emoticons* para definir se o *tweet* representava um sentimento positivo, ':)', ou negativo, ':('. Nesse estudo, após as publicações serem categorizadas de acordo com o *emoticon* que possuía, estes foram removidos, e só então foram utilizados para o treinamento. Essa abordagem influencia o classificador a aprender a utilizar de outros atributos do *tweet* para realizar sua categorização (GO et al., 2009).

Os conjuntos de dados utilizados para o estudo estão descritos em detalhes na metodologia.

3 METODOLOGIA

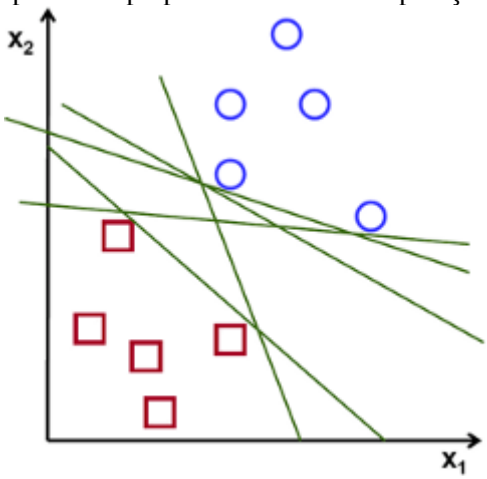
Nesse tópico serão abordadas as metodologias de classificação já existentes, assim como a abordagem que o estudo utilizará.

3.1 Metodologias de aprendizado de máquina

3.1.1 *Support Vector Machines – SVM*

Algoritmos de SVM (Máquina de Vetores de Suporte) têm como objetivo classificar um vetor de informações (números) por meio da criação de um hiperplano que divide o espaço de dados de uma forma binária. O modelo gera um vetor, ou seja, uma linha no plano de dados, que divide as o espaço amostral nas classes (JAMES et al., 2013; YANG et al., 2015), como pode ser visto na figura 1 (R. GANDHI, 2018).

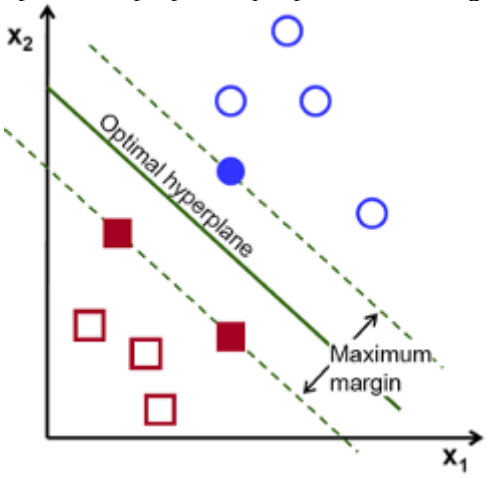
Figura 1 – Exemplos de hiperplanos criados na separação de duas classes



Fonte: Support Vector Machine — Introduction to Machine Learning Algorithms | by Rohith Gandhi | Towards Data Science

Um hiperplano que divide as classes é dado pela identificação de uma reta do formato $f(x) = wx + b$, sendo w um vetor de pesos do tamanho da amostra e b um escalar. O algoritmo de SVM tem como objetivo maximizar a margem entre os dados com classificação diferentes, como pode ser visto na figura 2. Para que isso seja possível o algoritmo utiliza funções de custo, gradiente e perda (R. GANDHI, 2018).

Figura 2 – Exemplos do hiperplano que possui sua margem maximizada



Fonte: Support Vector Machine — Introduction to Machine Learning Algorithms | by Rohith Gandhi | Towards Data Science

3.1.2 *Multilayer Perceptron - MLP*

O *Multilayer Perceptron* (MLP) é um algoritmo de aprendizado supervisionado baseado em redes neurais, que para cada conjunto de treinamento, aprende uma função baseada na quantidade de classes existentes, e quantidade de informações a serem consideradas para gerar a saída.

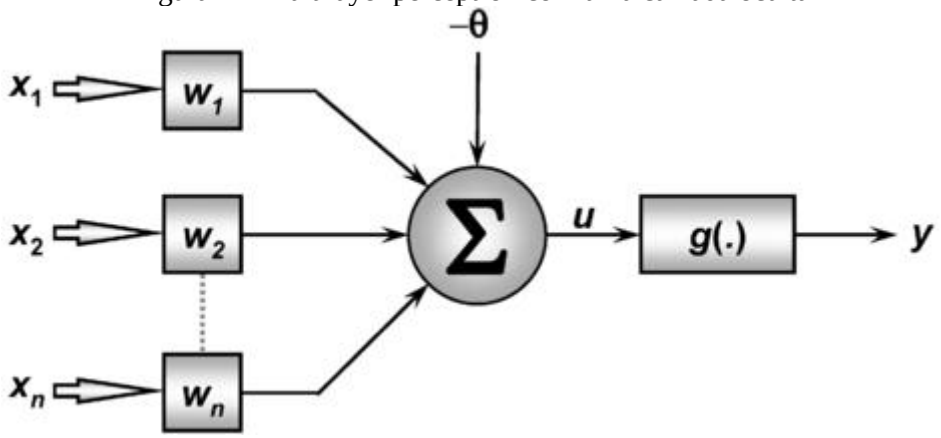
$$f(\cdot): R^m \rightarrow R^o$$

Onde m é o número de dimensões da entrada, o é o número de dimensões da saída, e $f(\cdot) = y$ a saída. Ou seja, dado um conjunto de entradas $X = x_1, x_2, \cdots, x_m$, e um objetivo y, o algoritmo aprende uma função não linear, que funciona tanto em situações de classificação quanto de regressão ("1.17. Neural network models (supervised) — scikit-learn 0.23.2 documentation", [s.d.]). Na figura 4 está representada uma rede *perceptron* com uma única camada oculta.

A primeira camada da esquerda representa um conjunto de neurônios $\{x_i | x_1, x_2, \cdots, x_n\}$, que representa os dados de entrada. Cada neurônio na camada oculta aplica a função linear, adicionando um limiar de ativação θ, resultando na função: $u = \sum_{i=1}^{N} w_i * x_i - \theta$, se o potencial de ativação u for maior ou igual ao limiar de ativação θ, a saída será 1, caso contrário, 0. Seguido por uma função não linear de ativação do tipo $g(\cdot): R \rightarrow R$, sendo esta uma função sigmoide do tipo:$g(u) = 1/(1 + e^{-u})$, sendo u a soma ponderada das entradas e do limiar de ativação. Por fim a última camada oculta recebe as informações da camada anterior e a transforma no valores de saída, a figura 4 demonstra de maneira simplificada o processo de

uma rede MLP (“1.17. Neural network models (supervised) — scikit-learn 0.23.2 documentation”, [s.d.]).

Figura 4 – Multilayer perceptron com uma camada oculta



Fonte: Wikipedia

3.1.3 Árvores aleatórias

Para explicar como os algoritmos de floresta aleatória (*Random Forest*) é importante citar brevemente o funcionamento de Árvores de Decisão. Utilizando a figura 5 é possível imaginar que o conjunto de dados são os números em cima {1, 1, 0, 0, 0, 0}, sendo os três primeiro vermelhos, os três últimos azuis, e os dois primeiros, além de vermelhos, também estão sublinhados (YIU, 2019).

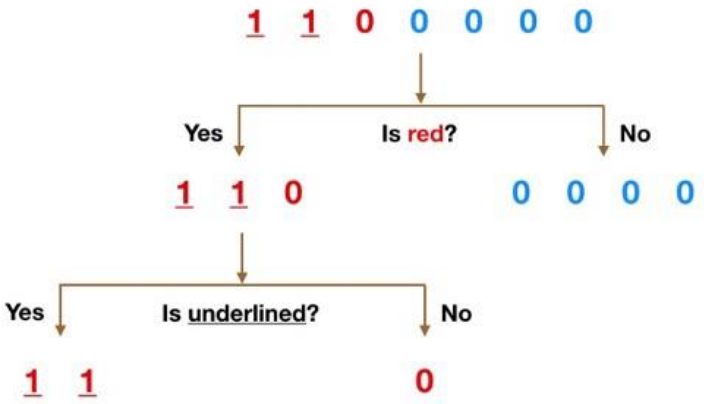
A árvore de decisão da figura possui duas validações:

- É vermelho?
- É sublinhado?

A partir dessas verificações o algoritmo segue realizando as escolhas, usando o primeiro 1 (vermelho, sublinhado), temos que na primeira validação, “é vermelho”, a resposta é “sim”, seguindo para o galho da esquerda. Na segunda validação, “é sublinhado?”, a resposta também é sim, seguindo então para a esquerda, o que pode-se entender como ele pertencente à classe “vermelho, sublinhado” (YIU, 2019).

O algoritmo de florestas aleatórias consiste em várias árvores de decisão geradas de maneira aleatória utilizando-se os dados de entrada. Cada uma dessas árvores constitui um modelo de decisão e são relativamente não-correlacionadas entre si, agindo como um comitê na decisão de classificação da entrada (YIU, 2019).

Figura 5 – Exemplo de árvore de decisão



Fonte: (YIU, 2019)

A chave para esse método trazer bons resultados está no fato de cada árvore trabalhar de maneira individual, e a pouca correlação entre elas faz com que elas estejam protegidas de erros individuais (YIU, 2019).

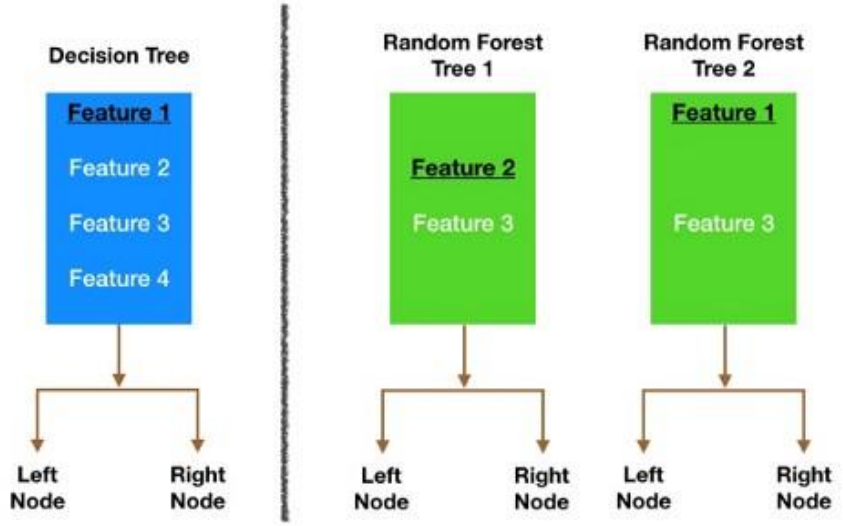
Para a criação das árvores de decisão utiliza-se um dos seguintes métodos:

1. *Bagging (Bootstrap Aggregation)* – Cada árvore é criada com uma amostra aleatória dos dados, com reposição, por exemplo, com uma base de treinamentos sendo [1, 2, 3, 4, 5, 6], é possível que uma das árvores apresente a combinação [1, 2, 3, 3, 6], e outra [1, 2, 4, 5, 6] (YIU, 2019).
2. *Feature Randomness* – Quando é feita uma árvore de decisão normal, todos os elementos e entrada são considerados, por outro lado, em uma floresta aleatória, um subconjunto aleatório das entradas será considerado para cada árvore.

   Como é possível identificar na figura 6, na esquerda temos uma árvore de decisão comum, com 4 elementos de entrada, onde esses elementos serão classificados.

   Do lado direito é apresentado um exemplo de como é feita a divisão por meio de uma floresta aleatória, uma das árvores possui os elementos [1, 3] e outra [2, 3], tendo o elemento 4 não sido utilizado na criação das árvores (YIU, 2019).

Figura 6 – Exemplo de divisão de um nó em uma floresta aleatória



Fonte: (YIU, 2019)

3.1.4 Regressão Logística

O conceito de regressão logística é amplamente utilizado em diversas situações, como por exemplo, um modelo que identifica se um tumor é benigno ou maligno. Para uma decisão ser feita por regressão logística é definido um limiar, um valor que separa a classe A de B, no caso do exemplo anterior, caso o algoritmo resulte em um valor maior que 0,5, o tumor é considerado maligno (SWAMINATHAN, 2018).

Existem 3 tipos de Regressão Logística:

- Binária – Quando a resposta só possui dois resultados, em geral 0 ou 1;
- Multinominal – Quando a resposta possui três ou mais categorias;
- Ordinal – Quando a resposta possui três ou mais categorias que funcionam de forma ordinal, como por exemplo a classificação de um filme.

A margem de decisão para que seja definido em qual classe o dado pertence pode ser definida por uma função linear ou não linear, podendo ainda ser utilizada funções polinomiais para uma definição de margem mais complexa (SWAMINATHAN, 2018).

3.2 API *twitter*

O Twitter possui uma API REST que possibilita baixar uma grande quantidade de *tweets* (publicações dos usuários) públicos de maneira simples. Por meio dessa API é possível capturar informações como quantidade de vezes que a publicação foi visualizada, horário de publicação, a própria publicação, entre outras informações que possam ser relevantes a cada tipo de estudo a ser realizado (“Twitter/analize”, [s.d.]).

3.3 Métricas para avaliação de algoritmos de classificação

A seção a seguir apresenta brevemente os conceitos das métricas utilizadas para a avaliação dos resultados obtidos no estudo.

3.3.1 Matriz de Confusão

A matriz de confusão é uma matriz com os valores reais e preditos pelo classificador, ela apresenta, na diagonal principal, os elementos que o algoritmo classificou corretamente, e nas outras posições, o que foi classificado de maneira errônea.

A tabela a seguir exemplifica uma matriz de confusão genérica, a partir da qual serão baseadas as explicações das métricas seguintes.

Tabela 2 – Matriz de confusão da regressão logística

	Negativo (Predito)	**Positivo (Predito)**
Negativo (Real)	Verdadeiro Positivo (VP)	Falso Positivo (FP)
Positivo (Real)	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Fonte: (NARKHEDE, 2018)

Em problemas desbalanceados é possível obter uma acurácia boa do modelo, porém, quando é observada a matriz de confusão, o classificador pode estar errando 100% da classe minoritária, a seguir é apresentado um exemplo onde as duas matrizes possuem mesmo valor de acurácia, porém no primeiro a classe minoritária está sendo classificada erroneamente.

Tabela 3 – Matriz de confusão: exemplo de classe minoritária classificada incorretamente

	Negativo (Predito)	**Positivo (Predito)**
Negativo (Real)	70	10
Positivo (Real)	20	0

Fonte: (Do Autor)

Tabela 4 – Matriz de confusão: exemplo de classes desbalanceadas

	Negativo (Predito)	**Positivo (Predito)**
Negativo (Real)	55	25
Positivo (Real)	5	15

Fonte: (Do Autor)

No exemplo da tabela 3 a classe positiva foi classificada inteiramente de forma errada, levando a um f1 score para a classe negativa de 0,82, e de 0 para a classe positiva, porém possui uma acurácia de 70%. Por outro lado, na tabela 4, o modelo possui um f1 score de 0,78 para a classe negativa e 0,5 para a positiva, e sua acurácia também é 70%.

Os cálculos para cada uma das métricas citadas será explicado na próxima seção, entretanto, é possível observar que dois modelos podem possuir a mesma acurácia, mas um não ser ideal pois possui 100% de erro para uma classe específica. (NARKHEDE, 2018)

3.3.2 Erro tipo I e tipo II

Uma hipótese (H_0) pode ser verdadeira ou falsa. Devido à essa característica, é possível identificar dois tipos de erros, erro tipo I e erro tipo II.

O erro tipo I ocorre quando H_0 é rejeitada (classificada como falsa), porém deveria ter sido aceita como verdadeira. Do outro lado, o erro tipo II ocorre quando H_0 é aceita (classificada como verdadeira), mas é falsa (NARKHEDE, 2018). A tabela a seguir ilustra o descrito anteriormente.

Tabela 5 – Erro tipo I e erro tipo II

	Aceitar H_0	**Rejeitar H_0**
H_0 verdadeira	Decisão correta	Erro tipo I
H_0 falsa	Erro tipo II	Decisão correta

Fonte:(“Erros cometidos nos testes de hipóteses - Inferência”, [s.d.])

No objeto de estudo desse trabalho existem duas classes, positiva e negativa (4 e 0), identificando o sentimento demonstrado pelo texto do *tweet* a ser classificado, sendo H_0 a presença de sentimento negativo. Logo, no material os erros ficam especificados da seguinte forma:

- Erro tipo I: Identificar como positivo um texto negativo;
- Erro tipo II: Identificar como negativo um texto positivo.

3.3.3 Acurácia

A acurácia de um modelo é o quanto de todas as classes existentes foram classificadas corretamente. No tópico anterior, foi citada que a acurácia de dois classificadores pode ser igual, mesmo em uma situação em que um deles erra uma das classes por completo, nas tabelas 3 e 4 o valor obtido para a acurácia foi 70%, sendo esse valor calculado pela seguinte fórmula:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}$$

3.3.4 Revocação

A revocação (*recall*) é a medida utilizada para quando o quão frequentemente o modelo classifica corretamente a classe positiva. É esperado que esse valor seja o mais elevado possível, demonstrando que o modelo frequentemente identifica a classe de maneira correta (NARKHEDE, 2018).

Essa métrica é calculada através da seguinte fórmula:

$$Revocação = \frac{VP}{VP + FN}$$

3.3.5 Precisão

A precisão (*precision*) é a medida que traz informação sobre quanto da classe que o modelo classificou como positiva, são realmente positivas (NARKHEDE, 2018).

Esse valor pode ser obtido através da seguinte fórmula:

$$Precisão = \frac{VP}{VP + FP}$$

3.3.6 F1 score

Idealmente busca-se um equilíbrio entre precisão e revocação nos modelos, o f1 score é uma métrica criada para ser possível indicar a qualidade de um modelo, levando em consideração a precisão e a revocação. Quanto mais próximo de 1 é o f1 score, melhor é a performance do modelo (NARKHEDE, 2018).

Essa métrica pode ser obtida através da seguinte fórmula:

$$F1\_score = \frac{2 * Precisão * Revocação}{Precisão + Revocação}$$

4 METODOLOGIA APLICADA NO ESTUDO

Nesse estudo serão utilizados dados de postagens da plataforma Twitter. Em cada *tweet* será avaliado a prevalência de sentimentos positivos e negativos, utilizando-se das técnicas de NPL existentes, focando, principalmente em uma abordagem bilingue (português e inglês). Esse resultado será combinado com as abordagens de aprendizado de máquina que melhor se enquadram no problema.

4.1 Descrição da base de dados

Para os testes iniciais foi utilizada a base disponível no site Kaggle ("Sentiment140 dataset with 1.6 million tweets | Kaggle", [s.d.]). Essa base contém 1,6 milhões de *tweets* em inglês classificados em positivos (4) e negativos (0), na figura 7 é possível visualizar os primeiros registros da base.

A base é composta por 6 colunas, sendo elas:

- "target" – Contém a classificação do conteúdo do *tweet*;
- "ids" – identificador do *tweet*;
- "date" – Data de publicação do *tweet*;
- "flag" – *Query* utilizada, indicador de "*NO_QUERY*" caso não possua uma *query* específica;
- "user" – Usuário que publicou o *tweet*;
- "text" – Texto publicado.

A base contém 50% de seus dados classificados como positivos, e 50% negativos, tornando-a equilibrada para que testes iniciais pudessem ser realizados.

Das informações que a base contém, foram utilizadas apenas as informações de "*target*" e "*text*". No site do *Kaggle* não estão disponibilizadas mais informações sobre o procedimento utilizado para a classificação de cada *tweet*, e da mesma forma não existem informações sobre os critérios utilizados para a busca dos dados por meio da API do *twitter*.

Além da base previamente citada, para esse estudo foi utilizada uma base de *tweets* em português (CAVALCANTE, 2017). A captura dos dados ocorreu por meio da API e os critérios foram os descritos na tabela 6.

Figura 7 – Primeiras 5 linhas da base de dados em inglês

	target	ids	date	flag	user	text
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....

Fonte: (Do Autor)

Tabela 6 – Regras para aceitar tweets e exemplos de tweets aceitos/recusados

Regra	**Exemplo aceito**	**Exemplo recusado**
Deve conter ao menos um dos seguintes Emoticons: “:)”, “:-)”, “:(“, “:-(“	Olá! :)	Olá!
Não pode conter um Emoticons positivos e negativos ao mesmo tempo	Opa! :)	Opa! :) :(
O idioma deve ser o português	Oi! :)	Hi! :)
Não pode ser composto apenas por Emoticons	Tudo bem? :)	:)
Não pode ser composto apenas por links acompanhados de Emoticons	Onde pesquisar: https://bing.com :)	https://bing.com :)
Não pode ser composto apenas por nomes de usuários acompanhados de Emoticons	Olá @pauloemmilio :)	@pauloemmilio :)
Não podem possuir o mesmo texto. Implicando em recusar *retweets* ou tweets com mesmos IDs.	Bom dia! :)	Bom dia! :)

Fonte: (CAVALCANTE, 2017)

A classificação dos *tweets* se deu por meio de identificação de emoticons, aqueles com emoticons ':)' ou ':-)' foram rotulados como positivos e os com emoticons ':(' ou ':-(' foram rotulados como negativos. A base contém 44.593 *tweets*, sendo 23.850 classificados como positivos e 20.743 como negativos, como é possível perceber, a base contém aproximadamente 53,5% de uma classe.

É possível considerar a base um pouco desbalanceada, porém para o estudo, optou-se por trabalhar com ela sem alterações para equilibrar os dados para cada uma das classes.

Finalizando, a terceira base utilizada é uma combinação das duas anteriores, criando um *dataset* com dados em português e inglês. Para sua montagem, foram escolhidos, aleatoriamente, 25.000 *tweets* classificados como positivos e 25.000 como negativos, da base do *Kaggle* (em inglês), e a base inteira em português, chegando a um *dataset* final composto por 94.573 *tweets*, sendo 48.850 positivos e 40.743 negativos.

Figura 8 – Primeiras 5 linhas da base de dados em português

	text	target	ids
0	@caprichOreality Fica assim não miga <3 Tud...	4	858793053972307968
1	Amanhã é dia de dar um trato na palestra para ...	4	858793027871154180
2	@thankovsky @patorebaichado eu também tenho :)...	4	858793021177040896
3	@JoseAbrantes6 Prefiro amar o meu clube nas vi...	4	858793273757970433
4	@Bel_Reedus Recomendamos que vá até uma loja C...	4	858793272017387521

Fonte: (Do Autor)

4.2 Análise dos algoritmos utilizados

Com intuito de identificar o melhor modelo para a classificação dos *tweets*, foram realizados testes com os seguintes algoritmos, configurados para uma classificação binária, seguidos pela configuração utilizando a seguinte modelagem:

- Regressão Logística – *random_state*=0, *solver*='lbfgs', *multi_class*='ovr'
- SVM – configuração padrão da função LinearSVC do pacote *sklearn*
- Árvores aleatórias – *n_estimators*=100, *max_depth*=2 , *random_state*=0
- Redes neurais (MLP) – *solver*='lbfgs', *alpha*=1e-5, *hidden_layer_sizes*=(5, 2), *random_state*=1

Inicialmente, os testes foram realizados utilizando a base de *tweets* apenas em inglês, porém apenas uma parcela de 100.000 registros, sendo metade classificada como positivo e outra metade como negativo, mantendo uma base final equilibrada. Todos os dados foram limpos, tiveram caracteres diferentes de letras, espaços extras e outras informações que apareceriam como ruído na classificação, removidos, além disso, todos os caracteres foram passados para minúsculo.

Em todos os casos foram utilizados 70% dos dados para treinamento e 30% para teste, de forma a apresentar as informações para cada algoritmo da forma mais igualitária possível.

A tabela 7 apresenta os resultados obtidos com o teste.

No teste inicial é possível identificar que o algoritmo de SVM e de Regressão Logística obtiveram uma acurácia superior aos outros métodos, porém outro teste foi realizado, alterando apenas a quantidade de dados disponíveis para treinamento e teste, dessa vez utilizando 150.000 *tweets*. O resultado pode ser observado na tabela 8.

Tabela 7 – Resultados obtidos na análise dos diferentes algoritmos na classificação dos dados (100.000)

Algoritmo	**Matriz de Confusão**	**Acurácia**	**F1 score (classe 0)**	**F1 score (classe 4)**
Regressão Logística	[10986 3923] [3468 11623]	0,75	0,75	0,76
SVM	[10882 4027] [3417 11674]	0,75	0,75	0,76
Árvores aleatórias	[12910 1999] [8810 6281]	0,64	0,7	0,54
MLP	[14909 0] [15091 0]	0,50	0,66	0,00

Fonte: (Do Autor)

Novamente identificamos o algoritmo de SVM e o de Regressão Logística como aqueles que trouxeram uma melhor acurácia do modelo e um maior equilíbrio entre o f1 score referente à classe negativa e da positiva.

A partir dessa informação, foram executadas diversas iterações, com alteração apenas na quantidade de dados disponível para o algoritmo, com intuito de identificar a quantidade ótima de dados para que os modelos apresentem a maior acurácia possível.

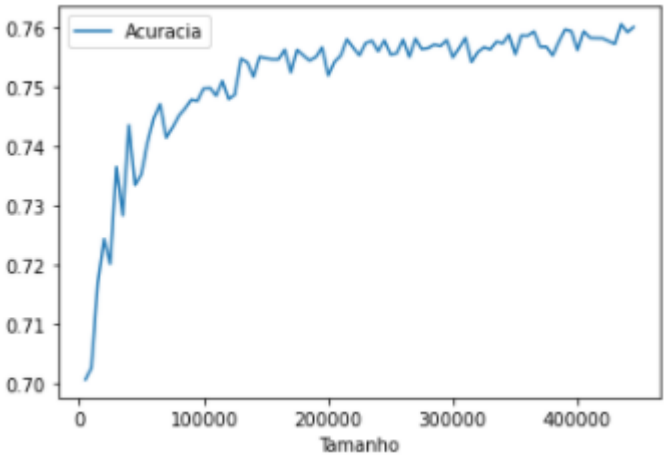
Para o algoritmo SVM, foi possível realizar 89 iterações resultando no gráfico da figura 9, que contém todos os valores de acurácia para cada uma das iterações.

Tabela 8 – Resultados obtidos na análise dos diferentes algoritmos na classificação dos dados (150.000)

Algoritmo	**Matriz de confusão**	**Acurácia**	**F1 score (classe 0)**	**F1 score (classe 4)**
Regressão Logística	[16481 5837] [5136 17546]	0,76	0,75	0,76
SVM	[16358 5960] [5006 17676]	0,76	0,75	0,76
Árvores aleatórias	[19305 3013] [13022 9660]	0,64	0,71	0,55
MLP	[22318 0] [22682 0]	0,50	0,66	0,00

Fonte: (Do Autor)

Figura 9 – Gráfico apresentando a evolução da acurácia do modelo SVM



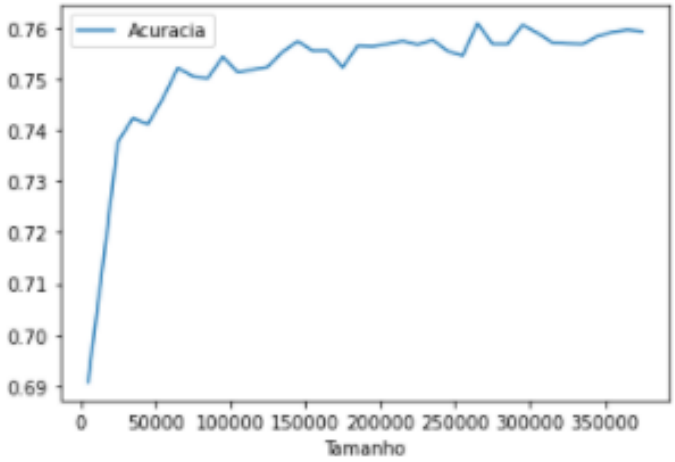
Fonte: (Do Autor)

E para o modelo de Regressão Logística foram realizadas 40 iterações, iniciando com 5.000 dados, incrementando de 10.000 em cada uma delas, resultando no gráfico presente na figura 10.

Visualizando o gráfico gerado pelo resultado da acurácia do algoritmo SVM obtido a cada iteração, é possível identificar uma estabilizada na acurácia a partir de 150.000 dados, mantendo a acurácia do modelo entre 75 e 76%, sendo essa, inicialmente, o melhor resultado obtido nas análises.

De maneira similar, é possível observar no gráfico gerado pelo modelo de regressão logística, que a partir de 100.000 dados, o modelo apresenta uma acurácia de mais de 75%, e aumentando a quantidade de dados para treino e teste não traz muita melhora para a acurácia do modelo.

Figura 10 – Gráfico apresentando a evolução da acurácia do modelo de Regressão Logística



Fonte: (Do Autor)

O segundo conjunto de testes foi realizado com a base em português. Como citado previamente, a base consistiu em 44.593 *tweets*: 23.850 classificados como positivos e 20.743 como negativos.

Todos os dados foram utilizados para treinamento e teste, sem nenhum tratamento prévio com relação ao seu desbalanceamento, porém foi efetuada a limpeza de caracteres que seriam considerados como ruído para o aprendizado do algoritmo.

As modelagens testadas foram as mesmas já descritas, entretanto, para essa análise, utilizou-se 80% dos dados para treinamento, e 20% para teste, obtendo-se os resultados que podem ser observados na tabela 9.

Como é possível observar, para o *dataset* em português, o algoritmo que apresentou melhores resultados foi o de regressão logística, apresentando uma acurácia de aproximadamente 76% com os dados de teste, muito mais elevada que outros métodos, além de apresentar um f1 score relativo à classe negativa de 0,74 e à classe positiva de 0,78. O resultado é bem próximo ao obtido com o algoritmo de SVM, que também é um modelo muito utilizado em problemas binários.

Os resultados obtidos, demonstram que é possível usar lógicas de aprendizado mais simples (Regressão logística e Máquina de Vetores de Suporte) para classificar *tweets* em positivos e negativos. Entretanto, esse estudo possui como foco, a classificação de *tweets* em dois idiomas, português e inglês, consequentemente, *tweets* que contenham ambos os idiomas presentes.

Tabela 9 – Resultados obtidos na análise dos diferentes algoritmos na classificação para os dados em português

Algoritmo	**Matriz de confusão**	**Acurácia**	**F1 score (classe 0)**	**F1 score (classe 4)**
Regressão Logística	[2978 1187] [957 3797]	0,76	0,74	0,78
SVM	[2967 1198] [1004 3750]	0,75	0,73	0,77
Árvores aleatórias	[61 4104] [5 4749]	0,54	0,03	0,70
MLP	[0 4165] [0 4754]	0,53	0,00	0,70

Fonte: (Do Autor)

Para que seja possível uma classificação em ambos os idiomas, foi necessário criar o terceiro *dataset*, que foi constituído de:

- 25.000 *tweets* em inglês classificados como positivos – Escolhidos aleatoriamente da base;
- 25.000 *tweets* em inglês classificados como negativos – Escolhidos aleatoriamente da base;
- 23.850 *tweets* em português classificados como positivos;
- 20.743 *tweets* em português classificados como negativos.

Totalizando em 94.573 *tweets*, sendo 48.850 positivos e 40.743 negativos. A decisão de quantidade levou em consideração a combinação que pudesse deixar o *dataset* mais completo possível, conforme os dados que estavam disponíveis.

Novamente a base foi limpa de caracteres especiais, e os mesmos modelos foram executados, sendo 80% dos dados para treinamento e 20% para teste. Não foram incluídos critérios para que a base de treino e teste possuísse a mesma proporção para as classes e idiomas. Os valores obtidos para a acurácia dos modelos podem ser observados na tabela 10.

Novamente identificamos o modelo de regressão logística como aquele que apresenta a melhor acurácia entre eles, aproximadamente 74%, e f1 score para classe negativa de 0,72 e positiva de 0,76, o que indica um bom índice de revocação e de precisão, demonstrando que é possível incluir dados de idiomas diferentes ao treinar modelos de classificação.

Tabela 10 – Resultados obtidos na análise dos diferentes algoritmos na classificação para os dados em português e inglês

Algoritmo	**Matriz de confusão**	**Acurácia**	**F1 score (classe 0)**	**F1 score (classe 4)**
Regressão Logística	[6421 2720] [2154 7624]	0,74	0,72	0,76
SVM	[6399 2742] [2193 7585]	0,74	0,72	0,75
Árvores aleatórias	[588 8553] [80 9698]	0,54	0,12	0,69
MLP	[0 9141] [0 9778]	0,52	0,00	0,68

Fonte: (Do Autor)

4.3 Algoritmos baseados em dicionários léxicos

Devido à pouca disponibilidade de dicionários léxicos gratuitos, nesse estudo foi utilizado apenas o VADER. Inicialmente, foi utilizada uma base de 100.000 *tweets* positivos e 100.000 negativos, em inglês, utilizando a função "*sid.polarity_scores*", disponibilizada pela biblioteca "*nltk.sentiment.vader*" no Python, classificando como positivo, quando a pontuação obtida for maior ou igual a zero, e negativo quando menor que zero. A partir desses parâmetros foram obtidos os resultados apresentados na tabela 11.

Tabela 11 – Resultados obtidos na análise do dicionário léxico VADER, com 200.000 *tweets*

Algoritmo	**Matriz de confusão**	**Acurácia**	**F1 score (classe 0)**	**F1 score (classe 4)**
VADER (0,0)	[42191 57809] [10072 89928]	0,66	0,55	0,73

Fonte: (Do Autor)

É possível observar que o algoritmo baseado no dicionário léxico frequentemente classifica um texto como positivo, sendo ele negativo (Erro tipo I alto). Para tentar equilibrar esse problema observado, foi realizado um segundo teste, onde a pontuação para considerar se um texto é positivo foi para maior ou igual a 0,1 e em seguida outro com 0,05, obtendo-se os resultados descritos na tabela 12.

Tabela 12 – Resultados obtidos na análise do dicionário léxico VADER

Algoritmo	**Matriz de confusão**	**Acurácia**	**F1 score (classe 0)**	**F1 score (classe 4)**
VADER (0,1)	[70319 29681] [39083 60917]	0,66	0,67	0,64
VADER (0,05)	[68780 31220] [38454 61546]	0,65	0,66	0,64

Fonte: (Do Autor)

Quando alterado o limiar para a classificação do texto entre positivo e negativo, a acurácia se mantém muito próxima, o erro tipo I diminui, mas o erro tipo II aumenta consideravelmente, e o f1 score apresenta um melhor equilíbrio entre as classes.

Após essas observações, é possível identificar que o uso do dicionário para a classificação dos *tweets* se mostra eficiente, apresentando uma taxa de erro de aproximadamente 35%, o que é próximo aos outros modelos apresentados no estudo, porém mais alto que o observados nos algoritmos de classificação como SVM e Regressão Logística.

Para o teste realizado na base em português foi utilizado a adaptação LeIA, trazendo os seguintes resultados:

Tabela 13 – Resultados obtidos na análise da adaptação LeIA na base em português

Algoritmo	**Matriz de confusão**	**Acurácia**	**F1 score (classe 0)**	**F1 score (classe 4)**
LeIA (0,05)	[19437 1306] [1923 21927]	0,93	0,92	0,93

Fonte: (Do Autor)

Em contraste com a versão original em inglês do VADER, a adaptação apresentou resultados muito satisfatórios, apresentando uma acurácia de mais de 90%, possuindo o f1 score também acima de 0,9. O resultado se mostrou melhor que os obtidos a partir de modelos de classificação treinados a partir da base.

No presente momento não foram identificados dicionários léxicos que compreendem os dois idiomas, portanto não foram realizados testes de classificação por dicionários léxicos com a base final (base que contém textos em português e inglês).

4 CONCLUSÃO

Após executados os testes, observou-se que a metodologia de classificação que mais se adequou a base disponível foi a de Regressão Logística principalmente quando ambas as bases foram combinadas, resultando em um *dataset* português-inglês.

Em seguida pode-se dizer que o modelo de SVM se adequou igualmente bem ao problema proposto pelo estudo, pois ambos os algoritmos têm boa performance com classificações binárias, no caso classe positiva e negativa.

Por outro lado, é visível a diferença de performance quando comparados com árvores aleatórias e rede MLP, que tiveram desempenho muito baixo com todos os *dataset* criados para o estudo.

Conclui-se também, que a acurácia de modelos não se perde entre os *datasets* estudados, em todos os casos, foi possível identificar resultados acima de 70% em pelo menos um dos modelos estudados, o que traz a informação que é possível um modelo ser treinado com dados de mais de um idioma, sem que haja uma discriminação (nenhum ponto informando ao modelo qual idioma deverá ser considerado para a classificação do dado), a princípio trabalhando-a de forma que seja equilibrada (contenha aproximadamente a mesma quantidade de dados para treino de cada idioma).

A utilização de base classificada a partir de dicionários léxicos pode trazer mais clareza com relação aos sentimentos expressados no texto. Um bom exemplo é a adaptação do dicionário VADER, LeIA, que alcançou mais de 90% de acurácia na base de dados em português testada nesse estudo.

Por fim, faz-se necessários estudos posteriores, com utilização de outras metodologias para classificação da base, e utilização de redes neurais profundas para a tarefa de identificação de sentimentos. Também seria muito vantajoso realizar estudos em conjunto com profissionais da área da psicologia, sendo possível alinhar outras informações disponibilizadas pela API, como horário de postagem, estações do ano, localização do usuário, o que possibilita uma análise de comportamento individual, na identificação de transtornos psicológicos.

REFERÊNCIAS

1.17. Neural network models (supervised) — scikit-learn 0.23.2 documentation. Disponível em: <https://scikit-learn.org/stable/modules/neural_networks_supervised.html>. Acesso em: 4 out. 2020.

ALMEIDA, R. J. A. **LeIA (Léxico para Inferência Adaptada)**. Disponível em: <https://github.com/rafjaa/LeIA>. Acesso em: 8 jan. 2021.

CALVO, R. A. et al. **Natural language processing in mental health applications using non-clinical texts**. [s.l: s.n.].

CAVALCANTE, P. E. C. Um dataset para análise de sentimentos na língua portuguesa. 2017.

COPPERSMITH, G.; DREDZE, M.; HARMAN, C. **Quantifying Mental Health Signals in Twitter**. [s.l: s.n.].

Depression. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/depression>. Acesso em: 11 jul. 2020.

Diferença entre Dados Estruturados e Não Estruturados - Cultura Analítica. Disponível em: <https://culturaanalitica.com.br/diferenca-entre-dados-estruturados-e-nao-estruturados/#oque-sao-dados-semi-estruturados>. Acesso em: 19 jul. 2020.

Emoção - Dicio, Dicionário Online de Português. Disponível em: <https://www.dicio.com.br/emocao/>. Acesso em: 14 jul. 2020.

Erros cometidos nos testes de hipóteses - Inferência. Disponível em: <http://www.portalaction.com.br/inferencia/511-erros-cometidos-nos-testes-de-hipoteses>. Acesso em: 7 jan. 2021.

FRANÇA, T. C. DE; OLIVEIRA, J. Análise de Sentimento de Tweets Relacionados aos Protestos que ocorreram no Brasil entre Junho e Agosto de 2013. **Proceedings of the III Brazilian Workshop on Social Network Analysis and Mining (BRASNAN)**, p. pág. 128-139, 2014.

Global social media research summary 2020 | Smart Insights. Disponível em: <https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/>. Acesso em: 11 jul. 2020.

GO, A.; BHAYANI, R.; HUANG, L. Twitter Sentiment Classification using Distant Supervision. **Processing**, p. 1–6, 2009.

HAMILTON, M. A RATING SCALE FOR DEPRESSION. **J. Neurol. Neurosurg. Psychiat.,** p. 23:56-62, 1960.

HUTTO, C. J.; GILBERT, E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. **Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014**, p. 216–225, 2014.

JAMES, G. et al. **An Introduction to Statistical Learning: with Applications in R**. [s.l: s.n.]. v. 1

JUNG, C. G. **Sobre sentimentos e a sombra: Sessões de perguntas em Zurique**. [s.l.] Editora Vozes, 2014.

KEITH, L. **Childhood and Adolescent Depression**. [s.l: s.n.].

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

Mental disorders. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/mental-disorders>. Acesso em: 11 jul. 2020.

Mental Health - Our World in Data. Disponível em: <https://ourworldindata.org/mental-health>. Acesso em:

11 jul. 2020.

MOWERY, D. et al. Understanding depressive symptoms and psychosocial stressors on twitter: A corpus-based study. **Journal of Medical Internet Research**, v. 19, n. 2, p. 1–17, 2017.

NARKHEDE, S. **Understanding Confusion Matrix.** Disponível em: <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>. Acesso em: 7 jan. 2021.

Persistent depressive disorder (dysthymia) - Symptoms and causes - Mayo Clinic. Disponível em: <https://www.mayoclinic.org/diseases-conditions/persistent-depressive-disorder/symptoms-causes/syc-20350929>. Acesso em: 11 jul. 2020.

R. GANDHI. **Support Vector Machine — Introduction to Machine Learning Algorithms**. Disponível em: <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>. Acesso em: 4 out. 2020.

RODRIGUES, M. C.; NASCIMENTO, C. S. A antropomorfização cromática da emoção: Análise da longa metragem "Inside Out" da Disney/Pixar. **Revista Psicologia e Educação**, v. 2, n. 2, p. 47–55, 2019.

SARTRE, J.-P. **Esboço para uma teoria das emoções**. [s.l: s.n.].

Sentiment140 dataset with 1.6 million tweets | Kaggle. Disponível em: <https://www.kaggle.com/kazanova/sentiment140>. Acesso em: 19 jul. 2020.

Sentimento - Dicio, Dicionário Online de Português. Disponível em: <https://www.dicio.com.br/sentimento/>. Acesso em: 18 jul. 2020.

SOCIAL NETWORK | meaning in the Cambridge English Dictionary. Disponível em: <https://dictionary.cambridge.org/dictionary/english/social-network>. Acesso em: 11 jul. 2020.

SWAMINATHAN, S. **Logistic Regression — Detailed Overview**. Disponível em: <https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc>. Acesso em: 4 out. 2020.

Twitter/analize. Disponível em: <https://developer.twitter.com/en/use-cases/analyze>. Acesso em: 19 jul. 2020.

YANG, C. et al. Evaluating unsupervised and supervised image classification methods for mapping cotton root rot. **Precision Agriculture**, v. 16, n. 2, p. 201–215, 2015.

YIU, T. **Understanding Random Forest.** Disponível em: <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>. Acesso em: 4 out. 2020.