UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE RUSSAS

CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CÍCERO MARCELO OLIVEIRA MENDES

CLASSIFICAÇÃO E CLUSTERIZAÇÃO DE DADOS DE MÍDIAS SOCIAIS OBTIDOS ATRAVÉS DE WEB CRAWLERS PARA ANÁLISE DE SENTIMENTOS E IDEAÇÕES SUICIDAS

RUSSAS

2022

CÍCERO MARCELO OLIVEIRA MENDES

CLASSIFICAÇÃO E CLUSTERIZAÇÃO DE DADOS DE MÍDIAS SOCIAIS OBTIDOS ATRAVÉS DE WEB CRAWLERS PARA ANÁLISE DE SENTIMENTOS E IDEAÇÕES SUICIDAS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus de Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Orientador: Prof. Dr. Pablo Luiz Braga Soares

RUSSAS

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

M49c Mendes, Cícero Marcelo Oliveira.
Classificação e clusterização de dados de mídias sociais obtidos através de web crawlers para análise de sentimentos e ideações suicidas / Cícero Marcelo Oliveira Mendes. – 2021.
39 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Russas, Curso de Ciência da Computação, Russas, 2021.
Orientação: Prof. Dr. Pablo Luiz Braga Soares.

1. Web Crawlers. 2. Mídias Sociais. 3. Aprendizado de máquina. I. Título.

CDD 005

CÍCERO MARCELO OLIVEIRA MENDES

CLASSIFICAÇÃO E CLUSTERIZAÇÃO DE DADOS DE MÍDIAS SOCIAIS OBTIDOS ATRAVÉS DE WEB CRAWLERS PARA ANÁLISE DE SENTIMENTOS E IDEAÇÕES SUICIDAS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus de Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. Pablo Luiz Braga Soares (Orientador)
Universidade Federal do Ceará - UFC

Prof. Dr. Bonfim Amaro Júnior
Universidade Federal do Ceará - UFC

Profª. Drª. Tatiane Fernandes Figueiredo
Universidade Federal do Ceará - UFC

AGRADECIMENTOS

Primeiramente à Deus, pelo dom da vida, por ter sido meu refúgio em momentos de incertezas, por ter me auxiliado na resolução das dificuldades encontradas em meio ao curso e por ter sempre proporcionado saúde e proteção a mim e às pessoas que amo.

À minha família, em especial minha mãe Maria Goretti de Oliveira, meu pai Antônio Agostinho Mendes e meu irmão Antônio Marcos de Oliveira, que sempre me apoiaram e cuidaram de mim, além de terem se esforçado para que eu pudesse alcançar minhas metas e objetivos. Vocês são a minha base.

Ao meu orientador, Prof. Dr. Pablo Luiz Braga Soares, pela sua paciência, suporte e por todo aprendizado adquirido. Obrigado, de coração, por ter acreditado em mim.

Aos Psicólogos do Campus de Russas que dispuseram de tempo para se reunir, pensar e executar estratégias favoráveis para a realização da classificação manual dos dados que foram utilizados para análise de ideações suicidas nesse trabalho.

A todos os membros do melhor grupo de pesquisa da UFC, o NEMO. Por ter me proporcionado um acolhimento único e poder contar com todos como em uma família, pelo aprendizado alcançado e por cada momento vivido.

Ao Professor Bonfim Amaro Júnior e à Professora Tatiane Fernandes Figueiredo, que aceitaram o convite para fazer parte da banca avaliadora, além das contribuições diretas e indiretas que auxiliaram o desenvolvimento deste trabalho por meio do grupo NEMO.

À todos os Professores que dispuseram de seu tempo em prol da disseminação do conhecimento, o qual rendeu bons frutos e que um dia pretendo contribuir da mesma forma.

Aos amigos fiéis e sinceros que pude fazer com o convívio ao longo do curso, os quais me proporcionaram uma melhor experiência na faculdade e que, com toda certeza, irei levar para a vida.

Por último, mas não menos importante, tal como disse Snoop Dogg, agradeço a mim, por ter acreditado em mim. Agradeço a mim por ter feito todo esse trabalho duro. Agradeço a mim por não ter tido dias livres. Agradeço a mim por nunca desistir.

“Se eu vi mais longe, foi por estar sobre ombros de gigantes.”

(Isaac Newton)

RESUMO

Com a ascensão das mídias sociais como consequência do advento da internet, os usuários dessas plataformas geram grande quantidade de dados que possibilitam sua exploração em diferentes formatos. Há inúmeros meios e técnicas que auxiliam na extração de informações presentes no ambiente web, além de regras para tratamento e aplicação de modelos de aprendizado de máquina com a finalidade de analisar e descobrir padrões em meio aos registros. Pesquisas comprovaram que as postagens de um dado usuário em uma mídia social pode expressar diferentes emoções, ou mesmo, a mensagem pode conter ideações suicidas, sejam estas claras ou não. Trabalhos recentes utilizaram a análise de sentimentos como estratégias para identificação da polaridade de um texto. Outros estudos demonstram o uso da escala de avaliação de risco suicida como ferramenta para a definição do grau suicida contida em um dado conteúdo. Este trabalho tem como objetivo o uso de algoritmos de automação para construção de uma base de dados contendo como fonte as mídias sociais Facebook, Instagram, Twitter e Mundo Psicólogos. Além disso, foca na aplicação de algoritmos da área de aprendizado de máquina para execução de análise de sentimentos e de ideações suicidas. Como resultado desse estudo, foi construída uma base contendo 32.794 mensagens através da implementação de *web crawlers*. Também foram rotuladas 1.600 mensagens de forma manual em 5 diferentes grupos representativos de níveis suicidas, e 31.974 de forma automática em polaridade positiva ou negativa. Foram utilizados quatro algoritmos de aprendizado de máquina, sendo o *Random Forest* com o *Naive Bayes* destinados a classificação sobre a análise de sentimentos e o *K-means* e *DBSCAN* para clusterização sobre ideações suicidas. Dos algoritmos utilizados para análise de sentimentos, o *RandomForest* apresentou os melhores resultados segundo as métricas utilizadas. Sobre a clusterização, nenhuma das classes previamente definidas foi agrupada integralmente a algum dos grupos gerados.

Palavras-chave: Web Crawlers; Mídias Sociais; Aprendizado de máquina;

ABSTRACT

With the rise of social media the advent of the internet, users of these platforms generate a large amount of data that makes it possible to explore them in different formats. Numerous means and techniques help extract information in the web environment and rules for handling and applying machine learning models to analyze and discover patterns in the records. Research has shown that a given user's posts on social media can express different emotions, or even the message can contain suicidal ideation, whether they are clear or not. Recent works have used sentiment analysis as strategies to identify the polarity of a text. Other studies demonstrate the use of the suicidal risk assessment scale as a tool to define the degree of suicide contained in a given content. This work aims to use automation algorithms to build a database containing as source the social media Facebook, Instagram, Twitter and Mundo Psicólogos. In addition, it focuses on the application of machine learning algorithms to analyze feelings and suicidal ideation. As a result of this study, a base containing 32,794 messages was built through the implementation of web crawlers. Also, 1,600 messages were manually labelled in 5 different groups representing suicidal levels, and 31,974 automatically in positive or negative polarity. This paper used four machine learning algorithms: Random Forest with Naive Bayes for classification on the analysis of feelings and K-means and DBSCAN for clustering on suicidal ideation. Of the algorithms used for analyzing feelings, RandomForest showed the best results according to the metrics used. About clustering, none of the previously defined classes was fully grouped with any of the generated groups.

Keywords: Web Crawlers; Social media; Machine learning;

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	*Applications Protocol Interface*
BFS	*Breadth-First-Search*
DBSCAN	*Density-based spatial clustering of applications with noise*
MD5	*Message-Digest Algorithm 5*
ML	*Machine Learning*
NLTK	*Natural Language Toolkit*
Pandas	*Python Data Analysis Library*
SA	*Sentiment Analysis*
UFC	*Universidade Federal do Ceará*
URL	*Uniform Resource Locator*

SUMÁRIO

1 INTRODUÇÃO

Com os avanços tecnológicos voltados à Internet, esta tem se tornado um ambiente alvo para uma série de novas aplicações, entre elas, as mídias sociais. Jones *et al.* (2009) mostra em seus estudos que as mídias sociais, constituídas pelo conteúdo criado e disseminado via interação social, se tornou uma das atividades online mais populares. Dessa forma, utilizar as mídias sociais como meio para o estudo têm se mostrado um caminho favorável em decorrência do grande volume de dados disponibilizados de forma pública pelos usuários em geral.

Facebook, Instagram, Twitter e Mundo Psicologos são exemplos de mídias sociais com as quais usuários comuns interagem de formas específicas, povoando essas plataformas com informações que tangem o entretenimento, em especial sobre as redes sociais Facebook, Instagram e Twitter, além de debates, tira dúvidas e desabafos a respeito de assuntos mais sensíveis ligados ao emocional encontrados em fóruns como o do Mundo Psicologos. Desta maneira, problemas relacionados a coleta, processamento e análise de dados têm ganhado cada vez mais espaço no âmbito acadêmico.

Muitos desses problemas, como o descrito em Burnap *et al.* (2015), tratam de encontrar e organizar sentenças específicas que sejam cruciais para o processo de análise do conteúdo. Como exemplo, considere uma rede de amigos que passam a postar mensagens relacionadas a um determinado evento em suas redes sociais, ou mesmo, uma pessoa depressiva que está pensando em cometer suicídio e utiliza as mídias sociais para compartilhar suas mágoas, anseios e pensamentos, mesmo que de forma indireta. Nota-se necessário a análise das mensagens disponibilizadas por tais usuários a fim de que possam ser identificados os sentimentos do grupo de amigos sobre o evento, ou mesmo, das intenções suicidas compartilhadas pelo usuário que está sofrendo com a depressão. O que utilizar para extrair essas informações tão relevantes desses conteúdos?

Em Rosa (2015) é definida a Análise de Sentimentos, que consiste em determinar a polaridade do sentimento de um determinado conteúdo a fim de que o mesmo possa ser classificado como positivo, negativo ou neutro. Neste trabalho, a autora apresenta implementações de algoritmos distintos que tornam possível a obtenção de resultados satisfatórios, assim como uma comparação com ferramentas já disponibilizadas para tal finalidade.

Em (WHO) *et al.* (2006) é definida a Escala de Avaliação de Risco Suicida, a qual é utilizada comumente por Psicólogos na interação com seus pacientes. Essa escala consiste em classificar as mensagens de um indivíduo suicida em quatro níveis, os quais variam do leve ao

extremo. Nesse trabalho, também é explicado em detalhes cada um dos níveis e a importância que há em identificar mensagens que mascaram ideações suicidas.

É importante mencionar que os trabalhos apresentados na literatura se concentram majoritariamente em tratamentos diretos, não correlacionando análises de forma acumulativa. Dessa forma, este trabalho tem como foco implementar e utilizar *web crawlers* para a coleta de informações em mídias sociais, definir regras e executar tratamentos, além de empregar algoritmos de aprendizado de máquina para realizar classificação e clusterização de mensagens textuais no escopo da análise de sentimentos e de ideações suicidas.

1.1 Organização do Trabalho

Este trabalho está dividido em sete seções. No segundo capítulo são apresentados os objetivos, tanto o geral quanto os específicos. No terceiro capítulo é demonstrada a fundamentação teórica, que explana de forma clara e concisa, conceitos chaves e relevantes para uma melhor compreensão do corrente trabalho. O tema é contextualizado através da apresentação de trabalhos relacionados no quarto capítulo. O quinto capítulo descreve o procedimento metodológico aplicado para a realização da pesquisa. Os resultados e discussões são expostos no sexto capítulo. Por fim, o sétimo capítulo demonstra a conclusão e os trabalhos futuros.

2 OBJETIVOS

2.1 Objetivo geral

Construir uma base de dados com informações presentes nas mídias sociais Facebook, Instagram, Twitter e Mundo Psicólogos a partir da implementação de *web crawlers* e realizar análise de sentimentos e de ideações suicidas a partir de técnicas de classificação e clusterização de dados com algoritmos de aprendizado de máquina.

2.2 Objetivos específicos

- Implementar *web crawlers* para extrair informações das mídias sociais e construir uma base de dados com registros textuais gerados pelos usuários;
- Tratar a base para realização de análise de sentimentos e de ideações suicidas;
- Rotular as mensagens da base de dados;
- Aplicar algoritmos de aprendizado de máquina para realização de análise de sentimentos e de ideações suicidas.

3 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta de forma clara e concisa, conceitos chaves e relevantes para uma melhor compreensão do corrente trabalho. Na seção 3.1 é apresentada a abordagem utilizada por este trabalho na coleta de informações a partir de mídias sociais, *web crawler*. Na seção 3.2 é demonstrada a estratégia de busca utilizada pelos *web crawlers* implementados neste trabalho, conhecida por *Breadth-First-Search* (BFS). Na seção 3.3 é abordado o conceito de mídias sociais, além de apresentados alguns exemplos. Na seção 3.4 é conceituado o aprendizado de máquina, tal como mencionado diferentes tipo e modelos dessa categoria, tendo como exemplo os aplicados nesse trabalho, sendo o *Random Forest*, *Naive Bayes*, *K-means* e *DBSCAN*. Na seção 3.5 é abordado o conceito de análise de sentimentos, o qual é aplicado nesse trabalho por meio de algoritmos de aprendizado de máquina sobre os dados extraídos das mídias sociais.

3.1 Web Crawler

De acordo com Kausar *et al.* (2013), um *web crawler* é um programa de computador que navega pela internet de forma metódica e automatizada. Também conhecido pelos termos Rastreador da Rede, Indexador Automático, Aranha da Rede, Escutador da Rede, ou mesmo, Robô. Um *web crawler* é utilizado comumente para a realização de cópias e extração de conteúdos disponíveis em ambientes web para a consolidação de uma base de dados.

Diante da grande variedade e volume de conteúdo distribuído em diferentes fontes pela internet, é inviável guardar todos esses dados de forma centralizada em um único lugar. Pelo fato que os *web crawler* extraem apenas partes específicas da informação contida em uma página na internet, é importante que essa porção capturada contenha significado real para o propósito desejado. Essa atividade possui um alto grau de dificuldade, pois é necessário que trabalhe corretamente com informações irregulares, além de que em muitas das vezes as páginas processadas não são conhecidas em meio ao processo de extração.

O comportamento dado pelo rastreamento e extração de informações ocorre da mesma maneira que uma pessoa faria se estivesse buscando construir uma base de dados de forma manual com informações presentes em páginas da web. Tal como mostra Pinkerton (2000), o Robô, ao obter um link específico, coleta os dados contidos na página, armazena-os em algum meio digital, obtém os links seguintes e finaliza o procedimento quando alguma condição de parada específica é atendida. Essa dinâmica no comportamento do processo não é exclusiva, mas

é comumento utilizada devido à sua similaridade ao comportamento humano.

3.2 Breadth-first-search (BFS)

O algoritmo *Breadth-first-search* (BFS), ou Busca em Largura, faz parte do conjunto de algoritmos de busca, os quais possuem como característica principal, a sua aplicação sobre grafos. O processo de busca é caracterizado pelo comportamento de exploração dos arcos de um grafo, onde o algoritmo cumpre com a tarefa ao partir de um vértice a outro sobre toda a estrutura. A ordem em que os vértices são visitados define a maneira de organização da busca. Por se caracterizar como um algoritmo de busca não informado, não há necessidade de informação para realização do seu procedimento, bem como, os dados podem ser tratados de forma atômica.

Cormen *et al.* (2009) mostra que o algoritmo de Busca em Largura é um dos mais simples para se realizar a busca em um grafo, além de que é utilizado como base para a implementação de inúmeros outros importantes algoritmos. Dado um grafo $G = (V, E)$ e um vértice s, o algoritmo de Busca em Largura varre as arestas de G, alcançando todos os vértices possíveis. O processo gera uma árvore de busca em largura, onde o vértice s é a raiz contendo todos os demais vértices alcançáveis a partir dele.

Funcionando em grafos dirigidos e em não dirigidos, o algoritmo pode utilizar o conceito de pintura dos vértices, o qual atribui a cada vértice uma cor, tendo como significado o conhecimento dos vértices visitados e não visitados. As três cores utilizadas por Cormen *et al.* (2009) são a branca, cinza e preta, onde a primeira representa os vértices não descobertos e as demais os vértices descobertos. Para definir se um vértice descoberto será cinza ou preto, é necessário analisar seus vértices adjacentes, pois para atribuir a cor cinza ao vértice corrente, se faz necessário que ao menos um dos vértices adjacentes a ele seja branco, caso contrário, o vértice passará a ter a cor preta. A seguir, o algoritmo 1 mostra uma adaptação do pseudocódigo do modelo genérico de um BFS:

Algoritmo 1: Breadth-first-search

```
1  for cada vértice u ∈ V[G] − s do
2      u.cor ← BRANCO;
3      u.d ← ∞ ;
4      u.p ← NIL ;
5  end
6  s.cor ← CINZENTO ;
7  s.d ← 0 ;
8  s.p ← NIL ;
9  Q ← ∅ ;
10 ENQUEUE( Q, s ) ;
11 while Q ≠ ∅ do
12     u ← DEQUEUE( Q ) ;
13     for cada v ← Adj[u] do
14         if v.cor = BRANCO then
15             v.cor ← CINZENTO ;
16             v.d ← u.d + 1 ;
17             v.p ← u ;
18             ENQUEUE( Q, v ) ;
19         end
20     end
21     u.cor ← PRETO ;
22 end
```

3.3 Mídias sociais

De acordo com Barz *et al.* (2020), as mídias sociais são canais on-line que conectam pessoas em todo o mundo ao permitirem a comunicação, relacionamento e compartilhamento de conteúdo entre os usuários. Como apresentado também por Polo e Polo (2021), esses espaços podem conter diferentes motivações para o seu uso, desde a interação social, até o marketing e vendas, sendo principalmente utilizado por seus usuários para a publicação de pensamentos e ideias, através de textos, fotos, vídeos e em outros diferentes formatos que surgem ao longo do

avanço tecnológico.

Em Morandin (2021) são mencionados exemplos de redes sociais, tal como Facebook, Instagram e Twitter, que em suma, são mídias sociais que possuem a característica de entretenimento. Por conter numerosos usuários ativos, a quantidade de conteúdo gerado por meio da utilização dessas plataformas se estende a interesse para estudos no campo de *big data*. Esses exemplos são algumas das redes sociais que foram utilizadas nesse trabalho com a finalidade da construção de uma base de dados.

3.4 Conceitos e algoritmos de Aprendizado de Máquina

De acordo com Monard e Baranauskas (2003), o Aprendizado de Máquina, também conhecido como *Machine Learning* (ML), é um subcampo da Ciência da Computação que tem ganhado um espaço mais acentuado no campo da computação. Como apresentado também por Alpaydin (2009), o crescente número na quantidade de dados disponíveis em todas as áreas, seja ela social ou organizacional, oferecem razões suficientes para justificar a ideia de que a análise inteligente dessas informações deve se tornar ainda mais abrangente, se transformando cada vez mais em uma ferramenta importante para o desenvolvimento tecnológico.

O Aprendizado de Máquina torna possível a identificação de padrões e características que seriam dificilmente percebidas de forma clara por um humano, através da utilização de técnicas diversas de análise de dados. Para Samuel (1959), o Aprendizado de Máquina é o campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados, utilizando da construção de algoritmos que aprendem com seus erros e fazem previsões sobre dados. Esses algoritmos são aplicados no processamento de linguagem natural com a tradução automática de documentos, na visão computacional com reconhecimento facial, na mecânica com automação de carros e em várias outras áreas. Há três abordagens que definem os algoritmos, sendo elas: Aprendizado supervisionado; Aprendizado não supervisionados; Aprendizado por reforço.

O aprendizado de máquina supervisionado acontece sobre a tentativa de prever uma variável dependente tomando como base uma lista de variáveis independentes. Ou seja, a característica básica desse tipo de aprendizado é a de que a lista de dados utilizada para o processo de treinamento contenha também um rótulo ou classe, a qual se faz pela resposta desejada, seja ela de caráter classificatório ou regressiva.

Rosa (2015) demonstra que o aprendizado de máquina supervisionado é uma das

áreas que engloba a maior parte das aplicações bem sucedidas e que possui uma gama de problemas bem definidos, o que permite a utilização de variadas técnicas para solução de problemas, dentre elas a regressão linear, regressão logística, árvore de decisões e outras.

O aprendizado de máquina não supervisionado por sua vez é um dos ramos do aprendizado de máquina e acontece sobre a identificação de padrões previamente desconhecidos, onde o algoritmo analisa os exemplos de teste não categorizados passados como entrada e identifica as semelhanças nos dados, atuando na presença ou ausência de tais semelhanças sobre os novos dados.

De acordo com Pimentel *et al.* (2003), entre as técnicas aplicadas com o aprendizado de máquina não supervisionado, a clusterização possibilita a divisão de forma automática dos dados presentes no conjunto em grupos separados de acordo com a similaridade, sendo esta a técnica implementada neste trabalho sobre os dados coletados das redes sociais. Além dessa técnica, uma outra importante é a de detecção de anomalias, que visa descobrir pontos de dados inusitados sobre a série de dados, sendo útil em contextos de fraudes e discrepâncias ocasionadas por erro humano.

Segundo Biau (2012), o algoritmo Random Forest, também conhecido pelo termo em português Árvore Aleatória, é um algoritmo flexível de Aprendizado de Máquina que possui fácil utilização. Faz parte dos métodos *emsembler* que possuem como característica principal a combinação de diferentes modelos na obtenção de um único resultado. Pertence também ao grupo dos algoritmos supervisionados e é amplamente utilizado em problemas que envolvem classificação e regressão.

De acordo com Breiman (1999) o algoritmo Random Forest cria uma combinação de árvores de decisão de maneira aleatória. Uma árvore de decisão, fundamental para a construção de uma floresta aleatória, cria de forma abstrata uma estrutura similar a um fluxograma. A validação em cada um dos nós na árvore é realizada sobre uma condição que direciona o fluxo para os ramos seguintes. Sobre os dados utlizados para o treinamento do modelo, o algoritmo busca as melhores condições. A aleatoriedade na geração das árvores internas do conjunto geral visa uma maior diversidade, o que, comumente, produz melhores resultados.

Não obstante, o algoritmo Naive Bayes, apesar de simples, dito isto em conformidade com a sua aparente “ingenuidade”, trabalha de forma exemplar, atingindo resultados concorrentes à outros métodos mais aprimorados. Há pressuposições de que sua definição *naive*, ou seja, ingênua, é dada pela forma como trata o conteúdo trabalhado. Além da independência das partes,

há indiferença na ordem de cada uma das palavras em um determinado documento, como se não houvesse importância alguma, o que não se caracteriza verdade em virtude dos fatos, mas da mesma forma, isso não apresenta contrariedade nos resultados além de tornar mais fácil sua implementação.

Assume como base o Teorema de Bayes na classificação de textos, dessa forma, pode ser visto como um algoritmo probabilístico. Tal como qualquer outra estratégia supervisionada de Aprendizado de Máquina, ele contém uma fase de treino com a finalidade de calcular estimadores ótimos para os parâmetros do modelo.

Segundo Amaral (2016), em meio a fase de treinamento é definido uma tabela de valores e concedido um peso a cada atributo de cada classe de classificação. Ao entregar uma nova instância para a classificação, o modelo trata de somar os pesos que foram atribuídos e a classe que contiver o maior resultado será a classe classificadora do item.

O agrupamento é uma técnica muito útil em diversos cenários, e para aplicá-la, pode ser utilizado o algoritmo K-means, o qual se baseia no conceito de similaridade, buscando relacionar itens semelhantes em "k"grupos de acordo com seus atributos. A similaridade é calculada de acordo com a distância entre os valores dos atributos de um dado com os demais, sendo que o algoritmo de distância pode variar de acordo com o problema trabalhado.

Dentre os tipos de agrupamentos que podem ser trabalhados a partir de um processo de clusterização, três deles são comumente utilizados: O tipo de grupo exclusivo separa cada dado em um cluster distinto, não possuindo intercepções entre os conjuntos de dados; No agrupamento sobreposto, um ou mais dados podem está contidos em dois ou mais grupos de conjuntos de dados, possuindo dessa forma, intercepções; Com o agrupamento hierárquico, existe a representação macro e micro, onde um dado grupo pode conter outros subgrupos em sua composição.

Hamerly e Elkan (2004) mostra que o algoritmo do K-means pode ser compreendido em um primeiro momento com uma fase de inicialização, onde há a definição aleatória de pontos centrais com quantidade igual ao parâmetro "k", os quais são utilizados como base para o cálculo de distância entre os atributos dos dados. Após construir os pontos centrais, a fase de atribuição ao cluster ocorre como a aplicação de uma função de distância sobre o dado não rotulado da base de teste e os pontos centrais, sendo que o dado pertencerá ao ponto que mais se assemelha aos seus atributos. Ao final, ocorre a fase de movimentação dos pontos centrais, a qual se dá através da correção dos pontos centrais por meio da média entre os dados pertencentes a cada ponto.

O algoritmo *Density-based spatial clustering of applications with noise* (DBSCAN) é utilizado para a clusterização de dados baseado em densidade. Birant e Kut (2007) mostra que, ao passar um conjunto de pontos em um espaço, estes são agrupados de acordo com a sua proximidade, sendo que cada grupo é determinado com base em uma quantidade de vizinhos pré-determinada, o que permite por sua vez a identificação de outliers, ou seja, pontos que não fazem parte de nenhum grupo específico.

De acordo com Tran *et al.* (2013), o algoritmo DBSCAN necessita de dois parâmetros. O primeiro parâmetro é um valor *epsilon*, denotado como a distância máxima entre dois pontos para que um seja considerado o vizinho do outro. O segundo parâmetro é o número mínimo de pontos que devem ser considerados para que um grupo seja consolidado. Dessa forma, sobre cada ponto no espaço, são obtidos todos os demais que respeitam a distância *epsilon*, sendo que o grupo é construído se ao menos a quantidade mínimo de vizinhos for respeitada.

3.5 Análise de Sentimentos

De acordo com Wang *et al.* (2011), no campo da Ciência da Computação, todos os modelos, algoritmos e técnicas incumbidos por realizarem o tratamento de opiniões são mantidos pelo campo da Análise de Sentimentos, também conhecido como *Sentiment Analysis* (SA). Refere-se a uma área em desenvolvimento que une conceitos de processamento de linguagem natural, aprendizado de máquina, análise textual e mineração de informações.

Sendo uma das áreas que mais tem atraído estudos e pesquisas, buscando auxiliar usuários na definição da polaridade de opiniões, sendo esta segundo Wang *et al.* (2011) "uma declaração subjetiva, com uma visão pessoal, que expressa atitude, emoção ou apreciação sobre uma entidade ou um aspecto de uma entidade, um parecer", ou seja, determina se um conteúdo específico possui significância positiva, neutra ou negativa.

A forma como é calculada a polaridade de uma mensagem, tal como a extração das emoções contidas nela, pode variar de acordo com a abordagem utilizada. Neste trabalho é realizada a Análise de Sentimentos através de algoritmos supervisionados, tal como o Random Forest e Naive Bayes, os quais usufruem de um dicionário contendo pesos previamente atribuídos de forma automática a cada uma das mensagens.

4 TRABALHOS RELACIONADOS

Este capítulo descreve trabalhos da literatura mais relevantes para a contextualização do problema proposto nesta monografia. Na Seção 4.1 são apresentadas duas alternativas de procedimento para *crawling* no Facebook. Na Seção 4.2 é apresentada uma abordagem de integração por meio de *crawlers* entre as redes sociais Twitter e Instagram. Por fim, na Seção 4.3 é apresentado um algoritmo utilizado para a classificação de sentimentos em conteúdo presente no Facebook.

4.1 Crawling no Facebook com propósito de análise da rede social

Foi visto no trabalho de Catanese *et al.* (2011) a estruturação de um esquema para a extração de grande volume de informações aplicado ao Facebook, utilizando um *web crawler* como uma das principais ferramentas. São apresentadas as etapas utilizadas para a realização da coleta da amostra de dados, a qual foi dividida em seis partes únicas e aplicada em duas estratégias distintas, sendo estas o BFS e Uniforme.

Buscando conformidade com as normas de uso destinadas ao usuário final e seguindo as recomendações postas nos termos de serviço aceito logo antes ao cadastro na rede social, os autores optaram por não armazenar nenhuma informação conceitual referente aos perfis processados, registrando apenas o ID e a lista de amigos de cada perfil.

Ao final foi apresentado uma análise da estruturação da rede por meio de um grafo e exame com métricas estipuladas pelo autor. Tendo maior amplitude e maior densidade, o resultado obtido por meio do BFS se mostrou mais útil em termos de aproveitamento de relacionamentos entre os nós existentes no grafo.

Para o cumprimento dos objetivos desta monografia, foi implementada e utilizada uma ferramenta como fator principal na coleta de informações do Facebook, a qual também utiliza-se da técnica BFS, porém, diferentemente do trabalho de Catanese *et al.* (2011), há o registro de comentários e outras informações públicas contidas na rede social.

4.2 Integrando Twitter e Instagram para monitoramento de eventos do mundo real

No trabalho de Giridhar *et al.* (2017) é estruturado um serviço que possui como finalidade a identificação de eventos do mundo real pela análise de sinais existentes nas redes sociais Twitter e Instagram. Para a realização da captura dos dados presentes em tais mídias

sociais, foi necessário a implementação de *web crawlers*, os quais utilizaram funcionalidades presentes em *Applications Protocol Interface* (API) fornecidas por tais plataformas.

Diferentemente do que é desenvolvido no corrente trabalho pela implementação da busca em largura, em Giridhar *et al.* (2017) é utilizado um arquivo contendo palavras chaves definidas pelo autor com o propósito de realizar buscas no ambiente social e extrair os dados de acordo com o resultado obtido.

Buscando remover todas as redundâncias presentes nos resultados de algumas das palavras chaves informadas através do arquivo, foi utilizado um algoritmo de aprendizado de máquina não supervisionado em um módulo de clusterização/agrupamento dos dados textuais. No atual trabalho, também será aplicado um algoritmo de clusterização, mas com a finalidade de identificar padrões sobre as mensagens de texto dos usuários, tais quais possam corroborar com a já estruturada tabela de níveis de ideação suicida, suscitada na seção 5.4.1.

4.3 Análise de sentimentos no Facebook

Foi proposto por Rodrigues (2017) um modelo para análise de sentimentos supervisionada e outro para o pré-processamento de textos por meio de uma ferramenta implementada em *Python 3*, utilizada para um estudo de caso na página no Facebook do Senado Federal Brasileiro.

Foram classificadas duas bases de dados compostas por comentários a respeito da reforma do ensino médio e limitação de dados em banda larga fixa, as quais foram processadas através de uma terceira base, inicialmente treinada com o algoritmo *Naive Bayes Multinomial Text*, usado na classificação das sentenças.

Ao final, foram utilizados gráficos para demonstrar os resultados obtidos através do algoritmo escolhido, assim como os dados fornecidos pela ferramenta de extração, concluindo a satisfação dos cumprimentos dos objetivos e demonstrando ideias que podem ser usadas para melhorar ainda mais os resultados, tal como a análise em nível de aspecto.

Assim como em Rodrigues (2017), no atual trabalho será realizada a Análise de Sentimentos em informações originalmente capturadas a partir do Facebook, distinguindo-se principalmente pela implementação e comparação dos resultados obtidos através de algoritmos distintos, os quais demonstrarão uma melhor visão da abordagem.

5 PROCEDIMENTOS METODOLÓGICOS

Este capítulo descreve os procedimentos metodológicos utilizados para a realização deste trabalho. A pesquisa foi dividida em 5 etapas: A definição das mídias sociais utilizadas para a construção de uma base de dados necessária para o desenvolvimento do estudo; a implementação de *web crawlers* para captura de dados presentes nessas mídias sociais; limpeza e tratamento das informações; rotulamento de forma automática e manual do conteúdo textual presente na base; e por último, a implementação de algoritmos de classificação e clusterização para a realização dos experimentos. A Figura 1 ilustrada a seguir, demonstra o fluxo dos procedimentos metodológicos adotados na pesquisa.

Figura 1 – Fluxograma da Metodologia de Pesquisa

5.1 Definição das mídias sociais

Devido ao grande número de mídias sociais presentes na Internet, foram selecionadas três redes sociais e um fórum web. As redes sociais escolhidas para o desenvolvimento do estudo foram o Facebook, Instagram e Twitter, assim como o fórum web Mundo Psicólogos. Os parâmetros utilizados para escolher tais redes sociais foi excepcionalmente a popularidade e as atividades de seus usuários, assim como a especialidade do fórum Mundo Psicólogos, que possui áreas específicas para a expressão de questões emocionais.

5.2 Implementação e execução dos web crawlers

Algumas das redes sociais escolhidas para o desenvolvimento deste trabalho tornaram privadas ou descontinuaram API e componentes auxiliares que eram utilizadas para a coleta de informações presentes no seu ambiente. Notou-se então a necessidade de criar ferramentas que fossem úteis para extração das informações pretendidas. Com isso, após ter sido estudado métodos possíveis para a obtenção de dados em páginas e sites da internet, concluiu-se que a melhor opção seria a implementação de web crawlers capazes de extrair todas as informações de

forma eficaz.

5.2.1 Definição das tecnologias

Para o processo de implementação foi necessário pesquisar linguagens que fornecessem suporte para a simulação de forma fidedigna à atividade de um ser humano em páginas da internet. Não existindo alguma que fosse superior em todos os aspectos, foi escolhido realizar a implementação utilizando linguagem de programação *Python*, pelas seguintes razões:

- *Python*[1] é uma linguagem multi-paradigma, fácil, ágil, objetiva e de alto nível com uma grande comunidade de apoio aos seus desenvolvedores;
- Possui uma gama de módulos que foram desenvolvidos especificamente para tal finalidade, tais como os comumente utilizados *Scrapy*[2] e *Mechanize*[3], além de outros auxiliares que complementam o processo de raspagem dos dados, como o *Mechanicalsoup*[4], *Beautifulsoup*[5] e *Selenium*[6], os quais foram utilizados neste trabalho;
- Trabalhos relacionados, descritos nas seções 4.2 e 4.3, utilizaram a linguagem como principal ferramenta para a implementação.

5.2.2 Definição das características de cada mídia social

Pela independência entre as características presentes em cada uma das mídias sociais exploradas por este trabalho, foi necessário realizar o mapeamento do código HTML apresentado pelo navegador ao *web crawler*, tal qual representasse uma mensagem do usuário a ser coletada e também da página seguinte que o algoritmo deveria visitar após a coleta, além de definir uma condição de parada para o algoritmo, onde a abordagem da busca em largura fosse devidamente trabalhada. Segue a lista com as regras definidas sobre o mapeamento para cada mídia social:

- **Facebook**: Uma postagem válida do usuário foi compreendida como o conteúdo textual presente em tags “div” com atributo “data-ad-comet-preview” contendo o valor “message”. Para as páginas seguintes na busca em largura, foi realizado o carregamento da página de amigos do usuário que estava sendo processado e sobre ela capturado os links contidos como valores dos atributos “href” das tags “a” que tivessem o valor “auto” em um segundo

[1] https://www.python.org/
[2] https://scrapy.org/
[3] https://mechanize.readthedocs.io/en/latest/
[4] https://mechanicalsoup.readthedocs.io/en/stable/
[5] https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[6] https://www.selenium.dev/documentation/

atributo denominado “dir”. O procedimento de busca foi iniciado a partir do perfil utilizado para o login na rede social e finalizado de forma manual pelo encerramento do processo.

- **Instagram**: Uma publicação válida do usuário foi compreendida como o conteúdo textual presente em tags “li” com atributo “role” contendo o valor “menuitem”. Para as páginas seguintes na busca em largura, foi realizado o carregamento da página de seguidores do usuário que estava sendo processado e sobre ela capturado os links contidos como valores dos atributos “href” das tags “a” que tivessem o atributo “tabindex” com valor “0” e um segundo atributo “class” contendo três valores separados por espaço, sendo um deles “notranslate”. O procedimento de busca foi iniciado a partir do perfil utilizado para o login na rede social e finalizado de forma manual pelo encerramento do processo.
- **Twitter**: Um *Tweet* válido do usuário foi compreendida como o conteúdo textual presente em tag “div” com atributo “dir” contendo o valor “auto” e apenas um elemento filho do tipo “span”. Para as páginas seguintes na busca em largura, foi realizado o carregamento da página de seguidores do usuário que estava sendo processado e sobre ela capturado os links contidos como valores dos atributos “href” das tags “a” que tivessem o atributo “role” com valor “auto” e um segundo atributo “class” contendo sete valores separados por espaço. O procedimento de busca foi iniciado a partir do perfil utilizado para o login na rede social e finalizado de forma manual pelo encerramento do processo.
- **Mundo Psicólogos**: Uma publicação válida do usuário foi compreendida como o conteúdo textual presente em tag “div” com atributo “class” contendo o valor “Message UserContent”. Para as páginas seguintes na busca em largura, foi realizado o carregamento da página principal da comunidade e sobre ela capturado os links contidos como valores dos atributos “href” das tags “a” que tivessem o elemento “li” como pai. O procedimento de busca foi iniciado a partir do tópico mais recente presente no dia da execução e finalizado de forma manual pelo encerramento do processo.

5.2.3 Execução dos Web Crawlers

O processo de execução dos algortimos ocorreu de forma intercalada, a qual permitiu a coleta de uma amostra de cada mídia social inicialmente idealizada, contendo informações textuais do tipo mensagem, as quais foram suficientes para consolidar a base de dados utilizada para a progressão do estudo. Cada base foi estruturada em um banco de dados orientado a documentos, utilizando o *MongoDB*.

5.3 Limpeza e tratamento da base de dados

Após realizar a extração dos dados foi executado um processo de limpeza, devido às inúmeras ocorrências que precisam ser removidas ou normalizadas no texto, contribuindo dessa forma para a redução de dados sem relevância. Esses dados constituem-se em sentenças ou termos que não agregam nenhuma informação ao conjunto geral e fazem com que o tempo de processamento seja maior. Com isso, Gonçalves *et al.* (2013) fala que a etapa de pré-processamento é de extrema importância, pois irá eliminar termos irrelevantes.

A remoção desses termos possibilita que os modelos finais sejam mais precisos e demorem menos tempo no processo de treino. Logo, diante das estratégias que podem ser utilizadas, o procedimento de remoção de *stopwords* é comumente um dos métodos mais utilizados nessa etapa de pré-processamento da base.

De acordo com Gonçalves *et al.* (2013), o processo de remoção de *stopwords* pode ser especificado como a eliminação de palavras ou termos que não agregam informações importantes ou de alguma relevância ao sentimento de um determinado texto. Logo, todo idioma dispõe de suas próprias *stopwords*, sendo exemplo dos da língua portuguesa os termos "aos", "o"e "que", os quais podem ser encontrados na lista adotada pela biblioteca *Natural Language Toolkit* (NLTK) do *Python*.

A etapa de limpeza e tratamento da base de dados foi composta não apenas pela remoção de *stopwords*, mas também sobre a elaboração de algumas regras específicas, as quais podem ser vistas em conjunto a exemplos de aplicação por meio da Tabela 1. As regras adicionais foram necessárias para tratar o tipo de dado trabalhado, originados a partir de mídias sociais.

Tabela 1 – Tabela de regras para limpeza dos dados

Definição da regra	**Mensagem sem regra**	**Mensagem com regra**
A mensagem deve conter todas as letras em minúsculo	BOOOOM DIIA!!	boom diia!!
A mensagem não pode conter *Uniform Resource Locator* (URL)	Vejam que notícia incrível http://www.site.com/exemplo	Vejam que notícia incrível
A mensagem não pode conter caracteres duplicados	eeuuu queerooooo!!	eu quero!
A mensagem não pode conter acentuações	você é incrível!	voce e incrivel!
A mensagem não pode conter *stopwords*	quero te ver hoje à noite	quero ver hoje noite
A mensagem não pode conter menções à usuários	veja isso @fulanodetal que demais!	veja isso que demais!

Outro ponto importante relacionado ao tratamento dos dados corresponde à anonimização. Para manter a privacidade dos usuários dessas mídias sociais, foi realizada a desconexão de informações pessoais e gerais, tais quais pudessem indentificar de alguma forma o usuário responsável pela mensagem presente na base de dados. A integridade dos relacionamentos entre os dados coletados com seu responsável se deu pela atribuição de uma *hash* do tipo *Message-Digest Algorithm 5* (MD5), a qual foi calculada utilizando o dado de identificação do usuário na rede social com um valor arbitrário gerado no momento da execução do processo.

5.4 Rotulamento da base de dados

O rotulamento da base de dados é uma das etapas mais importantes desse estudo, pois, como alguns dos experimentos são realizados sobre a aplicação de algoritmos supervisionados, tal rótulo se torna necessário para o cumprimento desse passo. O termo "rótulo", ou mesmo, "classe" é utilizado para denotar uma caracteristica pertinente à informação presente na base de dados, sendo esta a propriedade responsável pelo aprendizado do algoritmo.

Com a intenção de construir um padrão sobre os dados para validar o desempenho das técnicas aplicadas neste estudo, foi realizado um procedimento de classificação dos dados de forma manual por meio humano e um outro, automático, através de algoritmo. Ambas abordagens foram propostas para o cumprimento da análise de sentimentos e de ideações suicidas.

5.4.1 Rotulamento Manual

O rotulamento manual da base foi pensado sobre a problemática de analisar ideações suicidas em mensagens de texto. Na literatura, não foram encontrados trabalhos que pudessem ser usados como fontes auxiliares, logo, foram convidados Psicólogos da *Universidade Federal do Ceará* (UFC) do campus russas para realizar essa atividade. Foi selecionada uma amostra da base de dados contendo 500 mensagens de cada mídia social e 100 do fórum, as quais foram organizadas em uma planilha compartilhada de forma online por onde os profissionais puderam realizar o procedimento de rotulação.

Em (WHO) *et al.* (2006) é definida a Escala de Avaliação de Risco Suicida, a qual contém quatro níveis e é comumente utilizada por profissionais da área de Psicologia para avaliarem o risco de suicídio em uma dada pessoa através de suas mensagens. Tal escala foi usada como base para o rotulamento dos textos contidos na amostra. Pelo fato da escala cobrir

apenas casos suicidas e, sabendo que a base não é composta apenas por mensagens de pessoas com ideações suicidas, um outro nível foi adicionado, sendo identificado por "inexistente". Todos os níveis utilizados nesse trabalho são apresentados a seguir, com uma breve descrição.

- **Inexistente:** Essencialmente, nenhum risco de se fazer mal.
- **Leve:** A ideação suicida é limitada, não há nenhum plano ou preparação definida para se fazer mal, e há poucos fatores de risco conhecidos. A intenção de cometer suicídio não é aparente, mas a ideação suicida está presente. O indivíduo não tem um plano concreto e não tentou cometer suicídio no passado.
- **Moderado:** São evidentes planos definidos e preparação, com visível ideação suicida. Há possivelmente história de tentativas anteriores, e pelo menos dois fatores de risco adicionais. Ou, mais do que um fator de risco para o suicídio está presente, a ideação suicida assim como a intenção estão presentes, mas é negado que haja um plano claro. O indivíduo está motivado para melhorar o seu estado emocional e psicológico atual se houver ocasião para tal.
- **Severo:** Os planos e a preparação para se infligir mal foram claramente definidos ou a pessoa é reconhecida como alguém que já tentou múltiplas vezes o suicídio com dois ou mais fatores de risco. A ideação e a intenção suicida são verbalizadas em conjunto com um plano bem estudado e com os meios de o levar a cabo. Este indivíduo demonstra inflexibilidade cognitiva e desesperança quanto ao futuro e nega o apoio social disponível. Houve tentativas de suicídio anteriores.
- **Extremo:** Um indivíduo que tentou o suicídio múltiplas vezes com diversos fatores de risco significativos. Atenção e ação imediata são imprescindíveis.

5.4.2 Rotulamento Automático

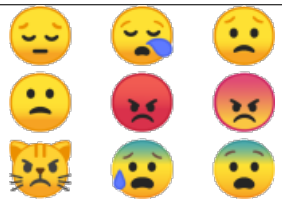
A abordagem do rotulamento automático foi idealizada como passo para o cumprimento da análise de sentimentos. No trabalho de Wood e Ruder (2016), é definida uma relação entre *emojis* e emoções básicas, as quais são agrupadas em seis classes denotadas por raiva, felicidade, surpresa, nojo, tristeza e medo. Essas classes foram utilizadas para o rotulamento automático de emoções sobre um conjunto de textos, que serviu para o treinamento de algoritmos de aprendizado de máquina.

Wood e Ruder (2016) também demonstra que as emoções podem ser compreendidas como boas ou ruins. Com isso, buscando determinar a polaridade de uma mensagem, foram

utilizados *emojis* para denotar se uma mensagem possui teor positivo ou negativo. A Tabela 2 apresenta os dois rótulos dessa abordagem além dos *emojis* utilizados em cada classe e suas representações em *unicode*.

Para que o algoritmo realizasse o rotulamento da mensagem, a mesma deveria respeitar uma dada condição: Não possuir dois ou mais *emojis* de classes diferentes. Dessa forma, mensagens que fossem "emocionalente ambíguas"foram excluídas, possibilitando assim uma classificação mais coerente.

Tabela 2 – Tabela com rótulos das mensagens para Análise de Sentimentos com emojis e sua representação em Unicode

Rótulo	**Unicode**	**Emoji**
Positivo	U+1F600, U+1F602, U+1F603, U+1F604, U+1F606, U+1F609, U+1F60A, U+1F607, U+1F60B	
Negativo	U+1F614, U+1F62A, U+1F61F, U+1F641, U+1F621, U+1F620, U+1F63E, U+1F630, U+1F628	

5.5 Implementação dos algoritmos de aprendizado de máquina

Previamente foi necessário realizar a instalação e configuração da distribuição Anaconda, a qual é destinada à computação científica e conta com um vasto grupo de pacotes já pré-configurados e que foram exclusivamente utilizados para a manipulação dos dados, tal como para implementar os algoritmos de aprendizado de máquina. Dentre os módulos dispostos pelo *framework*, os mais relevantes para o desenvolvimento deste trabalho foram o *Python Data Analysis Library* (Pandas) e o *Scikit-learn*.

Foi utilizada a biblioteca do *Scikit-learn* para realizar a implementação do modelo Random Forest através do uso da classe *RandomForestRegressor*. Dentre os parâmetros destacados pelo módulo, o principal é denotado por *n_estimators*, o qual possui como valor padrão o número 100 e se resume na quantidade de árvores de decisões que poderão ser criadas na construção da floresta. Com a finalidade de encontrar uma melhor configuração para o modelo,

foi realizada uma iteração sobre o parâmetro *n_estimators*, tendo início em 5 e fim em 50.

Para realizar a implementação do algoritmo Naive Bayes, também foi utilizada como auxílio a biblioteca do *Scikit-learn*, que disponibiliza três diferentes classes para o seu desenvolvimento, sendo a *MultinomialNB* a destinada para classificação de textos e, dessa forma, escolhida para tratar o problema deste trabalho. Tendo como objetivo uma melhor adequação do modelo no momento do treinamento sobre os dados de entrada, o parâmetro *fit_prior* foi configurado com o valor "true".

Os modelos de classificação Random Forest e Naive Bayes foram utilizados para realizar a análise de sentimentos adotando como parâmetro de divisão para os dados de entrada em 70% com a finalidade de treinamento e 30% para teste. Os treinamentos foram realizados com 10.125 mensagens, sendo 5.468 de polaridade positiva (54%) e 4.653 (46%) negativa. Os testes utilizaram 4.337 mensagens, sendo 2.485 (57%) de polaridade positiva e 1.852 negativa (43%).

As avaliações desses modelos de classificação foram realizadas através do uso das métricas *Accuracy*, *Recall* e *Precision*. Por meio da métrica *accuracy* é possível avaliar a taxa de acerto de todas as classes do modelo. *Recall* calcula a taxa de acerto da classe positiva e *Precision* a probabilidade de uma observação realmente pertencer à classe escolhida pelo modelo, dado que o modelo classificou uma determinada observação como positivo (*precision pos*) ou negativo (*precision neg*).

Para identificar padrões de ideações suicidas dentre as mensagens das mídias sociais e relacionar com as classificações realizadas de forma manual, o módulo *Scikit-learn* foi utilizado para implementar os algoritmos de clusterização K-means e DBSCAN. O parâmetro *n_clusters* do algoritmo K-means foi configurado com o valor "5", pois essa é a quantidade de classes em que os textos que servirão de entrada terão, segundo classificação realizada previamente de forma manual. O algoritmo DBSCAN teve os parâmentros *eps* e *metric* configurados respectivamente com os valores "0.79"e "euclidean", pois essa configuração resultou em 5 grupos, ou seja, a quantidade necessária para alinhar com os dados analisados. Os algoritmos foram alimentados com 1.600 mensagens. Estando os dados previamente rotulados pelos profissionais, o processo de avaliação dos modelos foi realizado através do comparativo entre a proporção de mensagens identificadas nos clusters gerados pelos modelos com a dos rótulos já conhecidos.

6 RESULTADOS E DISCUSSÕES

Nesta seção, serão apresentados e discutidos os resultados obtidos através da implementação dos *web crawlers*, tratamento realizado na base de dados, rotulamento das mensagens estruturadas e aplicação dos modelos de aprendizado de máquina para classificação de análise de sentimentos e agrupamento de mensagens no contexto de ideação suicida sobre o conteúdo textual de mídias sociais.

6.1 Implementação dos web crawlers e tratamento da base de dado

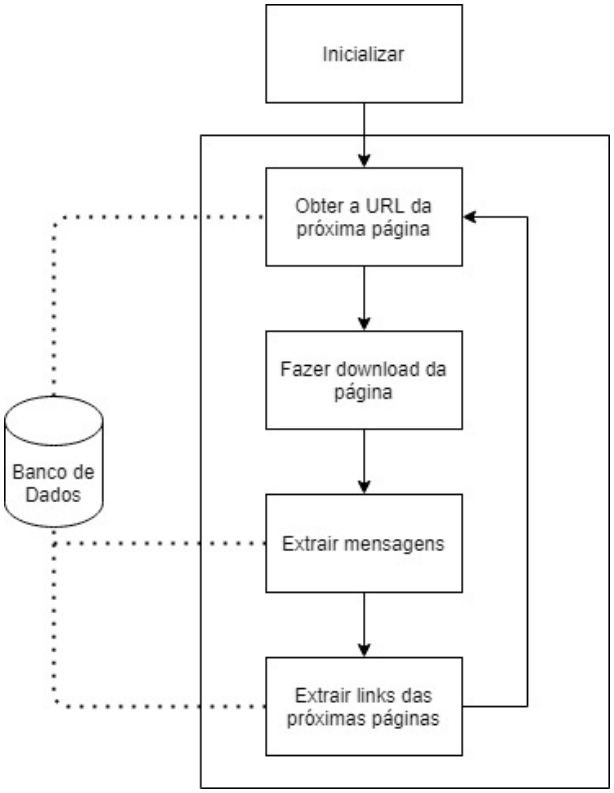


Figura 2 – Fluxo generalizado dos webcrawlers

A figura 2 mostra o fluxo generalizado de execução entre os módulos implementados para cada mídia social. Com exceção do fórum Mundo Psicólogos, o processo de extração foi dificultado nas demais mídias sociais, pelo fato de que nenhuma delas permitia a interação com alta velocidade nas páginas e publicações. Para contornar o problema, foi necessário acrescentar um passo de espera em segundos entre a captura de um dado e outro, o qual foi definido como um número aleatório no intervalo de 10 a 30. A execução desse processo deu origem a uma base de dados com 32.794 registros.

Tabela 3 – Tabela com quantidade e porcentagem de dados antes e depois do tratamento

Mídia Social	**Antes do tratamento**		**Depois do tratamento**	
	Quantidade	Percentual	Quantidade	Percentual
Facebook	10.094	30.80%	9.980	31.21%
Instagram	14.252	43.45%	13.795	43.15%
Twitter	5.725	17.45%	5.522	17.27%
Mundo Psicólogos	2.723	8.30%	2.677	8.37%
Total	**32.794**	**100%**	**31.974**	**100%**

Após consolidar a base de dados, o conteúdo extraído das mídias sociais foi então processado por uma etapa de tratamento, a qual preparou os registros para as fases seguintes do estudo. O conjunto de dados resultante desse processo foi de 31.974 mensagens. Por meio da Tabela 3 é possível acompanhar a quantidade em volume das mensagens relacionadas a cada mídia social, antes e depois do tratamento.

6.2 Rotulamento das mensagens

6.2.1 Manual

Foram rotuladas 1.600 mensagens, as quais ficaram distribuídas em 1.523 com rótulo inexistente, 15 com rótulo leve, 52 com rótulo moderado, 4 com rótulo severo e 6 com rótulo extremo. A princípio, foi notada a discrepância em termos de quantidade sobre cada rótulo, contudo, já era esperado valores menos recorrentes dos rótulos leves a extremos. Por meio da Tabela 4 é possível conferir algumas das mensagens classificadas na base, acompanhada de seus respectivos rótulos.

6.2.2 Automático

O processo de classificação automática foi realizado sobre a base tratada. Devido às restrições do método de rotulamento, as mensagens que não continham *emojis* foram ignoradas. Após processar 31.974 mensagens, foram rotuladas 14.462, sendo 7.953 (55%) de polaridade positiva e 6.509 (45%) negativa. Por meio da figura 3 é possível visualizar a distribuição em quantidade de mensagens em que cada *emoji* está presente.

Tabela 4 – Tabela com mensagens rotuladas com níveis de ideação suicida

Rótulo	**Frase**
Inexistente	jesus cristo, socoro, quero sumir dese país, não aguento mais essa política de corrupção
Leve	é hj q vou morer de tanto sofrer
Moderado	o meu pequeno problema com a depressão [ou talvez seja mera melancolia, não faço ideia do que eu tenho, pensei que passaria junto com a adolescência mas isso persiste] é que não tenho coragem suficiente pra me matar
Severo	em 2016/2017/2018 eu tive depressão, eu me cortava, eu me arranhava, ficava me culpando inteira. teve um dia que eu tentei me matar me jogando da varanda da minha avó, mas a minha mãe me pegou antes que eu pulasse e me levou pro hospital.
Extremo	cometi suicídio com anti-depressivos em grande quantidades e nem pra isso eu sirvo

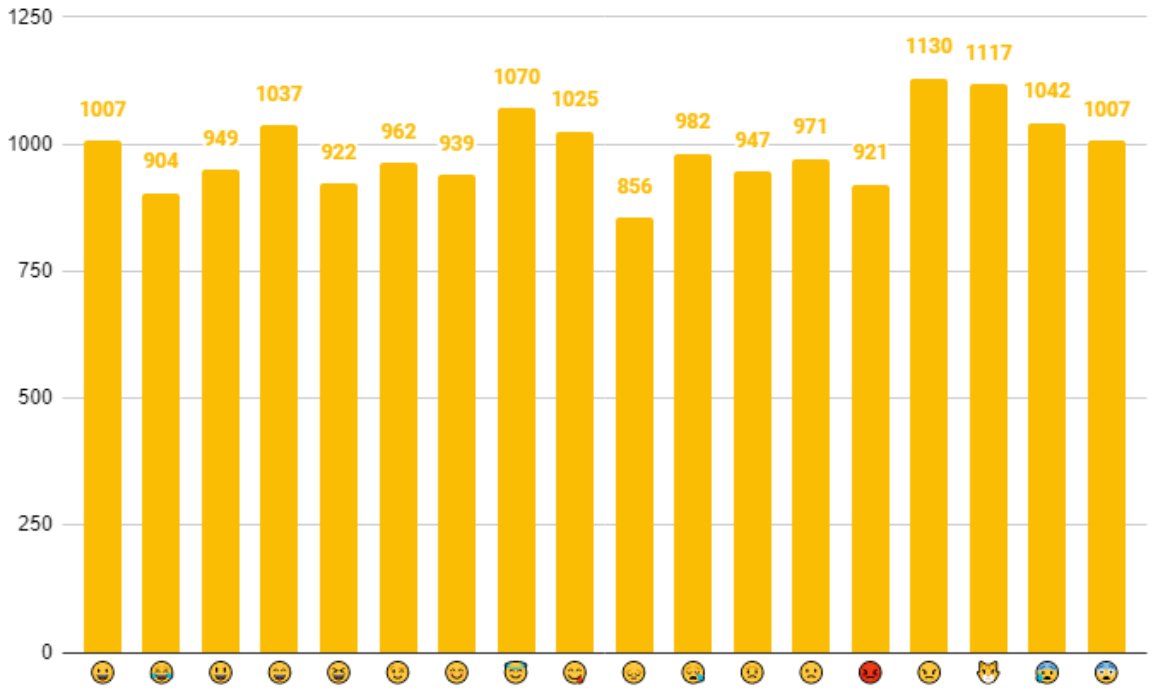


Figura 3 – Gráfico de linhas com a distribuição em volume dos emojis presentes nas mensagens após tratamento

Tabela 5 – Matriz de confusão do modelo Random Forest

Classificação prevista	**Classificação atual**	
	Positivo	**Negativo**
Positivo	1.838	445
Negativo	647	1.407

Tabela 6 – Tabela com os resultados das métricas aplicadas sobre diferentes estimadores com o algoritmo RandomForest

Número de estimadores	**Accuracy**	**Recall**	**Precision Pos**	**Precision Neg**
5	0,294	0,320	0,367	0,222
10	0,530	0,560	0,595	0,453
25	0,728	0,720	0,788	0,663
50	0,748	0,740	0,805	0,685

6.3 Aplicação dos modelos de aprendizado de máquina

6.3.1 Análise de sentimentos

A Tabela 5 é a representação da matriz de confusão originada sobre 50 estimadores do modelo RandomForest, a qual foi identificada como a melhor configuração de acordo com a análise das métricas da Tabela 6. Sobre a matriz de confusão é possível observar que o modelo previu 2.283 frases de polaridade positiva, sendo 80% corretas. Além disso, houve também a previsão de 2.054 frases com polaridade negativa, em que apenas 31% foram classificações incorretas.

O resultado das métricas utilizadas para avaliar o modelo Naive Bayes pode ser encontrado na tabela 7. Dessa forma, a matriz de confusão que está representada na Tabela 8 demonstra que o modelo previu 2.343 mensagens com polaridade positiva, onde aproximadamente 66% foram classificações corretas. O modelo também previu 53% das mensagens com polaridade negativa de forma correta.

Portante, ao realizar um comparativo entre as resoluções alcançados pelas métricas de cada modelo, é possível observar que nesse contexto da análise de sentimentos, o algoritmo Random Forest se sobressaiu com condições melhores do que o Naive Bayes, tendo resultado em valores superiores sobre todas as métricas utilizadas.

Tabela 7 – Tabela com os resultados das métricas obtidas do modelo Naive Bayes

Accuracy	**Recall**	**Precision Pos**	**Precision Neg**
0,608	0,630	0,668	0,539

Tabela 8 – Matriz de confusão do modelo Naive Bayes

Classificação prevista	**Classificação atual**	
	Positivo	**Negativo**
Positivo	1.565	778
Negativo	920	1.074

Tabela 9 – Distribuição de frequências dos grupos gerados pelo K-means segundo classes de ideações suicidas

grupos	**inexistente**		**leve**		**moderado**		**severo**		**extremo**		**total**
	n	%	n	%	n	%	n	%	n	%	**n**
g1	951	96,7	6	0,6	24	2,4	1	0,1	1	0,1	983
g2	128	92,8	9	6,5	1	0,7	0	0,0	0	0,0	138
g3	295	98,0	0	0,0	2	0,7	0	0,0	4	1,3	301
g4	98	83,1	0	0,0	19	16,1	0	0,0	1	0,8	118
g5	51	85,0	0	0,0	6	10,0	3	5,0	0	0,0	60
total	1523	95,2	15	0,9	52	3,3	4	0,3	6	0,4	1600

6.3.2 Ideações Suicidas

A tabela 9 demonstra a distribuição de frequências dos cinco grupos gerados pelo algoritmo K-Means. Observou-se que os grupos 1, 2 e 3 foram constituídos por mais de 90% de registros com classificação “inexistente”, sendo os grupos 4 e 5 formados com pouco mais de 80% dessa classe. Além disso, notou-se também que alguns grupos não foram formados com alguns dos demais rótulos, excepcionalmente o rotulado por “severo” que continha apenas 4 registros na base e dessa forma seria inviável possuir um registro para cada grupo.

Por meio da tabela 10 é possível acompanhar a distribuição de frequências dos cinco grupos gerados pelo algoritmo DBSCAN. Notou-se que o grupo 4 foi formado unicamente por registros da classe “inexistente”, porém, os rótulos dessa classe também foram organizados em outros grupos. A classe “severo” foi completamente organizada no grupo 2. Observou-se também que o grupo 5 foi estruturado com apenas 3 registros do rótulo “inexistente”.

Contudo, nenhum do dois modelos conseguiu identificar plenamente os mesmos padrões encontrados pelos psicólogos, caso contrário, cada grupo seria composto exclusivamente pelos registros de uma mesma classe. Porém, mediante agrupamento, o algoritmo K-means agrupou de forma mais próxima às expectativas dos profissionais os rótulos “inexistente” e “leve”, sendo “moderado” e “severo” melhor aproximado com o DBSCAN. O rótulo “extremo” foi proporcional em ambos os algoritmos.

Tabela 10 – Distribuição de frequências dos grupos gerados pelo DBSCAN segundo classes de ideações suicidas

grupos	**inexistente**		**leve**		**moderado**		**severo**		**extremo**		**total**
	n	%	n	%	n	%	n	%	n	%	**n**
g1	585	92,3	5	0,8	44	6,9	0	0,0	0	0,0	634
g2	644	97,9	5	0,8	3	0,5	4	0,6	2	0,3	658
g3	24	75,0	4	12,5	0	0,0	0	0,0	4	12,5	32
g4	267	100,0	0	0,0	0	0,0	0	0,0	0	0,0	267
g5	3	33,3	1	11,1	5	55,6	0	0,0	0	0,0	9
total	1523	95,2	15	0,9	52	3,3	4	0,3	6	0,4	1600

7 CONCLUSÃO E TRABALHOS FUTUROS

De forma concisa, apresentou-se neste trabalho o que se trata um projeto no campo da ciência de dados, demonstrando desde meios automatizados para extração de dados sobre mídias sociais, passando por diferentes regras aplicáveis para o tratamento de registros textuais, e chegando a aplicação de diferentes modelos de aprendizado de máquina em contextos específicos e bem definidos.

Sob esse ângulo, compararam-se algumas abordagens para a utilização de algoritmos de aprendizado de máquina para a classificação e clusterização de dados de mídias sociais, provenientes da implementação de *web crawlers*. Dessa forma, os algoritmos desenvolvidos no decorrer do estudo buscaram solidificar uma base para aplicação de análise de sentimentos e ideações suicidas sobre registros das mídias sociais Facebook, Instagram, Twitter e Mundo Psicólogos.

A quantidade de dados gerados por meio da execução dos *web crawlers* foram suficientes para aplicação de regras na etapa de tratamento e também pela classificação automática das mensagens no contexto da análise de sentimentos, porém, observou-se que as classificações realizadas manualmente pelos profissionais da universidade resultou em quantidades desproporcionais sobre os rótulos definidos. Nesse sentido, fica clara a performance dos algoritmos de clusterização.

Portanto, a partir das limitações citadas acima, propõe-se como trabalho futuro a obtenção de uma nova amostra de dados mediante execução de *web crawlers*, além de uma nova classificação manual contendo um maior número de registros. Ademais, recomenda-se o estudo de outros modelos criados com redes neurais, a utilizar uma abordagem focada na otimização dos parâmetros de treinamento.

REFERÊNCIAS

ALPAYDIN, E. **Introduction to machine learning**. [S.l.]: MIT press, 2009.

AMARAL, F. **Aprenda mineração de dados: teoria e prática**. [S.l.]: Alta Books Editora, 2016. v. 1.

BARZ, M.; KLAUS, E. do C.; RODRIGUES, D. F.; LAGO, M. L. P.; OELKE, C. A.; FRAGA, B. N. A extensão universitária através das mídias sociais. **Anais do Salão Internacional de Ensino, Pesquisa e Extensão**, v. 12, n. 3, 2020.

BIAU, G. Analysis of a random forests model. **The Journal of Machine Learning Research**, JMLR. org, v. 13, n. 1, p. 1063–1095, 2012.

BIRANT, D.; KUT, A. St-dbscan: An algorithm for clustering spatial–temporal data. **Data & knowledge engineering**, Elsevier, v. 60, n. 1, p. 208–221, 2007.

BREIMAN, L. Random forests. **UC Berkeley TR567**, 1999.

BURNAP, P.; COLOMBO, W.; SCOURFIELD, J. Machine classification and analysis of suicide-related communication on twitter. In: ACM. **Proceedings of the 26th ACM conference on hypertext & social media**. [S.l.], 2015. p. 75–84.

CATANESE, S. A.; MEO, P. D.; FERRARA, E.; FIUMARA, G.; PROVETTI, A. Crawling facebook for social network analysis purposes. In: ACM. **Proceedings of the international conference on web intelligence, mining and semantics**. [S.l.], 2011. p. 52.

CORMEN, T. H.; LEISERSON, C. E.; RIVEST, R. L.; STEIN, C. **Introduction to algorithms**. [S.l.]: MIT press, 2009.

GIRIDHAR, P.; WANG, S.; ABDELZAHER, T.; AMIN, T. A.; KAPLAN, L. Social fusion: Integrating twitter and instagram for event monitoring. In: IEEE. **2017 IEEE International Conference on Autonomic Computing (ICAC)**. [S.l.], 2017. p. 1–10.

GONÇALVES, P.; BENEVENUTO, F.; ALMEIDA, V. O que tweets contendo emoticons podem revelar sobre sentimentos coletivos? In: SBC. **Anais do II Brazilian Workshop on Social Network Analysis and Mining**. [S.l.], 2013. p. 128–139.

HAMERLY, G.; ELKAN, C. Learning the k in k-means. **Advances in neural information processing systems**, MIT Press, v. 16, p. 281–288, 2004.

JONES, S.; FOX, S. *et al.* **Generations online in 2009**. [S.l.]: Pew Internet & American Life Project Washington, DC, 2009.

KAUSAR, M. A.; DHAKA, V.; SINGH, S. K. Web crawler: a review. **International Journal of Computer Applications**, Foundation of Computer Science, v. 63, n. 2, 2013.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicações**, v. 1, n. 1, p. 32, 2003.

MORANDIN, J. L. P. L. Análise do desempenho altmétrico da revista movimento nas redes sociais facebook, twitter e instagram. 2021.

PIMENTEL, E. P.; FRANÇA, V. F. de; OMAR, N. A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2003. v. 1, n. 1, p. 495–504.

PINKERTON, B. **Webcrawler: Finding what people want**. [S.l.]: Citeseer, 2000.

POLO, F.; POLO, J. L. **# socialholic: Tudo o que você precisa saber sobre marketing nas mídias sociais**. [S.l.]: Editora Senac São Paulo, 2021.

RODRIGUES, A. C. F. Modelo para análise de sentimentos no facebook: um estudo de caso na página do senado federal brasileiro. 2017.

ROSA, R. L. **Análise de sentimentos e afetividade de textos extraídos das redes sociais.** Tese (Doutorado) — Universidade de São Paulo, 2015.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of research and development**, IBM, v. 3, n. 3, p. 210–229, 1959.

TRAN, T. N.; DRAB, K.; DASZYKOWSKI, M. Revised dbscan algorithm to cluster data with dense adjacent clusters. **Chemometrics and Intelligent Laboratory Systems**, Elsevier, v. 120, p. 92–96, 2013.

WANG, X.; WEI, F.; LIU, X.; ZHOU, M.; ZHANG, M. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In: ACM. **Proceedings of the 20th ACM international conference on Information and knowledge management**. [S.l.], 2011. p. 1031–1040.

(WHO), W. H. O. *et al.* Prevenção do suicídio: Um recurso para conselheiros. **Genebra: OMS. Recuperado de http://www. who. int/mental_health/media/counsellors_portuguese. pdf**, 2006.

WOOD, I.; RUDER, S. Emoji as emotion tags for tweets. In: **Proceedings of the Emotion and Sentiment Analysis Workshop LREC2016, Portorož, Slovenia**. [S.l.: s.n.], 2016. p. 76–79.