Fundação Educacional do Município de Assis
Instituto Municipal de Ensino Superior de Assis
Campus "José Santilli Sobrinho"

GUSTAVO HENRIQUE MELO SOUZA

MINERAÇÃO DE EMOÇÕES: ALGORITMO PARA DETECTAR PENSAMENTOS DEPRESSIVOS EM REDES SOCIAIS.

Assis/SP

2022

Fundação Educacional do Município de Assis
Instituto Municipal de Ensino Superior de Assis
Campus "José Santilli Sobrinho"

GUSTAVO HENRIQUE MELO SOUZA

MINERAÇÃO DE EMOÇÕES: ALGORITMO PARA DETECTAR PENSAMENTOS DEPRESSIVOS EM REDES SOCIAIS.

Projeto de pesquisa apresentado ao curso de Ciência da Computação do Instituto Municipal de Ensino Superior de Assis – IMESA e a Fundação Educacional do Município de Assis – FEMA, como requisito parcial à obtenção do Certificado de Conclusão.

Orientando(a): Gustavo Henrique Melo Souza

Orientador(a): Alex Sandro Romeo de Souza Poletto

Assis/SP

2022

FICHA CATALOGRÁFICA

S729m Souza, Gustavo Henrique Melo.

Mineração de emoções: algoritmo para detectar pensamentos depressivos em redes sociais / Gustavo Henrique Melo Souza – Assis, SP: FEMA, 2022.

62 f.

Trabalho de Conclusão de Curso (Graduação) – Fundação Educacional do Município de Assis – FEMA, curso de Ciência da Computação, Assis, 2022.

Orientador: Prof. Dr. Alex Sandro Romeo de Souza Poletto.

1. Mineração de Texto. 2. Descoberta de Conhecimento. 3. Mineração de Dados. 4. Classificador. I. Título.

CDD 005

Biblioteca da FEMA

MINERAÇÃO DE EMOÇÕES: ALGORITMO PARA DETECTAR PENSAMENTOS DEPRESSIVOS EM REDES SOCIAIS.

GUSTAVO HENRIQUE MELO SOUZA

Trabalho de Conclusão de Curso apresentado ao Instituto Municipal de Ensino Superior de Assis, como requisito do Curso de Graduação, avaliado pela seguinte comissão examinadora:

Orientador: Alex Sandro Romeo de Souza Poletto

Examinador: Diomara Martins Reigato Barros

Assis/SP
2022

DEDICATÓRIA

Dedico este trabalho aos meus pais.

AGRADECIMENTOS

Agradeço especialmente a minha família por me apoiar em todos os momentos da minha vida.

A todos que diretamente ou indiretamente adjuram-me nesta jornada acadêmica e a finalização desta monografia.

Ao meu professor e orientador Alex Sandro Romeo de Souza Poletto que me apoiou e auxiliou no desenvolvimento deste trabalho.

A médica e psiquiatra Juliane de Souza Cavassana, pelas orientações na definição das bases de dados dos termos depressivos e não depressivos.

Quando achamos a matemática e a física teórica muito difíceis, voltamo-nos para o misticismo.

Stephen Hawking

RESUMO

Levantamentos realizados por organizações da área da saúde apontam que os índices de pessoas diagnosticadas com depressão aumentaram tanto no mundo quanto no Brasil, já sendo possível ser observado em diversos locais, principalmente na Internet. Diariamente são produzidos números assustadores de dados, principalmente no formato desestruturado, composto em grande parte por textos que, por sua vez, são constantemente gerados por redes sociais. Apenas observando é possível traçar a relação das redes sociais com pessoas depressivas, com a era da tecnologia cada vez mais as pessoas estão perdendo contato físico e optando pelo contato virtual, logo surge um crescimento assustador de pessoas com tendências depressivas encontradas em comentários de redes sociais e que por se tratar de um ambiente de descontração e socialização, acabam por deixar suas mágoas registradas somente de forma online. A Mineração de Texto surge como uma estratégia que propõe ferramentas e métodos derivados e adaptados da Mineração de Dados que objetiva extrair, tratar, analisar e gerar conhecimento a partir de coleções de bases textuais, frases ou apenas palavras descritas em linguagem natural. Algoritmos classificadores surgem como um objeto preditor que, através de fórmulas matemáticas e códigos, são capazes de definir a categoria um de determinado texto com base em seus conhecimentos previamente definidos e supervisionados.

Palavras-chave: Mineração de Texto; Descoberta de Conhecimento; Mineração de Dados; Classificador.

ABSTRACT

Surveys carried out by health organizations indicate that the rates of people diagnosed with depression have increased both in the world and in Brazil, and it is already possible to be observed in several places, mainly on the Internet. Scary numbers of data are produced daily, mainly in the unstructured format, largely composed of texts that, in turn, are constantly generated by social networks. Just by observing it is possible to trace the relationship of social networks with depressive people, with the age of technology more and more people are losing physical contact and opting for virtual contact, soon there is a frightening growth of people with depressive tendencies found in social network comments and because it is an environment of relaxation and socialization, they end up leaving their grievances recorded only online. Text Mining emerges as a strategy that proposes tools and methods derived and adapted from Data Mining that aims to extract, treat, analyze and generate knowledge from collections of textual bases, phrases or just words described in natural language. Classifier algorithms appear as a predictor object that, through mathematical formulas and codes, are able to define category one of a given text based on their previously defined and supervised knowledge.

Keywords: Text Mining; Knowledge Discovery; Data Mining; Classifier.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	Application Programming Interface
BI	Business Intelligence
CEP	Código de Endereçamento Postal
CRISP-DM	Cross Industry Standard Process for Data Mining
DM	Data Mining
IDE	Integrated Development Environment
JSON	JavaScript Object Notation
KDT	Knowledge Discovered in Texts
MD	Mineração de Dados
NLTK	Natural Language Toolkit
OMS	Organização Mundial de Saúde
PDF	Portable Document Format
PLN	Processamento de Linguagem Natural
URL	Uniform Resource Locator
XML	eXtensible Markup Language

SUMÁRIO

1. INTRODUÇÃO

Mineração de Texto ou Text Mining é um processo de descoberta de conhecimento, que por meio de ferramentas computacionais, automatizam a análise e extração de informações a partir de coleções textuais, frases ou apenas palavras (MORAIS; AMBRÓSIO, 2007), sendo uma subárea da Mineração de Dados (Data Mining - DM). Segundo (AMARAL, 2016, p. 2), “mineração de dados são processos para explorar e analisar grandes volumes de dados em busca de padrões, previsões, erros, associações entre outros”.

A mineração pode ocorrer em textos livres, ou seja, textos que não seguem uma estrutura fixa, podendo ser um PDF, e-mail, livros, revistas, jornais e os famosos campos de textos encontrados em fóruns e redes sociais. Também pode ocorrer mineração em um texto semiestruturado que possui algumas regras para a organização das informações, os mais conhecidos são textos XML e JSON que possuem estruturas em *tags* e em atributos/valores respectivamente.

Atualmente a mineração de texto serve de suporte para diversas áreas do conhecimento, sendo usada no direito (sentenças judiciais), na construção de artigos científicos (intuito de filtrar tópicos pesquisados), relatórios de finanças e no marketing (classificação de opinião dos clientes) que, segundo Provost e Fawcett (2016, p. 33), provavelmente, é a área em que mais se aplica técnicas de mineração de dados. Mesmo com todas essas possibilidades de implementações, há pouco conteúdo que disserte sobre o uso de mineração de texto para capturar emoções depressivas em redes sociais e se existe alguma chance de tratar os dados filtrados para ajudar essas pessoas, cenário que pretendo explorar neste texto.

As redes sociais proporcionam um espaço para pessoas divulgarem seus sentimentos, pensamentos e opiniões, espaço esse utilizado para publicações alegres e momentos de descontração, mas, também usado para expelir dores e sentimentos amargos, que muitas vezes são pensamentos reprimidos ocasionados por um simples dia ruim ou, às vezes causados por uma doença crônica mental, também conhecida como depressão.

Segundo a Organização Mundial da Saúde (OMS), existem mais de 300 milhões de casos de depressão confirmados, aproximadamente 4,4% da população mundial, sendo 5,1% mulheres e 3,6% homens. Cerca de 96,8% dos casos de suicídio estão associados a transtornos psiquiátricos, dos quais quase 36% podem ser por transtornos de humor,

conjunto de sintomas que alteram a afetividade das pessoas e da qual a depressão está integrada, o que torna ainda mais importante a discussão sobre como pode-se usar a mineração de texto para classificar pensamentos depressivos em redes sociais. Em um cenário de sucesso pode-se salvar vidas.

O presente trabalho está disposto da seguinte forma:

- **Capítulo 1 –** São apresentadas a Introdução seguido dos Objetivos, Justificativas, Motivação e Perspectivas de Contribuição.
- **Capítulo 2 –** São apresentados alguns números sobre depressão.
- **Capítulo 3 –** São apresentados os conceitos referentes à Ciência de Dados.
- **Capitulo 4 –** São apresentados conceitos e métodos da Mineração de Dados.
- **Capitulo 5 –** São introduzidos conceitos fundamentais para o entendimento da descoberta de conhecimento em textos, apresentando o PLN (Processamento de Linguagem Natural), tarefas e métodos para minerar textos.
- **Capítulo 6 –** É apresentado o Teorema de Bayes e, posteriormente, o algoritmo Naive Bayes.
- **Capítulo 7 –** É apresentando a Proposta de Trabalho por meio de um organograma.
- **Capítulo 8 –** É apresentado a metodologia utilizada para construção do classificador e seus resultados.
- **Capítulo 9 –** É apresentado um resumo do tema, além da Problemática, Hipótese, Resultados, Trabalhos Futuros e Contribuições.

1.1. OBJETIVO

Este trabalho tem por objetivo geral desenvolver um algoritmo de mineração de textos, a princípio utilizando a rede social Twitter e sua Interface de Programação de Aplicação (API) para Python, com o intuito de identificar, filtrar e classificar os textos obtidos rotulando-os em depressivos ou não depressivos usando pesquisas oficiais sobre padrões em linguagem depressiva. Por fim, apresentar e discutir algumas possibilidades para uso dos dados com finalidade de ajudar as pessoas com comportamentos depressivos.

1.2. JUSTIFICATIVAS

A realização deste trabalho é considerada importante e relevante, já que se trata de um tema pouco comentado em língua portuguesa que abrange um problema social presente no mundo todo, a depressão, além de envolver Ciência de Dados uma das áreas mais bem faladas e estudadas da tecnologia no momento atual. Espera-se que este trabalho inspire mais pessoas a estudarem o assunto e criarem novas tecnologias com a mesma finalidade ou usarem a mesma metodologia aplicada para resolução de outro problema.

1.3. MOTIVAÇÃO

Fatores pessoais, profissionais e acadêmicos contribuíram ainda mais para inflamar o entusiasmo pelo tema tratado. A doença titulada como “mal do século” é um padrão encontrado em pessoas com tendências suicidas que, conforme a OMS (Organização Mundial de Saúde) é a segunda maior causa de morte entre jovens de 15 a 29 anos, com números crescentes a cada ano, estudiosos buscam cada vez mais descobrirem novas formas de detectar a doença e estimular a vontade das pessoas em procurarem ajuda profissional e por fim tratar a depressão de forma mais eficiente utilizando novas tecnologias e tratamentos. Somando a preocupação pessoal com a doença, ao interesse de aprender mais sobre Ciência de Dados e a realização de um Trabalho de Conclusão de Curso (TCC) surge a motivação pela realização deste trabalho.

1.4. PERSPECTIVAS DE CONTRIBUIÇÃO

Pretende-se com este trabalho, agregar a área de Ciência de Dados e simultaneamente acrescentar uma ideia de solução para detecção de pessoas com pensamentos depressivos compartilhados em redes sociais. Embora haja outros artigos e trabalhos sobre o assunto tratado, há pouca resposta na prática, tendo em vista que poucas redes sociais possuem mineração de texto com tal finalidade. Espera-se que este e outros trabalhos que virão sejam notados pelas empresas donas das redes sociais e que apliquem uma solução para ajudar pessoas com depressão.

2. DEPRESSÃO

Segundo a OMS, a Depressão ou Transtorno Depressivo Maior, nomenclatura descrita no DSM-5 (Manual Diagnóstico e Estatístico de Transtornos Mentais), é um transtorno mental que aflige homens e mulheres de todas as idades, tendo como principais sintomas, sentimentos de tristeza, perda de interesse ou prazer, sentimento de culpa ou baixa autoestima, sono e apetite alterados, cansaço, falta de concentração e ideação suicida.

Um levantamento realizado em 2021 pela Vigitel (Vigilância de Fatores de Risco e Proteção para Doenças Crônicas por Inquérito Telefônico) revelou que 11,3% dos brasileiros foram diagnosticados clinicamente com depressão e apontou que a diferença entre a frequência da doença nas mulheres era de 7,4% maior que a dos homens, representando 14,7% da população enquanto os homens 7,3%. Segundo o levantamento de 2017 da OMS a depressão afetava 5,8% da população brasileira, o que indica um aumento de 5,5% em quatro anos. A Vigitel ainda revela em sua pesquisa que no ano de 2021 teve mais diagnósticos de depressão do que de diabetes.

Com números tão alarmantes, a depressão já pode ser observada em diversos lugares como escolas, trabalhos e obviamente na Internet. A Internet vem se tornando um lugar de refúgio para muitas pessoas que precisam de alguma forma expelir seus sentimentos negativos, na maioria das vezes os desabafos surgem através de textos publicados em redes sociais. Entretanto, as maiorias das pessoas não se atentam aos próprios sinais depressivos e o quanto um comentário feito por elas podem indicar seu estado mental. Sabendo disso, novas tecnologias são criadas para auxiliar a análise desses comentários a fim de ajudá-las, e possivelmente revelar um diagnóstico precoce de depressão.

Segundo Gattaz (2014), presidente do Instituto de Psiquiatria do HCFMUSP, é importante identificar os sinais depressivos o mais cedo possível visto que, quanto antes se inicia o tratamento melhor será o prognóstico, consequentemente tornando o tratamento mais eficiente.

2.1. TRISTEZA X DEPRESSÃO

“A distinção primordial entre a tristeza e a depressão melancólica está relacionada aos diferentes níveis que elas ocupam na vida psíquica do sujeito” (SOUZA; MOREIRA, 2018, p. 117). Enquanto a tristeza está relacionada a um sentimento, o transtorno depressivo

maior está diretamente ligado ao humor do indivíduo, influenciado por inúmeros fatores como, ambientais, genéticos, alterações químicas no cérebro, traumas no qual a tristeza pode estar presente, dentre outros.

Segundo Viscott (1982, p. 11) “Os sentimentos são nossa reação ao que percebemos”, ou seja, o sentimento de tristeza está diretamente atrelado a situações presenciadas no dia a dia por um curto período de duração, como uma notícia de demissão por exemplo, logo não satisfaz o conceito de depressão maior descrito no DSM-5 (2014, p. 155) “caracterizado por episódios distintos de pelo menos duas semanas de duração (embora a maioria dos episódios dure um tempo consideravelmente maior) envolvendo alterações nítidas no afeto, na cognição e em funções neurovegetativas, e remissões interepisódicas”.

3. CIÊNCIA DE DADOS

Termo originado em 1960, Data Science (Ciência de Dados), surge apenas como um sinônimo de Ciência da Computação, se tornando um campo de estudo multidisciplinar somente depois do estrondo do Big Data (1990 - 2000) e dos Bancos de Dados Relacionais (1980 - 1990). Algumas fontes idealizam que seu advento foi logo após William S. Cleveland publicar um artigo em 2001 intitulado “Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics” no qual descreve a importância da ciência de dados no futuro e os problemas que os bacharéis em estatística e análise de dados terão se tais curso não incluírem Data Science em suas disciplinas, além de propor mudanças curriculares com base em estudos de caso.

Estatística, Matemática, Ciência da Computação, Inteligência Artificial, Mineração de Dados e Big Data são algumas das áreas do conhecimento que integram a Ciência de Dados e que em alguns casos se comunicam diretamente para se obter um resultado, nesse cenário, graças a combinação de conhecimento necessário para se atingir um objetivo ela acaba se tornando também uma área interdisciplinar, ou seja, uma área de estudo se relaciona com outra.

Devido a enorme ligação com outros campos de estudos, até hoje há dificuldades para se definir Ciência de Dados ou o que é um cientista de dados, tendo em vista que, um profissional poderá denominar-se cientista de dados possuindo apenas uma especialização ou várias (GRUS, 2016, p. 26). Grus (2016, p. 26) também menciona, “... um cientista de dados é alguém que sabe mais sobre estatística do que um cientista da computação e mais sobre ciência da computação do que um estatístico ...”.

Segundo Escovedo e Koshiyama (2020, p. 22), “Data Science não é uma ferramenta, mas sim um conjunto de métodos com o objetivo de apoiar decisões de negócio baseadas em dados”. Em outras palavras, Ciência de Dados é um aglomerado de regras e técnicas que buscam compreender quantidades massivas de informações tornando o processo automatizado (PROVOST; FAWCETT, 2016, p. 40) visando dar algum valor importante para tais informações. Os dados são retirados de diferentes fontes e sistematizados podendo ser estruturados (ex: banco de dados, planilhas de Excel), não estruturados (ex: fotos, vídeos) e semiestruturados (ex: XML, JSON) com o objetivo final de corroborar nas tomadas de decisões (ESCOVEDO; KOSHIYAMA, 2020, p. 22).

Com o aumento massivo de informações, técnicas antes utilizadas por grandes empresas de tecnologia e comércio foram aos poucos substituídas por Ciência de Dados e Algoritmos de Machine Learning, visto que, grande parte do trabalho realizado de marketing, soluções e análises estatísticas era manual ou contava com pouco poder computacional, além de um número muito menor de dados coletados e tratados que possuíam. Com as ferramentas certas e automatizando processos, diversas empresas foram capazes de mudar seus planos de negócios e criarem novas estratégias para alcance de público, vendas e injeção de verba em setores carentes apenas utilizando técnicas para descobrir padrões em *feedbacks* de seus clientes. A franquia McDonald's possui diversos casos envolvendo Big Data e Data Science, dentre eles, a empresa foi capaz de identificar com algoritmos preditivos quais os dias e horários em que é necessário ter mais funcionários trabalhando para que o drive-thru trouxesse menos insatisfação para os clientes devido a enormes filas. Outro caso envolvendo o McDonald's decorreu quando eles decidiram, graças à coleta de dados de seus clientes, deixar o cardápio de café da manhã durante todo o dia, tornando tal mudança responsável por 3,7% das vendas nos Estados Unidos.

Apesar de ser muito popular entre empresas que visam impulsionar seus negócios transformando dados em *insights* determinantes para seus propósitos, as técnicas de ciência de dados também são utilizadas em outros diversos ramos, como agricultura, saúde, segurança e até mesmo na política. Grus (2016, p. 27), comenta que em 2012, grande parte do sucesso da reeleição de Obama deve-se ao presidente ter contrato diversos cientistas de dados para minerar informações ao seu favor, coletando dados cruciais de eleitores que necessitavam de mais atenção, bem como otimizar seus programas e recursos de captação de fundos de doadores específicos.

Sendo assim, a ciência de dados tem a capacidade de sistematizar problemas analíticos de dados e trabalhar cada segmento de modo independente em que cada tarefa possui técnicas e ferramentas específicas disponíveis (PROVOST; FAWCETT, 2016, p. 67).

4. MINERAÇÃO DE DADOS

Enquanto ciência de dados se comporta como o cérebro de toda a operação, lidando com conceitos multidisciplinares, a mineração de dados (MD) é uma dentre várias de suas ferramentas utilizadas para concluírem seus objetivos. Minerar dados se trata de vários processos com estágios muito bem definidos, no qual sua tarefa global é encontrar padrões a partir de dados (PROVOST; FAWCETT, 2016, p. 65) e relacioná-los.

Silva, Peres e Boscarioli (2016, p. 7), conceituam o processo de mineração de dados como:

> Trata-se, portanto, da aplicação de técnicas, implementadas por meio de algoritmos computacionais, capazes de receber, como entrada, um conjunto de fatos ocorridos no mundo real e devolver, como saída, um padrão de comportamento, o qual pode ser expresso, por exemplo, como uma regra de associação, uma função de mapeamento ou a modelagem de um perfil.

Castro e Ferrari (2016, p. 25) propõe uma analogia ao garimpo de minérios no qual o foco está em extrair o máximo de pedras preciosas possíveis. A ideia central é de que a mina seria uma base de dados volumosa, as ferramentas se comparavam aos algoritmos e técnicas e os minerais precisos obtidos seriam os conhecimentos adquiridos.

4.1. MINERAÇÃO DE DADOS X CIÊNCIA DE DADOS

Comumente ambos os termos são confundidos em momentos que se busca relacionar técnicas, metodizar etapas ou até explicar a teoria por trás de seus objetivos, entretanto, saber distinguir suas diferenças é essencial para extrair o máximo de seus potenciais singulares, embora não seja totalmente um erro intercambiar as duas áreas.

Em poucas palavras, ciência de dados é um conglomerado de princípios substanciais que buscam sistematizar a extração de conhecimento a partir de dados, enquanto mineração de dados é o uso de ferramentas das quais incorporam tais princípios (PROVOST; FAWCETT, 2016, p. 34).

Por ser uma subárea da ciência de dados seus aspectos multidisciplinares e interdisciplinares se mantém (CASTRO; FERRARI, 2016, p. 28), ampliando-se dessa forma ainda mais a relação com outras áreas de estudos, além das que norteiam a Data Science. Dentro do enorme escopo da mineração de dados encontra-se um conjunto menor de conceitos fundamentais que integram a ciência de dados, sendo eles elementos gerais que

envolvem a essência da mineração de dados e a análise de negócios (PROVOST; FAWCETT, 2016, p. 63).

Figura 1: Multidisciplinaridade da MD (CASTRO; FERRARI, 2016, p. 28).

4.2. CRISP-DM: METODOLOGIA PARA MINERAR DADOS

Compreender o processo de minerar dados influenciará significativamente a probabilidade de um projeto ter sucesso, além de fixar conceitos importantes o CRISP-DM (Cross Industry Standard Process for Data Mining ou Processo Padrão Inter-Indústrias para Mineração de Dados) demonstra de forma estruturada etapas consistentes, objetivas e repetitivas (PROVOST; FAWCETT, 2016, p. 79) do processo para extração de conhecimento dos dados. O refinamento dos dados obtém-se da exaustão frequentativa de cada etapa do processo.

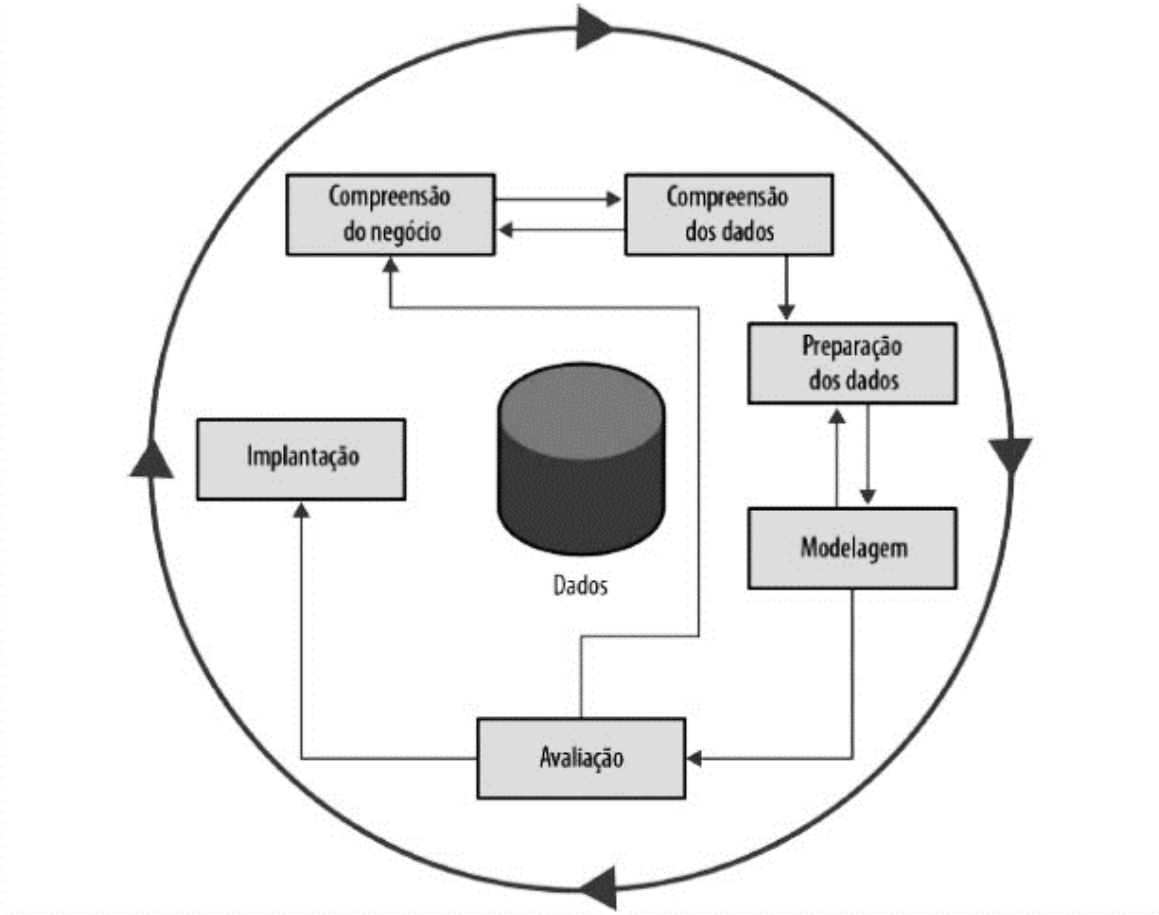

Figura 2: O processo da MD CRISP-DM (PROVOST; FAWCETT, 2016, p. 79).

4.2.1. COMPREENSÃO DO NEGÓCIO

Interpretar as necessidades e dificuldades, os problemas e objetivos são práticas cruciais do processo de mineração de dados. Embora pareça uma etapa óbvia, muitos projetos levam tempo para desenrolar-se devido à deficiência da percepção do que se almeja alcançar com o resultado final. Trata-se de compreender os requisitos fundamentais que atenderão ao propósito desejado durante todo o processo. Os requisitos variam dependendo da complexidade do projeto, entretanto, deve-se atentar previamente em verificar a disponibilidade de ferramentas e técnicas para a resolução da demanda, possuir profissionais capacitados e acima de tudo assimilar eximiamente os interesses das partes, “empresas onde os empresários não compreendem o que os cientistas de dados estão fazendo acabam em substancial desvantagem, porque perdem tempo e esforço, ou pior, porque, em última análise, tomam as decisões erradas.” (PROVOST; FAWCETT, 2016, p. 80).

Por ser uma das etapas mais importantes é uma das mais revisitadas do processo e que demanda grande criatividade dos envolvidos (PROVOST; FAWCETT, 2016, p. 80). “A formulação inicial pode não ser completa ou ideal, de modo que diversas repetições podem ser necessárias para que uma formulação de solução aceitável apareça” (PROVOST; FAWCETT, 2016, p. 80).

4.2.2. COMPREENSÃO DOS DADOS

Entender os dados e sua essência é de suma importância para as próximas etapas, é necessário planejar e combinar técnicas para análise, documentação e organização dos mesmos. Obtenção, custos e integridade são algumas das complicações mediante a compreensão dos dados (PROVOST; FAWCETT, 2016, p. 81).

Como e de onde serão coletados os dados, em que lugar serão armazenados, de quantas fontes serão coletados ou a existência dos próprios dados (eventualmente necessita-se criar os dados) são passos para obtenção dos dados.

Os custos para obtenção dos dados variam, alguns simplesmente serão gratuitos enquanto outros serão cobrados, além de existir dados que exigirão maior empenho aquisitivo (PROVOST; FAWCETT, 2016, p. 81).

Identificar a qual estrutura pertencem os dados e seu volume, realizar análises de viés qualitativos e quantitativos além de documentá-los são passos que garantem a integridade dos dados, ou seja, confiabilidade e consistência. “Mesmo depois de todos os conjuntos de dados serem adquiridos, conferi-los pode exigir esforço adicional” (PROVOST; FAWCETT, 2016, p. 81) tornando essa etapa ainda mais passível de confusões e erros significativos.

4.2.3. PREPARAÇÃO DOS DADOS

Embora coletados e conferidos, muitos dados apresentam inconsistências ou particularidades indesejáveis possivelmente revertidas como valores em branco, variáveis desnecessárias, formatos e caracteres especiais ou numéricos. Portanto, preparar os dados significa aumentar sua confiabilidade para construção do modelo futuro.

Algumas técnicas e algoritmos são utilizados nessa fase para transformar os dados de acordo com os problemas encontrados, espaços em brancos em colunas de uma tabela de

banco de dados por exemplo, são padronizados por outros símbolos ou até mesmo excluídos do modelo final, tipos de variáveis incorretas também podem ser tratadas, por exemplo, valores inteiros atribuídos como *strings*. Vale ressaltar que independentemente da limpeza realizada no conjunto a característica dos dados jamais poderá ser alterada.

4.2.4. MODELAGEM

Refere-se à aplicabilidade das técnicas e algoritmos decididos na etapa de compreensão do negócio para modelar os dados, logo dependendo das dificuldades encontradas é necessário retornar à primeira fase e reanalisar seus requisitos. “É importante ter alguma compreensão das ideias fundamentais de mineração de dados, incluindo os tipos de técnicas e algoritmos existentes...” (PROVOST; FAWCETT, 2016, p. 85).

Por apresentar seus primeiros resultados é comum haver repetições em demasia entre a etapa anterior e a de modelagem, tal processo é descrito em algumas fontes como “calibragem”. Em algumas ocasiões cria-se mais de um modelo utilizando diferentes tecnologias e métodos, os resultados obtidos são comparados e avaliados para identificar qual modelo apresenta mais harmonia com aquilo que se espera.

4.2.5. AVALIAÇÃO

Ainda que todas as fases sejam conferidas e avaliadas, esse procedimento é individual e não se atenta muito aos resultados alheios, portanto cabe à fase de avaliação gerar um panorama dos resultados como um todo e observar se os objetivos foram atendidos como esperado. “Gostaríamos de ter a confiança de que os modelos e padrões extraídos dos dados são regularidades verdadeiras e não apenas idiossincrasias ou anomalias de amostra” (PROVOST; FAWCETT, 2016, p. 85).

Caso os resultados apresentados sejam indesejados ou não atendam às expectativas do projeto, os responsáveis identificarão as inconsistências e suas respectivas etapas encarregadas de suas funções não cumpridas, em algumas situações o processo retornará direto para a etapa defeituosa não sendo necessário partir do início. “A fase de avaliação pode revelar que os resultados não são bons o suficiente para implantação, e precisamos ajustar a definição do problema ou obter dados diferentes” (PROVOST; FAWCETT, 2016,

p. 89).

Deve-se atentar principalmente em avaliações realizadas com dados inventados ou manipulados em laboratório, mesmo que a satisfação do resultado seja maior que 99% ainda há o risco do projeto ser um fracasso quando colocado em prática (PROVOST; FAWCETT, 2016, p. 85) isso porque os dados são voláteis, suas características se alteram dependendo da época ou até mesmo de onde são retirados.

4.2.6. IMPLANTAÇÃO

Em geral, é nessa etapa que o modelo estabelecido é analisado e colocado no ambiente de produção final. “Os casos mais claros de implantação envolvem a implementação de um modelo preditivo em algum sistema de informação ou processo de negócios” (PROVOST; FAWCETT, 2016, p. 87).

O modelo construído é avaliado outra vez com um novo conjunto de dados em um cenário real para qual foi desenvolvido e intencionado desde o início. Mesmo que o modelo implantado tenha êxito é comum retornar a fase de entendimento do negócio diversas vezes, tendo em vista que uma nova iteração minuciosa pode gerar uma solução ainda melhor para aperfeiçoar o modelo existente (PROVOST; FAWCETT, 2016, p. 89).

5. MINERAÇÃO DE TEXTOS

A Descoberta de Conhecimento em Textos (Knowledge Discovery in Texts - KDT) ou simplesmente Mineração de Texto, é um processo automatizado multidisciplinar que se baseia principalmente nos conceitos e técnicas da Mineração de Dados afim de extrair informações potencialmente relevantes e valiosas a partir de enormes bases de dados de documentos textuais, visando sobretudo fontes de dados desestruturados e semiestruturados. Todavia, convém evidenciar que a descoberta de conhecimento em texto não se limita somente a técnicas convencionais procedentes da mineração de dados, mas também consiste em criar novas técnicas específicas para o contexto em que os dados estão relacionados, além de adaptar diversas metodologias já suportadas provenientes da MD (WIVES, 2004, p. 24). O conhecimento gerado através do KDT não só atrai cientistas de dados como também desperta interesse em empresas que buscam melhorar seu posicionamento no mercado (GONÇALVES, 2012).

Ribeiro (2018, p. 57), descreve KDT da seguinte forma, “Compreende técnicas e ferramentas automáticas e inteligentes, responsáveis pelo auxílio na análise de grandes volumes de dados, com propósito de minerar conhecimento útil, aplicado a domínios que utilizem textos não estruturados”.

Para Corrêa et al., (2012, p. 1), “A Mineração de Textos é um processo que busca descobrir conhecimento útil a partir de coleções textuais, o que viabiliza sobremaneira a análise exploratória de documentos científicos”.

Estima-se que até em 2025 oitenta por cento (80%) dos dados mundiais serão desestruturados (KING, 2019), categoria a qual os documentos textuais pertencem, além de ser o tipo de dado que mais chama atenção das empresas inseridas no contexto tecnológico e econômico atual. No cenário corporativo os dados não estruturados já equivalem a 80% ou mais de todos os dados (SMALLCOMBE, 2022). Da perspectiva estatística, esses percentuais tendem a crescer pouco a pouco até que os dados estruturados se tornem mínimos, existindo apenas para cumprirem seus papéis básicos como esquematizar informações em linhas e colunas e de suprirem necessidades rotuladoras, ou seja, etiquetar os dados em campos predefinidos em tipos e tamanhos. Do ponto de vista comercial, empresas que não utilizam dados desestruturados para colaborar nas tomadas de decisões e impulsionar seus resultados estão em desvantagem no quesito inteligência de negócio (Business Intelligence - BI). “Com grandes quantidades de dados

disponíveis, as empresas em quase todos os setores estão focadas em explorá-los para obter vantagem competitiva” (PROVOST; FAWCETT, 2016, p. 33).

Independentemente dos benefícios obtidos, minerar documentos de texto não é uma tarefa fácil, adversidades complexas são encontradas em todas as etapas do processo exigindo coordenação e cooperatividade pluridisciplinar, abrangendo principalmente áreas como aprendizado de máquina, processamento de linguagem natural e estatística. A indisposição de uma estrutura inflexível e organizada gera dificuldades para análise dos documentos desde o primeiro momento do processo, não há nenhuma ferramenta global que trate os documentos textuais como as que organizam os dados nos famosos bancos de dados relacionais, sendo assim, novas soluções surgem para metodizar o processo de KDT dependendo do contexto em que os textos estão inseridos. Gonçalves (2012), adita outro desafio quanto a mineração de texto:

> Além disso, para tornar a coisa mais difícil, quando se trabalha com texto, é preciso conviver com uma série de problemas complicados para algoritmos computacionais, problemas estes inerentes aos processos de interpretação de texto: existência de sinônimos, erros ortográficos, diversidade de idiomas, recursos estilísticos (metáfora, metonímia, ironia, etc.), entre outros.

Complementando o raciocínio de Gonçalves, ainda há problemas com ambiguidade lexical ou estrutural de palavras, além de ser comum e predominante o uso de expressões no sentido figurado dependendo da fonte de onde são retirados os documentos. Santos (2015, p. 5), diz que o processo como um todo para geração de conhecimento através de dados desestruturados são mais complexos que o mesmo processo para dados estruturados, para ele, obter, tratar e analisar informações em formatos de texto são passos bem mais complicados do que quando realizados com informações rotuladas, por fim ele utiliza uma tabela e uma figura para demonstrar informações semelhantes só que em estruturas diferentes.

Nome	Idade	Morada	Telefone
Cédric	25	Coimbra	910000000
José	29	Viseu	918888888

Tabela 1: Informação estruturada (SANTOS, 2015, p. 6).

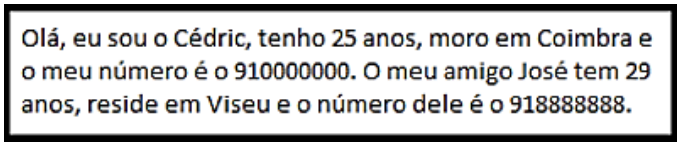

Figura 3: Informação não estruturada (SANTOS, 2015, p. 6).

Portanto, para solucionar os problemas mediante a descoberta de conhecimento em texto, há duas abordagens: a Análise Semântica e a Análise Estatística.

Análise Semântica	Análise Estatística
Ciência Cognitiva	Recuperação de Informação
Processamento de Linguagem Natural	Estatística
Mineração de Dados	Aprendizado de Máquina
Web Mining	Inteligência Computacional
	Mineração de Dados
	Web Mining

Tabela 2: As duas abordagens para a Análise de Textos e suas principais Áreas de Conhecimento (JUNIOR, 2007, p. 15).

Análise Semântica: "...apoia-se no tratamento de textos conforme o ser humano faz, através do significado das palavras, de conhecimentos morfológicos, sintáticos, semânticos, pragmáticos e do contexto em geral" (JUNIOR, 2007, p. 14). Para Gonçalves (2012), "...baseia-se no uso de algoritmos que tentam processar os textos da mesma forma que nós, seres humanos: através da interpretação sintática e semântica das frases".

Análise Estatística: "...os termos são valorados, basicamente, pela sua frequência de aparição na massa de dados, não importando a contextualização deste, como em que parágrafo está inserindo, que termos o antecedem ou que estão diretamente relacionados" (JUNIOR, 2007, p. 14). "Na realidade ela é menos trabalhosa de ser implementada e consegue resolver de forma eficiente a maioria das tarefas de mineração de texto, como a classificação, determinação de agrupamentos, descoberta de associações e recuperação de informação" (GONÇALVES, 2012).

5.1. DOCUMENTOS TEXTUAIS

Em suma, o termo documento alude a tudo aquilo que é capaz de registrar e transmitir informações que devem ser elucidadas e apresentar algum nível de utilidade. Segundo Wives (2004, p. 18), “o documento pode ser uma pintura, uma figura, um gráfico, uma escultura, um filme ou outro objeto qualquer, desde que ele transmita informação”. Portanto, um documento de texto, é todo objeto que em sua essência contenha ao menos um elemento de natureza textual, seja qual for sua estrutura. Compondo o grupo de documentos textuais estão: arquivos PDF, XML, TXT, páginas Web, e-mails, livros, artigos online, comentários em redes sociais, campos VARCHAR em tabelas relacionais, entre outros.

Na mineração de texto a expressão *corpus*, do latim corpo, é utilizada para referenciar uma coleção de documentos, sendo assim, *corpus* é um conjunto de elementos textuais selecionados e estruturados com características e funções próprias que tem como principal objetivo servir de modelo para análise linguística.

> Em bases de dados textuais, também conhecidas como corpora, cada exemplar é tratado como um “documento”. Cada documento em um corpus pode assumir diferentes características em relação a, por exemplo, tamanho do texto (sequência de caracteres), tipo de conteúdo (assunto que aborda), língua na qual é escrito (inglês, português, japonês, árabe etc.) ou tipo de linguagem adotada (formal, coloquial, poética, irônica etc.) (SILVA; PERES; BOSCARIOLI, 2016, p. 61).

5.2. PROCESSAMENTO DE LINGUAGEM NATURAL

O termo linguagem natural diz respeito a todo meio de comunicação formal desenvolvido espontaneamente pelos seres humanos com propósito estrito de criar relações dentre comunidades, seja por meio da escrita, fala, ou sinais, inclusive, a língua de sinais é considerada natural já que apresenta todas as características necessárias, possuindo estrutura gramatical complexa e peculiaridades próprias.

Ao contrário dos seres humanos, as máquinas não possuem naturalmente a capacidade de perceber e assimilar instintivamente informações a sua volta, para tal propósito seriam necessários receptores sensoriais igualmente aos dos humanos. Entretanto, os humanos não aprendem somente através de instintos, muitas vezes os conhecimentos são

adquiridos no decorrer da vida, por meio da leitura, assimilação, interação social e de experiências repassadas culturalmente ou por professores. Com base nas possibilidades para se obter conhecimento, diversas áreas de estudos compartilham recursos e fundamentos para alcançarem o mesmo objetivo, transformar máquinas em sistemas capazes de assimilar informações tal como os seres humanos, porém com uma velocidade extraordinária.

O Processamento de Linguagem Natural (PLN), é um conglomerado de métodos e algoritmos computacionais utilizados para dar às máquinas a capacidade de analisar e manipular um determinado conjunto de elementos equivalentes à linguagem natural humana. Em virtude disso, o PLN é um estudo que visa sobretudo identificar as limitações perceptivas dos computadores assim como explorar suas capacidades para processamento de informações, mesclando princípios da Linguística, Ciência da Computação e Ciência Cognitiva.

De acordo com Lopes e Vieira (2010, p. 185), os empenhos aplicados no Processamento de Linguagem Natural estão direcionados para cinco níveis de análises: fonético ou fonológico, morfológico, sintático, semântico ou pragmático. Cabe destacar que, cada nível de análise possui funcionalidades e qualidades próprias que são utilizadas de acordo com as necessidades preestabelecidas e que atendam ao propósito do PLN (LOPES; VIEIRA, 2010, p. 185). “Por exemplo, aplicações sobre textos científicos usualmente não têm preocupação com uma análise fonológica, por outro lado, aplicações que façam uma interface com reconhecimento de voz focam esse nível de análise” (LOPES; VIEIRA, 2010, p. 185). Basicamente, a fonética e a fonologia são análises de aspectos sonoros, enquanto a morfologia e a sintaxe se atentam a características estruturais, pôr fim a semântica e a pragmática se preocupam com o significado da informação (SANTOS, 2015, p. 7).

As áreas mais famosas em que o PLN atua são: análise de sentimentos, análise de risco de crédito, assimilação de tendências, plataformas de busca, assistentes virtuais e chatbots. Ressalta-se que o uso de PLN nem sempre estará ligado à mineração de textos, pois qualquer área que busque desenvolver soluções com sistemas em linguagens naturais pode utilizar o PLN. O maior exemplo que foge aos conceitos de KDT são os tradutores, inclusive, os tradutores são os pioneiros na área, surgindo por volta de 1940 com o intuito de quebrar códigos durante a Segunda Guerra Mundial (LOPES; VIEIRA, 2010, p. 186).

Os códigos desenvolvidos a partir dos métodos de PLN, são capazes de contornar a maioria

das adversidades gramaticais morfológicas, semânticas e sintáticas, detectadas em textos escritos em linguagem natural. Partindo do conceito proposto inicialmente para o uso do processamento de linguagem natural, são necessárias regras preestabelecidas criadas estrategicamente para padronizar e estabelecer relações entre os documentos selecionados, logo, há algumas abordagens que previamente devem atender aos requisitos mínimos para avaliação do modelo final. Dentre a vasta diversidade de abordagens propostas estão: organizar e distribuir documentos textuais baseados na frequência de repetição de palavras-chave, criar dicionários de palavras indesejadas e utilizar algoritmos para excluir tais palavras dos documentos, retirar números e caracteres dos documentos caso não tenham valor estrutural e retirar palavras vazias. Outras abordagens serão mencionadas na seção “Metodologia para minerar texto”.

5.3. TAREFAS DA MINERAÇÃO DE TEXTO

Devido ao enorme potencial que a KDT disponibiliza, novos interessados no mercado surgem a todo momento com propostas empreendedoras cercadas de desafios que dependem apenas de soluções inteligentes. Portanto, para satisfazer a demanda, torna-se necessário segmentar a KDT em diversas tarefas que deem apoio à análise das informações. As tarefas que serão abordadas nesta seção são: extração da informação, descoberta por agrupamento (“*Clustering*”) e classificação de textos.

5.3.1. EXTRAÇÃO DA INFORMAÇÃO

Os procedimentos realizados na tarefa de extração de informação têm como principal objetivo transformar os dados desestruturados em estruturados, os conhecimentos relevantes são extraídos por meio de regras específicas que identificam trechos textuais chaves para posteriormente organizá-los e armazená-los em *templates* categorizados. As informações são extraídas e armazenadas desse modo para que possam ser processadas e analisadas novamente recorrendo às técnicas convencionais de mineração de dados (WIVES, 2004, p. 25).

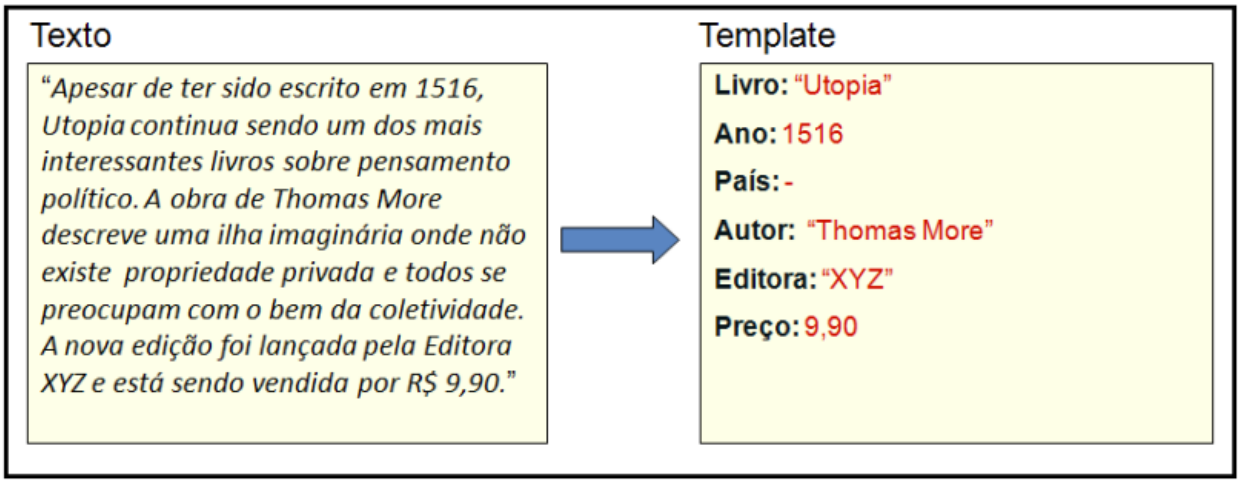

Figura 4: Extração da Informação (GONÇALVES, 2012).

Segundo Gonçalves (2012), os sistemas baseados nesta tarefa precisam ser capazes de identificar características contidas nos textos como: datas, substantivos, verbos, abreviações, endereços, URL's, caracteres que representam unidades de medidas, etc. Ele ainda complementa apresentando algumas técnicas para reconhecer elementos desejados dentro de um texto. A primeira técnica é utilizando expressões regulares, com elas a maioria das informações padronizadas como datas, telefones, CEPs e e-mails são encontradas facilmente. Outra técnica mencionada por ele é utilizar dicionários supervisionados compostos por substantivos, abreviações e acrônimos. Por fim, ele propõe regras de identificação de padrões e cita o exemplo: "se existe no texto uma sequência com duas ou mais palavras possuindo apenas a primeira letra em maiúsculo, é muito provável que se trate de um nome de pessoa ou local".

5.3.2. DESCOBERTA POR AGRUPAMENTO ("*CLUSTERING*")

Essa técnica consiste basicamente em separar documentos com alto nível de semelhança em um grupo (*cluster*). Elementos do mesmo cluster devem apresentar similaridades de conteúdos tão altas quanto clusters diferentes devem apresentar dissimilaridades, ou seja, um documento agrupado em um cluster deve ser similar aos outros documentos do mesmo grupo só que ao mesmo tempo deve ser diferente dos documentos pertencentes a outros clusters. Basicamente, o método é utilizado para descobrir padrões e associações entre documentos (WIVES, 2004, p. 27).

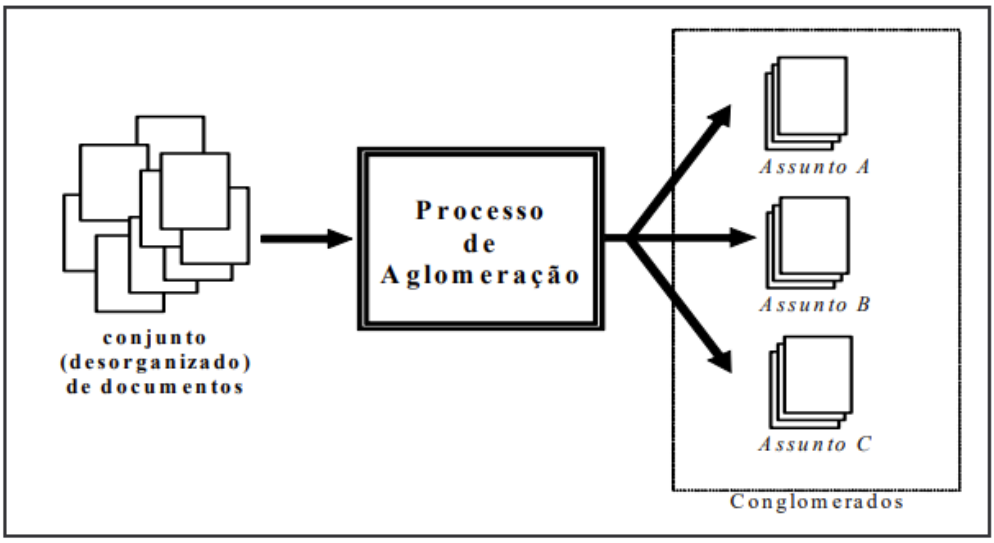

Figura 5: Objetivo do processo de agrupamento de documentos (WIVES, 2004, p. 28).

Gonçalves (2012), demonstra o processo de agrupamento utilizando arquivos XML para exemplificar a associação entre os documentos. Ele separa três arquivos, sendo que dois possuem informações relacionadas a livros e o outro arquivo relacionado a dados geográficos de um país, depois, simula como um sistema teria processado e relacionado os documentos.

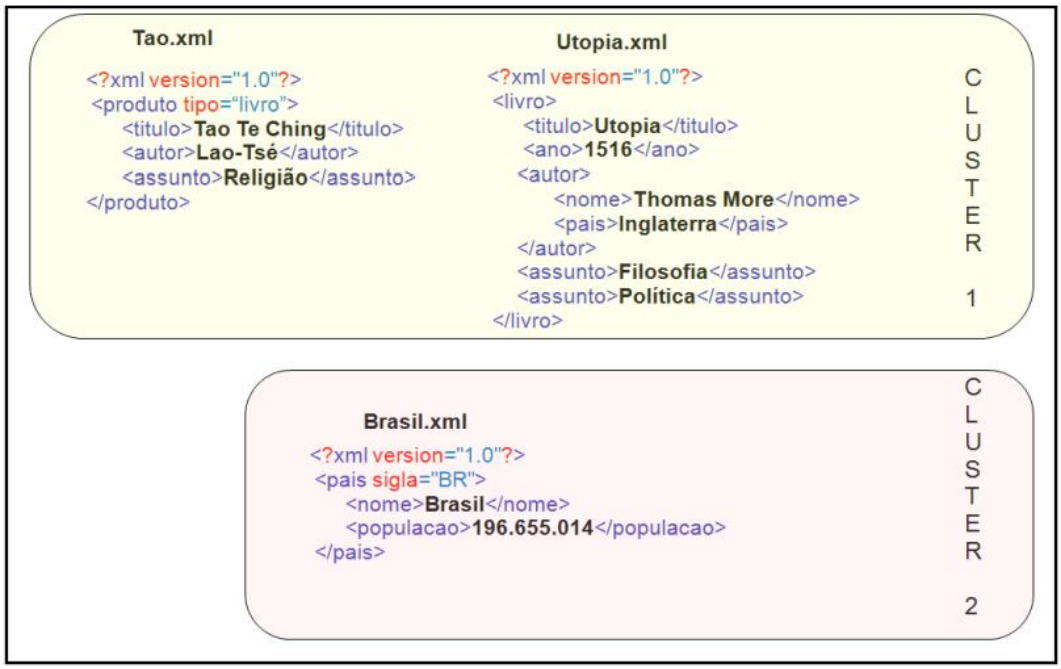

Figura 6: Agrupamento de documentos XML (GONÇALVES, 2012).

5.3.3. CLASSIFICAÇÃO DE TEXTOS

A tarefa encarregada por classificar textos é uma das mais utilizadas no KDT, por meio dela é possível criar sistemas utilizando algoritmos potentes que facilitam a categorização de dados não estruturados. Em síntese, classificar textos é um processo automático que determina a categoria (classe) de um texto em linguagem natural baseando-se em uma coleção de documentos predefinidos já classificados e supervisionados manualmente.

Basicamente, novos documentos são submetidos a um modelo estabelecido, muitas vezes treinados por processos analíticos, que mesclam aprendizado de máquina e PLN para classificar os textos em classes disponíveis que possuem mais características em comum com seu conteúdo. A filtragem de *spam* aplicada em correios eletrônicos é um exemplo claro de classificador de sucesso e que a maioria das pessoas utilizam diariamente (GONÇALVES, 2012). O classificador leva em conta o assunto e o corpo das mensagens recebidas para submetê-las a um programa composto por uma série de regras analíticas, que consideram cada fragmento da mensagem relevante para associar os padrões encontrados com os já conhecidos pelo modelo. Por fim, ao final de todo o processo, as mensagens são consideradas normais ou spams.

Figura 7: Classificador de spam (GONÇALVES, 2012).

5.4. METODOLOGIA PARA MINERAR TEXTO

Ao contrário da Mineração de Dados, o processo KDT não possui etapas rígidas para obter o conhecimento através da análise de textos, na verdade, em muitas soluções que dependem da KDT, a metodologia utilizada é criada de acordo com as necessidades, dependendo muito da criatividade dos envolvidos para esquematizar as etapas do processo. Dito isso, nessa seção será abordado o modelo proposto por Aranha (2007) em sua tese de doutorado, objetivando apresentar as etapas da Mineração de Texto de maneira resumida e integrada com ideias de outros autores.

Figura 8: Modelo do processo de mineração de textos (ARANHA, 2007, p. 19).

5.4.1. COLETA

Compreende-se por coleta, o procedimento que se preocupa com formação do *corpus*, o processo visa modelar uma coleção de documentos textuais criteriosamente selecionados, onde cada informação é apurada preocupando-se com características qualitativas e quantitativas. Denomina-se *corpus* somente coleções textuais constituídas por dados seletos, com o intuito de eventualmente servirem para análise e extração de conhecimento.

Segundo Aranha (2007, p. 41), coletar informações para minerar textos é uma das atividades mais trabalhosas, já que os dados disponíveis geralmente não estão em formatos apropriados para serem utilizados, além disso há uma dificuldade maior para descobrir onde estão armazenados. Complementando seu raciocínio, há diversos casos que a coleta terá que ser supervisionada por pessoas especialistas no assunto do conteúdo alvo, tendo em vista que o objetivo é ter um documento com o menor nível possível de ruído nas informações.

5.4.2. PRÉ-PROCESSAMENTO

Antes de submeter os documentos em processos analíticos, deve-se atentar ao seu nível de abstração, mesmo que a coleta tenha sido perfeitamente realizada, ainda há muitos elementos textuais sem relevância significativa para o processo. Basicamente, é uma etapa que se propõe aumentar a qualidade inicial dos dados (JUNIOR, 2007, p. 30). Pré-processar os dados significa subjugá-los a uma série de ferramentas tratativas que visam

padronizá-los para reapresentá-los em um formato mais adequado. O principal objetivo da etapa é preparar os dados para que sejam submetidos futuramente aos algoritmos de mineração (ARANHA, 2007, p. 43).

Nas palavras de Aranha (2007, p. 42), “o pré-processamento de textos consiste em um conjunto de transformações realizadas sobre alguma coleção de textos com o objetivo de fazer com que esses passem a ser estruturados em uma representação atributo-valor”.

Embora seja uma etapa elementar para minerar textos, deve-se tomar muito cuidado ao transformar os documentos em objetos concretos, deve-se sempre planejar rigorosamente, minimizar a perda de informações relevantes (ARANHA, 2007, p. 43). Dentre as técnicas utilizadas no pré-processamento estão: tokenização, *filtering*, *stemming* e remoção de *stopwords*.

Tokenização: Consiste em quebrar os textos em fragmentos (tokens) normalmente compostos por uma única palavra ou símbolo. Essa técnica permite tratar individualmente cada elemento de um texto, isso é possível já que a maioria das abordagens estatísticas tendem a se preocupar minimamente com a semântica (GONÇALVES, 2012).

“Zico foi o maior jogador da história!”
[Zico] [foi] [o] [maior] [jogador] [da] [história] [!]

Figura 9: Frase separada em *tokens* (JUNIOR, 2007, p. 31).

Filtering: Consiste em filtrar os elementos textuais que necessitam de algum tratamento específico ou que apenas devam ser excluídos. Dependendo do objetivo, números, caracteres, abreviações e palavras indesejadas são separadas ou simplesmente descartadas por meio de expressões regulares. É comum retirar acentos e transformar os textos em letras minúsculas caso a identificação de substantivos próprios não faça diferença.

Stemming: Consiste em reduzir cada palavra para seu estado base ou raiz, transformando suas variações em uma única representação. Basicamente, é o processo que remove das palavras os tempos verbais, gerúndios, plural, sufixos masculinas ou femininas, etc.

Palavra	Stem
saltar	salt
saltaram	salt
saltos	salt
saltei	salt
saltou	salt
saltará	salt

Tabela 3: Exemplo de Stemming (GONÇALVES, 2012).

Remoção de stopwords: Palavras vazias ou stopwords, são elementos textuais que possuem pouca relevância semântica bem como não precisam ser indexados e que por sua vez, se repetem diversas vezes em todo o documento textual. Segundo Aranha (2007, p. 44), as *stopwords* são constituídas de artigos, preposições, verbos auxiliares, advérbios, etc, tais como "que", "de/do/das", "o/os" ou "a/as", "e", "para", "com", "um". Para separar essas palavras, dicionários (*stoplist*) são criados manualmente de acordo com a língua desejada. A remoção ocorre verificando cada token do documento, nesta etapa é comum perceber palavras sem valor significativo e que não estão sendo removidas, logo deve-se atentar em preencher as *stoplists* a todo momento do processo de pré-processamento, tendo em vista que uma palavra indesejada repetida várias vezes pode influenciar negativamente e diretamente na análise final. A remoção das *stopwords* é uma das etapas que mais diminui o tamanho dimensional dos documentos, logo, ela é capaz de aumentar significativamente a velocidade de processamento nas etapas de mineração.

Stoplist				**Texto**
A	seu	De	Um	~~[A]~~[Casa&Lar]~~[se]~~[mudou]~~[.]~~[Agora]~~[,]~~
O	deve	do	uma	[atendemos] ~~[na]~~ [Av. Dom Casmurro]~~[,]~~ [nº
pelo	sua	da	sobre	200]~~[.]~~ [Você] ~~[também]~~[pode]—[comprar]
por	nosso	também	são	~~[pelo]~~ [nosso] [endereço] ~~[na]~~ [Internet]~~[.]~~
em	nossa	se	cada	[http://www.casaelar.com.br] ~~[.]~~
na	.	comigo	isso	[Comprando] ~~[por]~~ ~~[lá]~~~~[,]~~ [você] [tem] [até]
no	!	pela		[R$ 100,00] ~~[de]~~
como	;	?		[desconto]~~[em]~~[tubos]~~[e]~~[caixas d'água]~~[!]~~
lá	,	só		

Tabela 4: Identificação e Remoção de *Stopwords* (os *tokens* descartados estão tachados) (JUNIOR, 2007, p. 39).

5.4.3. INDEXAÇÃO

Embora formatados, os dados ainda não manifestam qualquer estrutura que facilite sua visualização, tanto humana quanto computacional, logo não estão aptos a serem submetidos ao processo de mineração. Portanto, busca-se realizar técnicas e utilizar ferramentas de indexação para transformar os elementos dos documentos em informações estruturadas.

A indexação de um documento é a categorização dos seus elementos com base em seus conteúdos, a fim de simplificar e acelerar sua busca e consequentemente sua recuperação. Em suma, é o processo que facilita a organização e localização de um ou mais documentos de acordo com a associação de seus atributos com os termos procurados.

Na mineração de dados, as informações geralmente estão armazenadas em tabelas estruturadas e categorizadas, logo não há tantas complexidades para recuperar os dados, basta utilizar ferramentas que pesquisem as informações pelos índices desejados, como por exemplo, pesquisar um cliente em um banco de dados, pelo nome, CPF, endereço e assim por diante. Na mineração de texto, como as informações estão desestruturadas, há uma dificuldade maior para recuperar informações, no exemplo de busca por um cliente, os dados estariam em um único bloco textual, sendo necessário percorrer caractere por caractere até encontrar as informações desejadas.

Para suprir a necessidade de indexar textos, foram criadas diversas técnicas que facilitam a categorização de documentos. Gonçalves (2012), menciona o método de índices invertidos, no qual consiste em criar uma tabela de dispersão com palavras-chave, onde cada elemento dessa tabela aponta para uma lista de documentos que possuem a palavra em questão.

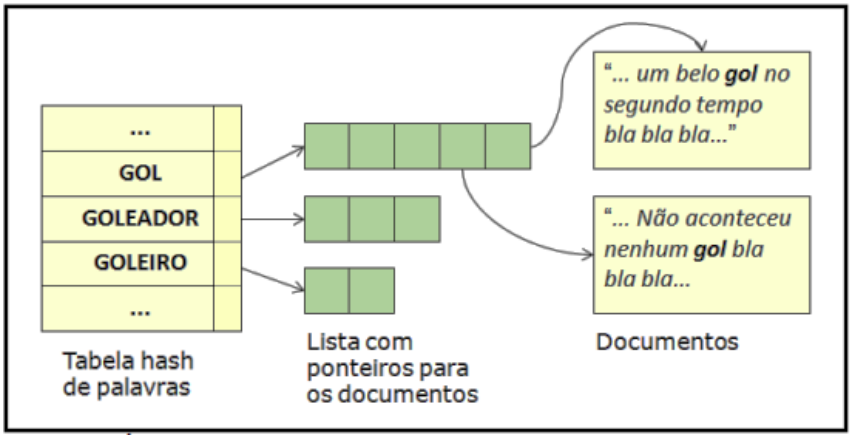

Figura 10: Índice invertido (GONÇALVES, 2012).

Segundo Junior (2007, p. 48), o Modelo de Espaço Vetorial é uma das técnicas mais utilizadas para representar documentos, tendo em vista que sua representação final atende aos requisitos estruturais de muitos algoritmos na etapa de mineração, como Naive Bayes e K-Means. Nas palavras dele os "documentos são representados como pontos em um espaço Euclidiano t-dimensional em que cada dimensão corresponde a um token do léxico". Gonçalves (2012) menciona o modelo da seguinte forma: "Considerando um conjunto de *d* documentos e um total de *n* termos distintos considerando todos esses *d* documentos, os documentos são modelados como uma matriz de dimensão *d* linhas x *n* colunas".

documentos / termos	t1	t2	t3	t4	t5
d1	6	14	0	5	15
d2	0	0	8	0	22
d3	17	10	1	0	6

Tabela 5: Matriz com a frequência de termos por documentos (GONÇALVES, 2012).

5.4.4. MINERAÇÃO

Essa etapa serve estritamente para decidir as tarefas que serão realizadas bem como os algoritmos conhecidos capazes de solucioná-las. Em síntese, significa aplicar o modelo da representação dos documentos sobre um determinado algoritmo, objetivando acima de tudo, extrair conhecimento relevante. Segundo Aranha (2007, p. 58), desde que atendam ao propósito do processo, os algoritmos escolhidos podem pertencer a qualquer área do conhecimento, por exemplo, Aprendizado de Máquina, Estatística, Redes Neurais e Banco de Dados.

Se o objetivo é associar automaticamente um determinado documento a uma classe específica, então o algoritmo escolhido deverá ser um classificador. Já se o objetivo for agrupar documentos por similaridades de conteúdo, então o algoritmo deverá ser um agrupador (*clustering*). Se a necessidade é transformar dados não estruturados em estruturados, então o algoritmo escolhido deverá ter peculiaridades em extrair informações.

Em projetos que envolvem KDT, é muito comum utilizar diversas técnicas para alcançarem o mesmo objetivo, logo, é importante avaliar e documentar cada resultado obtido, assim facilitará a escolha de um algoritmo melhor apenas observando seu desempenho.

5.4.5. ANÁLISE

Qualquer processo de descoberta de conhecimento seja de mineração de dados ou textos, são apresentados em esquemas cíclicos, ou seja, mesmo que os resultados obtidos sejam frustrantes, é possível retornar a qualquer etapa do processo e a qualquer momento, mesmo sem avaliar o modelo, e rever os pontos causadores dos problemas encontrados. Há casos que mesmo tendo avaliações satisfatórias, o processo é regredido até a primeira fase, dado que, o conhecimento extraído e avaliado, proporciona um novo panorama sobre os dados analisados, resultando em novas ideias e aprendizados.

Segundo Corrêa et al., (2012, p. 15), “a avaliação pode ser realizada de forma subjetiva, utilizando um conhecimento de um especialista de domínio, ou de forma objetiva por meio de índices estatísticos que indicam a qualidade dos resultados”.

Enfim, a etapa da análise servirá para deixar evidente se o modelo construído através de todo o processo, está pronto para sair do estudo laboratorial e partir para o uso real.

6. TEOREMA DE BAYES

O “Princípio da Probabilidade Inversa” (SIMÕES, 1981, p. 18) ou *Teorema de Bayes*, foi desenvolvido pelo matemático Thomas Bayes no Século XVIII e posteriormente publicado por Pierre-Simon, físico e matemático francês reconhecido por engendrar a equação de Laplace.

Pertencente ao ramo da estatística, o teorema busca encontrar a probabilidade de um evento ocorrer sabendo que outro evento ocorreu, portanto, é necessário ter informações prévias do evento ocorrido tendo em vista que o resultado do segundo evento é totalmente dependente do resultado do primeiro. O resultado do evento é uma predição de qual *classe* pertence o dado analisado. Segundo Rocha *et al*., (2012), se há um conjunto de dados (*dataset*) onde cada dado pertence a um rótulo e se tem conhecimento dos rótulos, então pode-se utilizar o termo *classe*.

Para Simões (1981), o teorema se baseia em encontrar a probabilidade de A ser causada por B sabendo que: há um espaço amostral formado por um conjunto *a priori* de $A's$ onde A são eventos exaustivos concomitante por eventos aleatórios $B's$.

Imagine um *dataset* que contenha dados sobre a eficiência do teste de Covid-19 no qual cada resultado positivo ou negativo pode pertencer as *classes comCovid-19* e *semCovid-19*. Ao analisar o *dataset*, percebeu-se que havia 100 testes no qual 40% representavam a classe *comCovid-19* e 60% a classe *semCovid-19*, entretanto 95% dos *comCovid-19* deram positivo e 30% dos *semCovid-19* também deram positivo. A pergunta é, qual a probabilidade de uma pessoa não estar com o vírus mesmo o teste dando positivo? Utilizando probabilidade comum chegou-se ao resultado de $\frac{18}{56}$ ou 32,14%:

$$100\ testes\begin{cases}60\%\ semCovid\begin{cases}30\%\ t.positivo\begin{cases}18\\42\end{cases}\\70\%\ t.negativo\end{cases}\\40\%\ comCovid\begin{cases}95\%\ t.positivo\begin{cases}38\\2\end{cases}\\5\%\ t.negativo\end{cases}\end{cases} = \frac{18}{18+38} = \frac{18}{56} \cong 0,3214\ \cong 32{,}14\% \quad (1)$$

Mas será que o teorema de Bayes mostra o mesmo resultado utilizando sua fórmula matemática?

Dada a fórmula:

$$P(A_{c1} \backslash B) = \frac{P(B \backslash A_{c1})\, P(A_{c1})}{P(B)} \quad (2)$$

$$P(B) = P(B \backslash A_{c1})\, P(A_{c1}) + P(B \backslash A_{c2})\, P(A_{c2}) + \cdots + P(B \backslash A_{cn})\, P(A_{cn}) \quad (3)$$

Onde:

- A_c representa cada classe do *dataset*, no caso, *semCovid-19* e *comCovid-19*.
- $P(A_c)$ representa a probabilidade *a priori* total de ocorrer cada *classe*.
- B representa o atributo (evento) em que temos conhecimento.
- $P(A_c \backslash B)$ representa a probabilidade *a posteriori* de ocorrer A_c sabendo de B, exatamente como a pergunta formulada acima.
- $P(B \backslash A_c)$ representa a probabilidade total de ocorrer B dado a classe.
- $P(B)$ representa a soma total dos produtos de $P(B \backslash A_c)\, P(A_c)$.

Logo:

- A_{c1} = *semCovid-19*, A_{c2} = *comCovid-19*.
- $P(A_{c1})$ = 60%, $P(A_{c2})$ = 40%.
- B = positivo.
- $P(B \backslash A_{c1})$ = 30%, $P(B \backslash A_{c2})$ = 95%.

$$P(semCovid \backslash \text{positivo}) = \frac{30\% \,.\, 60\%}{30\% \,.\, 60\% + 95\% \,.\, 40\%} \cong 0{,}3214 \cong 32{,}14\% \quad (4)$$

O resultado matemático mostra que a probabilidade de uma nova pessoa que testou positivo não estar com Covid-19 é de 32,14%. O teorema chegou no mesmo resultado sem ter a complexidade da interpretação dos valores dado um problema probabilístico. O exemplo dado acima é muito simples para o uso real do teorema, tendo em vista que, trabalhou-se apenas com dois atributos, há casos em que se terá vários atributos influenciando individualmente, diretamente o resultado de uma classe, logo seria inviável e praticamente impossível separar e analisar os dados manualmente usando probabilidade convencional.

O algoritmo aplicado para classificar o conjunto de dados utilizado neste trabalho é fundamentado no teorema de Bayes.

6.1. ALGORITMO DE NAIVE BAYES

Muito utilizado na área de Machine Learning o Algoritmo de Naive Bayes é um classificador, no qual utiliza conceitos de aprendizado de máquina supervisionado, ou seja, o classificador é treinado presumindo que considera confiável um conjunto de dados já analisado e classificado manualmente.

“Naïve Bayes é um algoritmo bayesiano, baseado na teoria das probabilidades e que supõe que os atributos vão influenciar a classe de forma independente” (AMARAL, 2016, p. 41). O impacto provocado pelo valor de um atributo na distribuição de classes de um conjunto de dados é independente do impacto provocado pelos valores de outros atributos sobre a mesma distribuição de classes, por isso o algoritmo recebe ingênuo (*naive*) no nome (SILVA; PERES; BOSCARIOLI, 2016). Segundo Amaral (2016), o modelo final criado pelo classificador é uma tabela mostrando a relevância de cada atributo sobre cada classe.

A aplicabilidade do algoritmo é vasta, sendo utilizado no ramo da medicina para predizer resultados de exames, por exemplo, ou até mesmo para descobrir a probabilidade de risco de empréstimo para clientes. “Estudos indicam que os algoritmos simples de classificação bayesiana, conhecidos como Naïve Bayes, possuem desempenho comparável a Redes Neurais Artificiais e Árvores de Decisão para alguns problemas” (CASTRO; FERRARI, 2016, p. 205).

Neste trabalho, será aplicado técnicas para processamento de Linguagem Natural, termo utilizado para o processo de análise da linguagem escrita ou falada dando significado para ela (ANDREATA, 2017), que servirá para preparar os dados antes da aplicação do algoritmo de Naive Bayes.

7. PROPOSTA DE TRABALHO

A proposta de trabalho tem por finalidade desenvolver um algoritmo preditivo utilizando metodologias de Processamento de Linguagem Natural e Mineração de Texto, a fim de classificar frases coletadas do Twitter e, através de um modelo probabilístico, identificar a possibilidade de um texto ter natureza depressiva. Em suma, realizou-se uma pesquisa aplicada de cunho quantitativo, com a finalidade de apresentar, em forma de gráficos e tabelas, a relevância do algoritmo desenvolvido na detecção de pensamentos depressivos em redes sociais. Na Figura 11 é apresentada a proposta do trabalho, de modo a ilustrar o processo realizado.

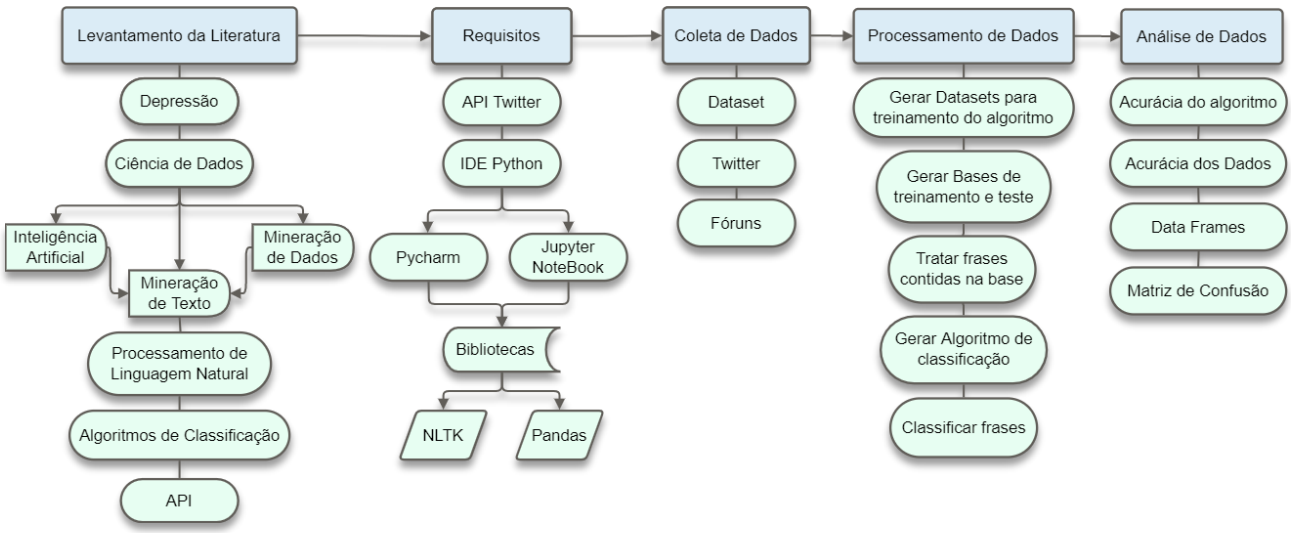

Figura 11: Organograma da “Proposta de trabalho”.

7.1. LEVANTAMENTO DA LITERATURA

Esta etapa teve por finalidade pesquisar a fundo cada tópico tratado, utilizando como base de conhecimento livros, artigos e fóruns, de modo a enriquecer o texto com contribuições de outros autores. A revisão da literatura serviu principalmente para garantir embasamento teórico suficiente para auxiliar na definição dos passos e na construção do algoritmo.

Para levantamento bibliográfico utilizou-se o Google Acadêmico com filtros para encontrar artigos em língua portuguesa e inglesa utilizando como palavras-chave “Mineração de Texto”, “Depressão”, “Text mining”, “KDT” e “Ciência de Dados”.

7.2. REQUISITOS

As ferramentas basilares para construir o classificador foram decididas previamente, entretanto, durante o processo de construção novas necessidades surgiram à medida que as complexidades dos requisitos aumentavam. Para contornar as dificuldades encontradas na utilização da IDE (Integrated Development Environment) Pycharm, foi optado por trocá-la pelo JupyterLab.

Os requisitos utilizados para realizar o trabalho foram: Python, API do Twitter, JupyterLab, NLTK e Pandas.

Python: É uma linguagem de programação interpretada, orientada a objetos e multiplataforma, de alto nível que apresenta uma sintaxe simples, versátil e legível para humanos. É muito utilizado no desenvolvimento de aplicações complexas como Aprendizado de Máquina, Mineração de Dados e Automatização, mas também em tarefas mais simplistas como Desenvolvimento Web e como recurso didático para ensinar programação.

API Twitter: Application Programming Interface ou simplesmente API, é um recurso muito utilizado na computação para criar comunicação entre sistemas distintos, basicamente é uma maneira de consumir os dados de outra aplicação sem ter a necessidade de implementar sua codificação. Desse modo, a API do Twitter possibilita a criação própria de um código que utilizará os dados disponibilizados para criar soluções independentes. Através desta API é possível ter acesso ao motor de busca usado no Twitter para pesquisar tweets, utilizando palavras-chave, *hashtags* e até datas.

JupyterLab: Segundo o próprio site *jupyter.org*, o JupyterLab é uma versão mais recente e robusta do Jupyter Notebook, que disponibiliza um ambiente de desenvolvimento interativo baseado na Web. Sua interface flexível permite a organização dos documentos em abas e possibilita que o usuário instale extensões externas, personalizando suas funcionalidades além das que têm de fábrica.

NLTK: Natural Language Toolkit, é um conjunto de bibliotecas escritas em Python que fornecem materiais computacionais suficientes para o processo de extração de conhecimento de documentos em linguagem natural humana, utilizando principalmente técnicas de PLN. Entre seus recursos para processamento de texto estão: classificação, tokenização, *stemming*, *tagging*, *parsing* e raciocínio semântico.

Pandas: É uma biblioteca gratuita e *open source* extremamente poderosa que fornece ferramentas para análise de dados utilizando Python. Ela trabalha manipulando Data Frames que são dispostos em formatos de tabelas que permitem indexar, manipular e organizar dados através de algoritmos sofisticados responsáveis por realizarem operações matemáticas, recuperar informações e abstrair os dados em esquemas legíveis.

7.3. COLETA DE DADOS

Baseando-se no modelo CRISP-DM apresentado na Seção 4.2, a coleta de dados é uma das etapas mais complexas, já que há diversos requisitos que garantem a qualidade dos documentos coletados, se a compreensão das informações apresentarem qualquer ruído, o algoritmo é completamente comprometido. Sabendo disso, foi preferido pedir ajuda a uma profissional para a indexação dos documentos depressivos, visto que o tema depressão é muito sensível e requer conhecimento específico.

Para a formação dos documentos não depressivos foram utilizados diversos meios como criar frases manualmente, retirar frases da Internet, do Twitter e por fim manipular os textos com palavras específicas que equilibrassem seus pesos para que padrões indesejados não atrapalhasse o classificador.

7.4. PROCESSAMENTO DE DADOS

Após coletados e conferidos as etapas para processar os dados seguem uma ordem específica que garante a estrutura das informações de acordo com o padrão necessário para classificá-las.

O primeiro passo utilizando Python, foi dividir os documentos em duas partes, uma para treinamento do algoritmo e outra para os testes, essa divisão é estritamente usada apenas para testes laboratoriais. O segundo passo foi extrair as informações dos documentos para dentro do Python. O terceiro passo foi organizar as informações em duas listas, uma para o treinamento e outra para teste, as listas, por sua vez, foram compostas por tuplas com os valores dos textos e sua respectiva classe. O terceiro passo foi aplicar as técnicas do PLN para transformar os dados e reduzir seu tamanho. Por fim, os textos tratados foram mapeados para que o Naive Bayes pudesse classificá-los.

7.5. ANÁLISE DE DADOS

A fim de transparecer a eficiência do algoritmo, seus erros e acertos foram apresentados em uma Matriz de Confusão que revela a quantidade de frases que foram confundidas e classificadas com classes erradas. Também foram realizados *prints* para que as frases rotuladas incorretamente pudessem ser identificadas a fim de melhorar a confiabilidade dos documentos.

Para descobrir a acurácia do classificador aplicado nas frases coletadas do Twitter foi preciso supervisão de uma profissional para conferir os valores obtidos e consequentemente suas frases.

8. METODOLOGIA

Como mencionado na Seção 5.4, o processo para minerar textos não possui etapas definidas, portanto, o presente trabalho baseou-se no modelo CRISP-DM (seção 4.2) para auxiliar principalmente na identificação dos problemas encontrados e das etapas comprometidas.

8.1. COMPREENSÃO DO NEGÓCIO

O levantamento dos requisitos partiu das dificuldades acerca dos conhecimentos necessários para a implementação do algoritmo. Basicamente, essa etapa foi composta em grande parte pelo Levantamento da Literatura que serviu como alicerce para o conhecimento adquirido. Como o trabalho não é um projeto real empresarial, algumas preocupações foram deixadas de lado como custos, metas e riscos, sendo assim, a atenção foi direcionada para o levantamento de softwares, capacidades intelectuais e pela busca de um profissional na área da saúde mental para cooperar no entendimento dos dados.

Nesta etapa, foi escolhida a rede social Twitter para que servisse como fornecedora dos dados tanto para classificação futura quanto para compor as bases textuais. A decisão foi tomada principalmente pela facilidade em se trabalhar com a API do Twitter e pela enorme disposição de textos que a plataforma oferece. Também foi decidido as ferramentas necessárias (seção 7.2) bem como o algoritmo classificador Naive Bayes.

8.2. COMPREENSÃO DOS DADOS

Para construir o conjunto de textos depressivos foi necessário a contribuição de uma especialista no assunto que orientou como as frases são formadas e alguns padrões encontrados na maioria dos textos que influenciam seus significados semânticos e sintáticos. A partir do entendimento adquirido, as frases foram construídas e armazenadas em um arquivo TXT.

Para formalizar o documento de textos depressivos foram necessárias diversas manipulações estruturais nas frases para compatibilizar com o contexto disponibilizado nas redes sociais como abreviar palavras, utilizar gírias e acrescentar palavrões, caso contrário

a frequência dessas palavras atrapalhariam o classificador.

O documento de textos não depressivos foi construído principalmente em cima de palavras que são raramente encontradas em frases depressivas como: cinema, cerveja, balada, séries, músicas, tecnologia, etc. Essas frases em grande maioria foram coletadas do próprio Twitter por meio de sua API, contudo, as coletas foram realizadas de modo que a frequência das palavras não atrapalhasse o classificador, limitando as buscas entre 4 e 8 frases que contivessem tais palavras.

Algumas regras foram criadas para manter padrões necessários para o classificador funcionar corretamente. A primeira regra foi estabelecer que os dados coletados tanto para composição das bases quanto para a classificação futura deveriam ser pesquisadas ignorando *tweets* com links ou marcações por motivos estruturais. Foi observado que frequentemente os tweets com links e/ou marcações eram muito curtos ou não apresentavam nenhuma relevância já que na maioria das vezes eram sobre algum vídeo ou música, o que traria novamente o problema com a frequência de palavras indesejadas. A segunda regra foi estabelecer uma palavra-chave muita utilizada em frases depressivas para treinar o algoritmo baseando-se na frequência dela, essa regra foi necessária já que o prazo para entrega do trabalho era curto e pela vastidão de palavras que podem caracterizar os textos depressivos, a palavra em questão escolhida foi “morrer”, entretanto também se utilizou outras palavras como “depressão”, “morte” e “suicídio”.

Ao final da coleta, a base de textos depressivos possuía 213 frases e a base não depressiva possuía 766 frases.

8.3. PREPARAÇÃO DOS DADOS

A primeira etapa consistiu em separar os documentos completos em documentos menores que seriam utilizados para treinamento e teste laboratorial, um algoritmo foi encarregado de separar os arquivos de acordo com a porcentagem desejada.

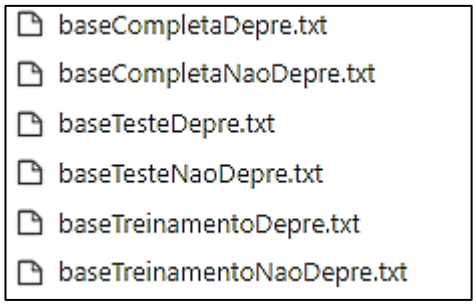

Figura 12: Documentos "Teste" e "Treinamento".

O segundo passo foi extrair as informações dos documentos para dentro do programa e separá-las em listas de treinamentos e testes. As listas foram compostas por tuplas (linhas) que na primeira posição representavam as frases e na segunda à sua categoria.

```
('cada vez que me corto, menos dor eu sinto', 'depressivo')
('Sextouuu! Sem cerveja Sem você ', 'nao_depressivo')
```

Figura 13: Estrutura das frases extraídas.

Com as frases estruturadas da maneira correta, se iniciou as etapas para processamento dos textos. Os procedimentos realizados foram: remoção das *stopwords*, tokenização, *steeming* e remoção de acentos.

Embora a *stoplist* disponibilizada pelo NLTK para língua portuguesa seja bem completa, ela não se preocupa com os problemas de acentuação e abreviação frequentemente encontrados em rede sociais e por esse motivo foi necessário complementá-la durante o processo com palavras como: “nao”, “pq”, “to”, “oq”, entre outras. Em conjunto com a operação de remoção das *stopwords* foi utilizado a técnica de tokenização utilizando expressão regular para remover números e caracteres especiais dos textos já que não possuíam relevância alguma.

```
('cada vez corto menos dor sinto', 'depressivo')
('sextouuu cerveja', 'nao_depressivo')
```

Figura 14: Remoção das *stopwords* e tokenização.

A última etapa para processamento dos textos foi reduzir as palavras a sua raiz ou radical e posteriormente retirar os acentos das palavras, essa decisão foi tomada já que diversos textos coletados do Twitter são ausentes de pontuação e acentos, logo para estabelecer um padrão, também se seguiu o mesmo raciocínio para as bases de treinamentos.

```
(['cad', 'vez', 'cort', 'menos', 'dor', 'sint'], 'depressivo')
(['sextouuu', 'cervej'], 'nao_depressivo')
```

Figura 15: *Stemming* e remoção de acentos.

A estrutura dos textos mudou de uma *string* única para uma lista em que cada posição ocupava uma palavra que pertencia respectivamente à sua frase, o motivo foi para facilitar o próximo passo da etapa de modelagem.

8.4. MODELAGEM

O primeiro passo da modelagem consistiu em criar uma única lista contendo todas as palavras das frases de treinamento, foi por esse motivo que se separou as *strings* em listas na etapa anterior, desse modo com um simples comando de *extend* do Python, as listas que ocupam a primeira posição da *tupla* se juntam, formando uma única lista. Ao todo, deram 8232 palavras que pertenciam somente aos documentos para treinamento.

```
vez fic imagin vid pesso trat transtorn tom medic direit
afog chuv morr sei inst cabec cois basic prefir mat faz
```

Figura 16: Algumas palavras juntas em uma única lista.

Para o modelo funcionar corretamente com o Naive Bayes é necessário trabalhar somente com palavras únicas, já que como cada elemento interfere individualmente na classificação, não faz sentido utilizar palavras repetidas, logo a segunda etapa da modelagem é formada por dois passos importantes: gerar a frequência de cada palavra e separar em uma lista somente uma palavra de cada.

Através da frequência das palavras é possível obter *insights* que ajudam a entender os padrões associados as bases textuais e assim criar soluções inteligentes, a frequência de palavras é muita utilizada para descobrir pedaços que estão confundindo o algoritmo além de ser frequentemente aplicada no estudo e formação de *Word Clouds*.

```
'morr': 126, 'mat': 86, 'faz': 75, 'ser': 62, 'tod': 56, 'fal': 53, 'vou': 53,
```

Figura 17: Frequência de algumas palavras.

Embora seja muito útil, esse processo também é utilizado para se obter somente as

palavras únicas, como é possível observar na Figura 17, a estrutura disponibilizada permite a extração somente das palavras, por ser um dicionário basta utilizar a função *keys* do Python e assim obter somente as chaves, descartando completamente o valor da frequência registrada. Ao todo deram 2292 palavras únicas contidas nas bases de treinamento.

```
 vid pesso trat transtorn tom medic direit fac
arec opc melhor sempr talv precis estud pens ma
s mund sum depred aut aniquil apert bot destru
 janel bibliotec engol tod remedi cas odi pec d
 mim inutil sint cans sent vai emb nad perd fra
cab log noss sim merec soluc problem suport mut
ic said ont infeliz acontec angust torn rez sai
```

Figura 18: Algumas palavras sem repetições em uma única lista.

Com as palavras únicas devidamente separadas e organizadas se iniciou a terceira etapa do processo segmentando-a também em dois passos: mapear e construir o classificador utilizando a base de treinamento e mapear as bases de testes para serem classificadas. Vale lembrar que essas etapas utilizando metade dos documentos são apenas utilizadas para os testes de laboratório, ao final dos testes as bases são encaixadas e mapeadas como um único documento de treinamento para poderem classificar as frases coletadas do Twitter.

A ideia por trás do Naive Bayes é construir uma tabela que cada frase é disposta em linhas e colunas, sendo que as linhas são compostas pelas próprias frases e as colunas são a representação de cada palavra única utilizada para o treinamento, formando assim um modelo que para toda palavra encontrada na frase é feita uma marcação de *True* ou *False*. A marcação *True*, indica que a palavra contida na frase pertence ao conjunto das palavras únicas definidas anteriormente e *False* indica a ausência das palavras únicas na frase em questão. A Figura 19 mostra como a primeira frase da Figura 15 (“[‘cad’, ‘vez’, ‘cort’, ‘menos’, ‘dor’, ‘sint’], ‘depressivo’”) foi mapeada.

```
'vid': False, 'pesso': False, 'trat': False,
False, 'surt': False, 'morr': False, 'faz': F
: False, 'sei': False, 'inst': False, 'cabec'
hor': False, 'sempr': False, 'talv': False, '
lse, 'mal': False, 'querr': False, 'alguem':
lse, 'suicid': False, 'desaparec': False, 'de
aniquil': False, 'apert': False, 'bot': False

'cad': True, 'vez': True, 'cort': True, 'menos': True, 'dor': True, 'sint': True,
```

Figura 19: Frase mapeada.

Após mapear cada frase, foi utilizado uma função chamada *train* do NLTK que é responsável por criar computacionalmente uma tabela probabilística que indica as ocorrências de cada palavra de acordo com sua *classe*, formando finalmente o classificador. O mesmo processo de mapeamento se repetiu para os documentos de testes com exceção da função *train*. Vale lembrar que as frases mapeadas ainda estão em uma *tupla* e que a segunda posição indica sua classe (“depressivo” ou “nao_depressivo”).

8.5. AVALIAÇÃO

O classificador foi avaliado duas vezes, uma para testes laboratoriais e outra com testes classificando *tweets*. Entende-se por testes em laboratório, um conjunto de documentos previamente preenchidos por dados manipulados e supervisionados considerados o objeto alvo, ou seja, o classificador é treinado esperando que os dados a serem classificados fora do laboratório apresentem características idênticas ou ao menos semelhantes aos textos contidos nos documentos de testes. Já para avaliar informações reais, os documentos de treinamento e testes são anexados e passam novamente pelas etapas de “Preparação dos Dados” e “Modelagem”, gerando assim um classificador completo.

8.5.1. AVALIAÇÃO EM LABORATÓRIO

Para avaliar o classificador em laboratório foram separados os documentos completos em duas bases de dados diferentes, uma para treinamento, e outra para testes, ambas com frases tanto depressivas quanto não depressivas. Foi realizado três avaliações dividindo os documentos em 10, 15 e 20 por cento para servir as bases de textos para testes, logo para treinamento o que restou foi, 90, 85, e 80 por cento dos documentos completos.

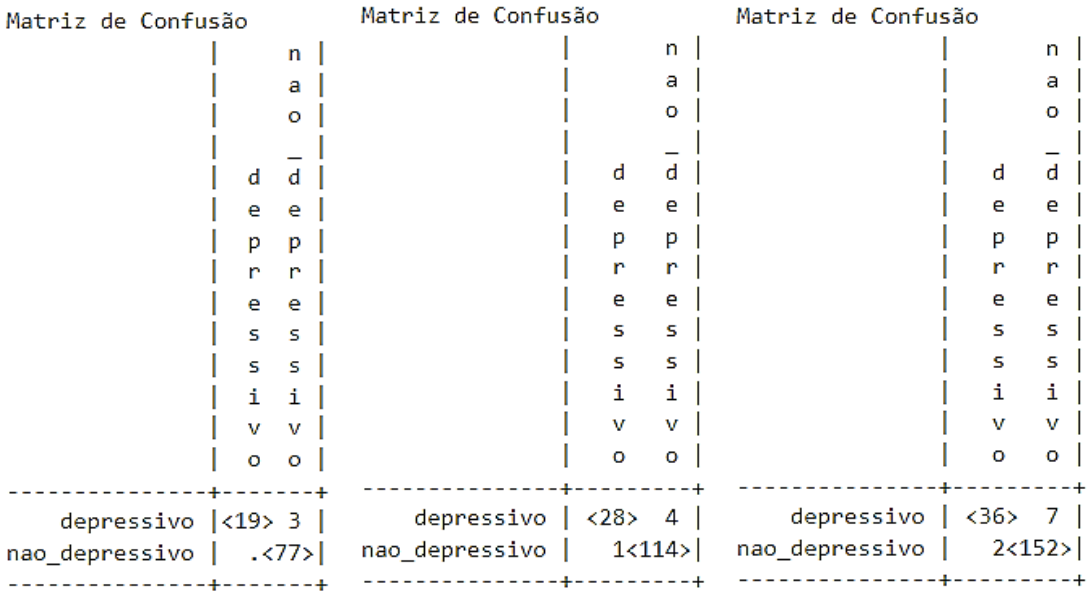

Figura 20: Matrizes de Confusão de 10%, 15% e 20%.

A matriz de confusão revela os resultados em um esquema bem simples onde as linhas representam as classes do modelo e as colunas revelam a classificação de cada elemento da classe, onde os números representados entre os sinais de menor e maior evidenciam os acertos e os números fora dos sinais apontam os erros.

Os testes apresentaram altos níveis de acurácia, mantendo uma precisão maior que 90%. A precisão do classificador para os testes de 10, 15, e 20 por cento foi de aproximadamente 96,96%, 96,59% e 95,43% respectivamente. Embora fosse um fenômeno esperado, foi possível observar que quanto mais frases eram retiradas do treinamento e usadas para teste, mais a precisão do algoritmo reduzia.

Essa divisão dos documentos na maioria das vezes revela padrões indesejados que estão confundindo o classificador, no caso dos testes realizados em laboratório foi possível observar que a palavra “feliz” contida em um texto depressivo estava confundindo o classificador fazendo com que ele classificasse erroneamente a frase, isso ocorreu, já que, a palavra em questão possuía frequência em excesso nas frases não depressivas, com esse padrão descoberto bastou diminuir um pouco a frequência da palavra “feliz” nas bases não depressivas e aumentar um pouco nas bases depressivas para que o algoritmo classificasse corretamente aquele texto.

8.5.2. AVALIAÇÃO DA CLASSIFICAÇÃO DOS TWEETS

Para testar o classificador utilizando frases do Twitter foram coletados 200 *tweets* obtidos de duas pesquisas realizadas através da API. A busca pelos *tweets* ocorreram durante o mês de julho de 2022 (07/2022), a primeira pesquisa filtrou somente *tweets* que continham obrigatoriamente o trecho “quero morrer” no corpo do texto, a segunda filtrou os textos que continham os trechos “quero morrer” e “quero me matar”.

Para identificar a acurácia do classificador referente às frases coletadas, foi solicitado ajuda a médica e psiquiatra Dra. Juliane, referenciada nos agradecimentos deste trabalho, que auxiliou na verificação das frases categorizadas e proporcionou conhecimento simplificado para o entendimento das características de um texto depressivo.

A precisão dos *tweets* coletados para a documentação deste trabalho e supervisionados pela Dra. Juliane foi de 74%. O classificador em geral apresentou uma acurácia relativamente boa, mantendo uma precisão de 68 a 87 por cento para as frases coletadas não documentadas.

De modo geral, as classificações das frases que continham “quero morrer” e “quero me matar” apresentaram resultados melhores, pois continham na maioria das vezes diversas palavras-chaves que caracterizavam textos depressivos. Já frases com poucos palavras-chaves apresentaram resultados menores, isso ocorreu pela grande quantidade de textos com sentidos opostos ao que se espera de uma frase que contenha “quero morrer” e “me matar”, por exemplo: “quero morrer quando me atraso”, “quero morrer só de pensar que amanhã é segunda”, “vou me matar se eles não vierem cantar na minha cidade”.

Segundo a Dra. Juliane alguns *tweets* foram classificados, certos ou errados, por competência das capacidades estatísticas do classificador, isso porque as frases são apresentadas em formatos textuais, logo não apresentam outros aspectos importantes para identificar uma pessoa com sinais depressivos, como contato visual, análise comportamental e principalmente a entonação que a frase seria pronunciada caso fosse falada e não escrita.

As frases coletadas não foram apresentadas neste trabalho para não discordar das políticas de privacidade do Twitter.

9. CONSIDERAÇÕES FINAIS

Enormes quantidades de dados são produzidas todos os dias e em todos os formatos seja por meio de fotos, vídeos, textos, planilhas, pesquisas e informações mantidas em bancos de dados, fazendo com que o volume disponibilizado já não permita mais análises profundas, descoberta de padrões e extração de conhecimento a partir de técnicas convencionais baseadas apenas em trabalhos manuais e na observação.

A Mineração de Dados surge como ferramenta principal para extrair e analisar informações atribuindo significado a elas, caso contrário, o trabalho manual seria lento e até mesmo ineficiente, levando em conta a grande quantidade de processos acerca das análises. Ainda que seja possível realizar os processos à mão, não seria viável, uma vez que levariam dias para se alcançar o mesmo resultado que os algoritmos de mineração produzem em pouco minutos.

Entre a imensa base de dados gerados diariamente estão os textos, que são informações descritas em linguagem natural que, por sua vez, se apresentam constantemente em formato não estruturado sendo frequentemente encontrados e produzidos em redes sociais, correios eletrônicos e sites. Vivemos em um momento que o cenário corporativo já é composto por mais de 80% de dados não estruturados, logo as empresas estão buscando cada vez mais tecnologias capazes de extrair, interpretar e revelar padrões contidos em informações desestruturadas.

Para suprir as necessidades científicas e empresariais em torno da exploração de bases textuais, surge a mineração de texto, que nada mais é que uma derivação conceitual e prática da mineração de dados, que disponibiliza estratégias e recursos para a descoberta de conhecimento a partir de textos com o propósito de criar soluções automáticas e inteligentes.

A problemática deste trabalho gira em torno do crescente número de pessoas depressivas no Brasil e no mundo todo que decorrente aos avanços tecnológicos passaram a utilizar ferramentas online de socialização, também chamadas de redes sociais, para exprimirem suas mágoas e tristezas, e que muitas vezes acabam por deixar tais registros somente online e privados não compartilhando-os com pessoas reais que poderiam disponibilizar ajuda e tratamento. Acerca do problema levantado, o presente trabalho buscou apresentar durante todas as seções, as técnicas e os conceitos da descoberta de conhecimento em

textos que comprovassem a hipótese de que um algoritmo classificador, se bem treinado e aplicado de maneira correta, seria possível analisar e classificar frases depressivas.

De acordo com os resultados obtidos e apresentados no capítulo de “Metodologia”, o algoritmo classificador construído foi capaz de predizer corretamente diversas frases apresentando um nível de acurácia relativamente bom, principalmente em testes laboratoriais em que os dados alvos são manipulados buscando maior semelhança possível com os dados testados fora de laboratório.

As maiores dificuldades encontradas na criação do modelo probabilístico foram as variações estruturais linguísticas encontradas nas redes sociais como palavrões, gírias, abreviações e erros ortográficos, semânticos e sintáticos. Como o modelo é dependente da frequência das palavras, muitas frases coletadas do Twitter foram classificadas erradas por que continuam inúmeras gírias além da maioria estarem carregadas de ambiguidades e conotações, além de frases com trechos claramente depressivos só que em contextos diferentes como “Que final de série chato, quero morrer”, “Vou me matar se esse casal fofo não ficar junto”, “Perdi essa merda desse ônibus, quero morrer”, entre outras.

Conforme o levantamento realizado neste capítulo e os resultados apresentados pelo classificador conclui-se que, um algoritmo classificador Naive Bayes se construído cooperativamente com ferramentas da Descoberta de Conhecimento em Textos e do Processamento de Linguagem Natural, pode apresentar resultados satisfatórios sendo capaz de detectar previamente pessoas com tendências depressivas através de textos publicados em redes sociais.

Espera-se que este trabalho influencie outras áreas do conhecimento a também criarem soluções inteligentes acerca do tema depressão e que sirva de modelo metodológico para ser utilizado em trabalhos futuros que desenvolvam uma solução criativa em cima dos resultados obtidos, por exemplo, criar uma plataforma de ajuda para as pessoas que o algoritmo identificar ou encaminhar os tweets classificados para outro classificador e assim conferir os resultados.

Para trabalhos futuros o maior objetivo será com certeza aprimorar o classificador, visando contornar a grande maioria dos problemas já citados neste capítulo bem como aumentar a qualidade dos dados treinados a um nível que a análise se torne a mais próxima da realizada por um ser humano.

REFERÊNCIAS

4 empresas que estão usando o Big Data para aumentar receitas e diminuir custos. **Digital House**. Disponível em: <https://www.digitalhouse.com/br/blog/4-empresas-que-estao-usando-o-big-data-para-aumentar-receitas-e-diminuir-custos/>. Acesso em: 20 Mar 2022.

AMARAL, Fernando. **Aprenda mineração de dados:** teoria e prática. Rio de Janeiro: Alta Books, 2016. 225 p. ISBN 9788576089889.

Aumenta o número de pessoas com depressão no mundo. **OPAS/OMS**. 2017. Disponível em: < https://www.paho.org/pt/noticias/23-2-2017-aumenta-numero-pessoas-com-depressao-no-mundo>. Acesso em: 20 Mar 2022.

ANDREATA, Guilherme Henrique Santos. **O Uso de Processamento de Linguagem Natural para a Análise de Sentimentos na Rede Social Reddit**. 2017. 52 f. TCC (Bacharelado) - Sistemas de Informação, Centro de Computação, Universidade de Caxias do Sul, 2018. Disponível em: < https://repositorio.ucs.br/xmlui/bitstream/handle/11338/3804/TCC%20Guilherme%20Henrique%20Santos%20Andreata.pdf?sequence=1&isAllowed=y>. Acesso em: 15 Mar 2022.

ARANHA, C.N. **Uma abordagem de pré-processamento automático para mineração de textos em português:** sob o enfoque da inteligência computacional, Tese de Doutorado, Departamento de Engenharia Elétrica, PUC/Rio, 2007. Disponível em: <http://docplayer.com.br/56945632-Uma-abordagem-de-pre-processamento-automatico-para-mineracao-de-textos-em-portugues-sob-o-enfoque-da-inteligencia-computacional.html>. Acesso em: 02 Jul 2022.

CASTRO, Leandro Nunes; FERRARI, Daniel Gomes. **Introdução à Mineração de Dados:** Conceitos Básicos, Algoritmos e Aplicações. Editora Saraiva, 2016. 978-85-472-0100-5. 369 p. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/978-85-472-0100-5/>. Acesso em: 02 Nov 2021.

CLEVELAND, William S. Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. **International Statistical Review / Revue Internationale de Statistique**. Published By: International Statistical Institute (ISI). Vol. 69, No. 1 (Apr., 2001), pp. 21-26 (6 pages). Disponível em: <https://www.jstor.org/stable/1403527>. Acesso em: 28 Mai 2022.

CORRÊA, Geraldo Nunes; MARCACINI, Ricardo Marcondes; REZENDE, Solange Oliveira. **Uso da Mineração de Textos na análise exploratória de artigos científicos**. São Carlos: Instituto de Ciências Matemáticas e Computação. 2012. Disponível em: < http://repositorio.icmc.usp.br/handle/RIICMC/6631>. Acesso em 01 jun. 2022.

DELOUYA, Daniel. **Depressão**. 5ª edição. São Paulo: Casa do Psicólogo, 2008.

Depressão. **OPAS/OMS**. Disponível em: < https://www.paho.org/pt/topicos/depressao#:~:text=A%20depress%C3%A3o%20%C3%A9%20um%20transtorno%20mental%20frequente.,a%20carga%20global%20de%20doen%C3%A7as.>. Acesso em 07 nov. 2021.

Depressão pode causar de tristeza a vontade de tirar a vida. **VivaBem uol**. Disponível em: < https://www.uol.com.br/vivabem/noticias/redacao/2018/07/10/depressao-sintomas-tratamento-tipos-e-relacao-com-suicidio.htm>. Acesso em 07 nov. 2021.

ESCOVEDO, Tatiana; KOSHIYAMA, Adriano. **Introdução a Data Science:** Algoritmos de Machine Learning e métodos de análise. Casa do Código, 2020. 656 p.

GATTAZ, W. F. Vida sem depressão. **At revista - A Tribuna**, Santos - SP, p. 6 - 10, 07 dez. 2014. Disponível em: <http://neurociencias.org.br/wp-content/uploads/2016/10/Vida-sem-Depressa%CC%83o-A-Tribuna-Atrevista.pdf>. Acesso em: 20 jan 2022.

GONÇALVES, Eduardo. Mineração de texto - Conceitos e aplicações práticas. **SQL Magazine**, v. 105, 2012, p. 31-44. Disponível em: <https://www.researchgate.net/publication/317912973_Mineracao_de_texto_-_Conceitos_e_aplicacoes_praticas>. Acesso em 02 nov. 2021.

GRUS, Joel. **Data Science do zero:** Primeiras regras com o Python. 1ª edição. Rio de Janeiro: Alta Books, 2016. 442 p.

JUNIOR, João Ribeiro Carrilho. **Desenvolvimento de uma Metodologia para Mineração de Textos**. 2007. 96f. Dissertação (Mestrado) – Curso de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007. Disponível em: <https://www.maxwell.vrac.puc-rio.br/colecao.php?strSecao=resultado&nrSeq=11675@1>. Acesso em: 05 jul. 2022.

How BI and Big Data is Essential to McDonald's Growth Strategy. **YourShortList**. Disponível em: < https://yourshortlist.com/how-bi-and-big-data-is-essential-to-mcdonald-s-growth-strategy/>. Acesso em: 28 Mai 2022.

KING, Timothy. 80 Percent of Your Data Will Be Unstructured in Five Years. **SolutionsReview**, 2019. Disponível em: <https://solutionsreview.com/data-management/80-percent-of-your-data-will-be-unstructured-in-five-years/>. Acesso em 02 jun. 2022.

LOPES, Lucelene; VIEIRA, Renata. **Processamento de Linguagem Natural e o Tratamento Computacional de Linguagens Científicas**. In: Cristina Lopes Perna; Heloísa Koch Delgado; Maria José Finatto. (Org.). Linguagens Especializadas em Corpora: modos de dizer e interfaces de pesquisa. Porto Alegre: EDIPUCRS, 2010, p. 183-201. Disponível em: < https://editora.pucrs.br/edipucrs/acessolivre//livros/linguagensespecializadasemcorpora.pdf>. Acesso em 02 jun. 2022.

MORAIS, Edison; AMBRÓSIO, Ana Paula L. **Mineração de Textos**. Instituto de Informática - Universidade Federal de Goiás, 2007. Disponível em: <https://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf>. Acesso em 02 nov. 2021.

PROVOST, Foster; FAWCETT, Tom. **Data Science Para Negócios:** O que você precisa saber sobre Mineração de Dados e Pensamento Analítico de Dados. 1ª edição. Rio de Janeiro: Alta Books, 2016. 592 p.

RIBEIRO, A. C. **Modelo de reconhecimento de padrões em ideias usando técnicas de descoberta de conhecimento em textos**. 2018. 172 p. Dissertação (Mestrado) - Curso de Engenharia e Gestão do Conhecimento, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2018. Disponível em: <https://repositorio.ufsc.br/handle/123456789/193500>. Acesso em: 05 jun. 2022.

ROCHA, Thiago; PERES, Sarajane Marques; BÍSCARO, Helton Hideraldo; MADEO, Renata C. B; BOSCARIOLI, Clodis. Tutorial sobre Fuzzy-c-Means e Fuzzy Learning Vector Quantization: abordagens híbridas para tarefas de agrupamento e classificação. ***Revista de Informática Teórica e Aplicada***, Porto Alegre, v. 19, n. 1, p. 120-163, 2012. Disponível em: < http://seer.ufrgs.br/rita/article/view/rita_v19_n1_p120/18115 >. Acesso em: 17 Mar 2022.

SANTOS, Cedric M. **Classificação de Documentos com Processamento de Linguagem Natural**. Tese (Mestrado em Informática e Sistemas). Departamento de Engenharia Informática e

de Sistemas do Instituto Superior de Engenharia de Coimbra. Coimbra. p. 217. 2015. Disponível em: <https://comum.rcaap.pt/bitstream/10400.26/15293/1/Cedric-Michael-Santos.pdf>. Acesso em: 17 jun 2022.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à Mineração de Dados:** Com aplicações em R. Editora Campus – Série SBC, ISBN: 9788535284461, 296p., 2016.

SIMÕES, Newman Ribeiro. **Estimadores de Bayes**. Aplicação ao modelo de regressão linear simples. 1981. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura Luiz de Queiroz, University of São Paulo, Piracicaba, 1981. doi:10.11606/D.11.1981.tde-20220208-021938. Acesso em: 2022-03-15.

SMALLCOMBE, Mark. Structured vs Unstructured Data: 5 Key Differences. **integrate.io**, 2022. Disponível em: <https://solutionsreview.com/data-management/80-percent-of-your-data-will-be-unstructured-in-five-years/>. Acesso em 02 jun. 2022.

SOUZA, C.; MOREIRA V. Tristeza, depressão e suicídio melancólico: a relação com o Outro. **Arquivos Brasileiros de Psicologia**; Rio de Janeiro, 70 (2): 173-185, 2018. Disponivel em: <http://pepsic.bvsalud.org/pdf/arbp/v70n2/13.pdf>. Acesso em: 15 Mar 2022.

Um pouco da história da Ciência de Dados. **ILUMEO**. Disponivel em: <https://ilumeo.com.br/todos-posts/2021/08/17/um-pouco-da-historia-da-ciencia-de-dados>. Acesso em: 28 Mai 2022.

VIGITEL BRASIL 2021. **ESTIMATIVAS SOBRE FREQUÊNCIA E DISTRIBUIÇÃO SOCIODEMOGRÁFICA DE FATORES DE RISCO E PROTEÇÃO PARA DOENÇAS CRÔNICAS NAS CAPITAIS DOS 26 ESTADOS BRASILEIROS E NO DISTRITO FEDERAL EM 2021**. Ministério da Saúde, Secretaria de Vigilância em Saúde, Departamento de Análise em Saúde e Vigilância de Doenças Não Transmissíveis. Disponivel em: <https://static.poder360.com.br/2022/04/pesquisa-saude-vigitel-2021.pdf>. Acesso em: 15 jun 2022.

VISCOTT, David. **A linguagem dos sentimentos**. São Paulo: Summus Editorial, 1982.

WIVES, L.K. **Utilizando Conceitos como Descritores de Textos para o Processo de Identificação de Conglomerados (clustering) de Documentos**. 2004. 136p. Tese (Doutorado em Ciência da Computação) – Universidade Federal do Rio Grande do Sul, Porto Alegre.