

UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO” – UNESP – CÂMPUS DE BAURU
FACULDADE DE ARQUITETURA, ARTES E COMUNICAÇÃO – FAAC
PROGRAMA DE PÓS-GRADUAÇÃO EM MÍDIA E TECNOLOGIA

ALCIANO GUSTAVO GENOVEZ DE OLIVEIRA

**PROPOSTA DE APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE
DADOS PARA ANÁLISE DO COMPORTAMENTO DAS PESSOAS EM
UMA REDE SOCIAL COM O INTUITO DE DETERMINAR POSSÍVEIS
INDÍCIOS DE SINDROMES SOCIOEMOCIONAIS**

BAURU – SP
FEVEREIRO 2023

UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO” – UNESP – CÂMPUS DE BAURU
FACULDADE DE ARQUITETURA, ARTES E COMUNICAÇÃO – FAAC
PROGRAMA DE PÓS-GRADUAÇÃO EM MÍDIA E TECNOLOGIA

ALCIANO GUSTAVO GENOVEZ DE OLIVEIRA

**PROPOSTA DE APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE
DADOS PARA ANÁLISE DO COMPORTAMENTO DAS PESSOAS EM
UMA REDE SOCIAL COM O INTUITO DE DETERMINAR POSSÍVEIS
INDÍCIOS DE SINDROMES SOCIOEMOCIONAIS**

Tese apresentada ao Programa de pós-graduação
em Mídia e Tecnologia da Faculdade de Arquitetura,
Arte e Comunicação, Unesp – Bauru, como requisito
para obtenção do título de Doutor em Mídia e
Tecnologia.

Orientador: Eduardo Martins Morgado

Coorientador(a): Nair Cristina Margarido Brondino

BAURU – SP
FEVEREIRO 2023

O48p

Oliveira, Alciano Gustavo Genovez de

Proposta de aplicação de técnicas de mineração de dados para análise do comportamento das pessoas em uma rede social com o intuito de determinar possíveis indícios de síndromes socioemocionais / Alciano Gustavo Genovez de Oliveira. -- Bauru, 2023

98 f. : il.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Arquitetura, Artes, Comunicação e Design, Bauru

Orientador: Eduardo Martins Morgado

Coorientadora: Nair Cristina Margarido Brondino

1. Síndrome socioemocional. 2. SAS Enterprise Miner. 3. SAS Visual Text Analytics. 4. Mineração de dados. 5. Rede social. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Arquitetura, Artes, Comunicação e Design, Bauru. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

ATA DA DEFESA PÚBLICA DA TESE DE DOUTORADO DE ALCIANO GUSTAVO GENOVEZ DE OLIVEIRA, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM MÍDIA E TECNOLOGIA, DA FACULDADE DE ARQUITETURA, ARTES, COMUNICAÇÃO E DESIGN.

Aos 24 dias do mês de fevereiro do ano de 2023, às 09:00 horas, por meio de Videoconferência, realizou-se a defesa de TESE DE DOUTORADO de ALCIANO GUSTAVO GENOVEZ DE OLIVEIRA, intitulada **Proposta de aplicação de técnicas de mineração de dados para análise do comportamento das pessoas em uma rede social com o intuito de determinar possíveis indícios de síndromes socioemocionais**. A Comissão Examinadora foi constituída pelos seguintes membros: Professor Associado EDUARDO MARTINS MORGADO (Orientador(a) - Participação Virtual) do(a) Departamento de Computação da Faculdade de Ciências / Universidade Estadual Paulista, Prof. Dr. FERNANDO AUGUSTO GARCIA MUZZI (Participação Virtual) do(a) Faculdade de Tecnologia de Lins / Centro Estadual de Educação Tecnológica Paula Souza, Prof. Dr. ADRIANO DE SOUZA MARQUES (Participação Virtual) do(a) Engenharia da Computação / Instituto Federal de São Paulo - IFSP - Câmpus Birigui. Após a exposição pelo doutorando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, o discente recebeu o conceito final: **APROVADO**. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

Professor Associado EDUARDO MARTINS MORGADO



ALCIANO GUSTAVO GENOVEZ DE OLIVEIRA

**PROPOSTA DE APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE
DADOS PARA ANÁLISE DO COMPORTAMENTO DAS PESSOAS EM
UMA REDE SOCIAL COM O INTUITO DE DETERMINAR POSSÍVEIS
INDÍCIOS DE SINDROMES SOCIOEMOCIONAIS**

Tese aprovada como requisito parcial para obtenção do Grau de Doutor em Mídia e Tecnologia da Unesp – Bauru, pela seguinte banca examinadora:

Professor Orientador: Prof. Dr. Eduardo Martins Morgado (Presidente)
PPG – Mídia e Tecnologia – FAAC – Unesp Bauru

Prof. Dr. Adriano de Souza Marques
Depto. de Computação – IFSP – Campus Birigui

Prof. Dr. Fernando Augusto Garcia Muzzi
Docente – Faculdade de Tecnologia de Lins

Bauru – 24 de fevereiro de 2023

Dedicatória

A meus pais Alceu e Sonia, esposa Edilma, filha Gabriela, tia Vera, avós e amigos.

AGRADECIMENTOS

Agradeço primeiramente a Deus pelo dom da vida e por minha família, é graças a ele que estou concluindo este Doutorado. É Deus quem tem me dado forças não só para concluir este trabalho, mas também para superar todas as adversidades que surgiram em minha vida.

A minha mãe Sonia e em memória de meu pai Alceu, por estarem sempre ao meu lado e por terem me ensinado virtudes que foram fundamentais para a formação de minha pessoa.

A minha esposa Edilma pela compreensão e pela ajuda me motivando, cobrando e dando forças para que eu não desistisse durante as dificuldades.

A minha filha Gabriela que é a minha joia preciosa, que com seu sorriso ao me ver quando chego em casa me abastece de energia e me dá forças para continuar sempre seguindo em frente.

Ao Prof. Dr. Eduardo Martins Morgado, por ter-me permitido fazer parte de seu grupo de trabalho e, ter acreditado na minha capacidade confiando a mim suas horas de trabalho me auxiliando e orientado, sempre que solicitei.

Agradeço, de igual forma, à Prof.^a Dr.^a Nair Cristina Margarido Brondino, pelas conversas e orientações, sendo sempre muito solícita e prestativa, não medindo esforços em me ajudar.

Ao Prof. Dr. Marcos Américo (Tuca) que enquanto gestor do programa de Pós-Graduação oferecido pela FAAC, foi muito atencioso e preocupado com nosso desempenho enquanto alunos, sempre nos, orientando e auxiliando no processo como um todo e, também à Prof.^a Dr.^a Vânia Cristina Pires Nogueira Valente pelo excelente trabalho realizado a frente do Programa de Doutorado.

Agradeço ao Helder e a todos os funcionários(as) da secretaria da FAAC que sempre foram muito prestativos e atenciosos realizando um trabalho de excelência.

Aos meus amigos que de várias formas também contribuíram para a conclusão de mais esta etapa de minha vida. À Camila e a todos da empresa SAS que foram de vital importância no suporte prestado no uso das soluções SAS utilizadas nesta Tese.

Ao Departamento de Pós-Graduação da FAAC da Unesp de Bauru por proporcionar a realização de um sonho pessoal de me tornar Doutor. É um sonho realizado!

*Combati o bom combate, acabei a carreira, guardei a fé.
Timóteo 4:7*

RESUMO

Segundo a OMS (Organização Mundial da Saúde), o índice de síndromes socioemocionais que vem acometendo as pessoas teve aumento significativo. Pesquisas indicam que, entre 2010 e 2020, esse índice teve aumento de 24%. Alguns fatores que podem contribuir para esse aumento são: a incapacidade em superar problemas emocionais, a desmotivação devido a problemas familiares e/ou financeiros, a cobrança excessiva por parte da sociedade com relação ao convívio social entre outros. Em muitos casos, só se percebe que uma pessoa está acometida de alguma síndrome socioemocional de forma tardia, e em alguns deles, após o extremo que pode chegar ao suicídio. Quando tratada adequadamente, a pessoa, por muitas vezes volta a ter uma vida considerada normal e seu convívio social bem como, seu comportamento melhoram significativamente. Pensando nesse tipo de problema e associando ao alerta da OMS sobre o aumento significativo de pessoas acometidas de síndromes socioemocionais, essa Tese de Doutorado, apresenta uma proposta de aplicação de técnicas de Mineração de Dados, aplicadas a dados provenientes de uma Rede Social. O objetivo é de contribuir, oferecendo uma forma de tentar detectar potenciais pessoas acometidas de síndromes socioemocionais. Para isso são utilizadas duas ferramentas de mineração de dados conhecidas por *SAS Enterprise Miner* e *SAS Visual Text Analytics*. O método computacional, bem como o resultado dos testes são fundamentados em métodos estatísticos que são implementados pelas próprias ferramentas de Mineração que têm por finalidade validar os métodos de busca, de classificação e de organização dos resultados obtidos. Dessarte, é apresentada uma proposta para determinar se as pessoas estão sofrendo de algum tipo de síndrome socioemocional (depressão, ansiedade e estresse) sem que elas saibam disso, evitando uma abordagem direta. A coleta, a mineração e a análise dos dados seguem a Lei 13.709 que regulamenta a Proteção dos Dados Pessoais. Com isso pretende-se contribuir para que psicólogos, orientadores educacionais e outros profissionais possam determinar se as pessoas possam estar passando por problemas psicológicos provenientes de Transtornos Emocionais, associando assim, a área de Psicologia à aplicação de técnicas de mineração de dados.

Palavras-chave: Síndrome Socioemocional; *SAS Enterprise Miner*; *SAS Visual Text Analytics*; Mineração de Dados; Rede Social.

ABSTRACT

According to the WHO (World Health Organization), the rate of socio-emotional syndromes that has been affecting people has increased significantly. Research indicates that, between 2010 and 2020, this index increased by 24%. Some factors that may contribute to this increase are: the inability to overcome emotional problems, demotivation due to family and/or financial problems, excessive demands by society regarding social interaction, among others. In many cases, it is only realized that a person is affected by some socio-emotional syndrome at a later date, and in some of them, after the extreme that can lead to suicide. When properly treated, the person often returns to a life considered normal and their social life, as well as their behavior, improve significantly. Thinking about this type of problem and associating it with the WHO alert about the significant increase in people suffering from socio-emotional syndromes, this Doctoral Thesis presents a proposal for the application of Data Mining techniques, applied to data from a Social Network. The objective is to contribute, offering a way to try to detect potential people affected by socio-emotional syndromes. For this, two data mining tools known as SAS Enterprise Miner and SAS Visual Text Analytics are used. The computational method, as well as the results of the tests, are based on statistical methods that are implemented by the Mining tools themselves, whose purpose is to validate the methods of searching, classifying and organizing the results obtained. Thus, a proposal is presented to determine whether people are suffering from some type of socio-emotional syndrome (depression, anxiety and stress) without them knowing it, avoiding a direct approach. The collection, mining and analysis of data follow Law 13,709 which regulates the Protection of Personal Data. This is intended to help psychologists, educational counselors and other professionals to determine whether people may be experiencing psychological problems arising from Emotional Disorders, thus associating the area of Psychology with the application of data mining techniques.

Keywords: Socioemotional Syndrome; SAS Enterprise Miner; *SAS Visual Text Analytics*; Data Mining; Social Network.

LISTA DE ILUSTRAÇÕES

Figura 1 – Etapas Operacionais do Processo KDD	30
Figura 2 – Modelo de Rede Neural	33
Figura 3 – Modelo de Árvore de Decisão	34
Figura 4 – Fluxograma de um Algoritmo Genético	34
Figura 5 – Matriz de representação do grafo de uma rede complexa	36
Figura 6 – Exemplo de grafo para redes complexas dirigidas e não dirigidas	37
Figura 7 – Redes complexas com coeficiente de aglomeração distintos	41
Figura 8 – Quadro de associação entre técnicas e tarefas utilizadas em mineração de dados	50
Figura 9 – Exemplo de diagrama de fluxo	66
Figura 10 – Diagrama de blocos no SAS Visual Text Analytics	67
Figura 11 – Interface do <i>Facepager</i>	70
Figura 12 – Base de dados no formato CSV	71
Figura 13 – Amostra da base de dados no formato <i>SAS Data Set</i>	72
Figura 14 – Interface de análise da base de dados	72
Figura 15 – Diagrama de blocos com a base de dados e o particionamento	74
Figura 16 – Diagrama de bloco com o bloco <i>Text Parsing</i>	74
Figura 17 – Diagrama de blocos com o bloco <i>Text Parsing</i>	76
Figura 18 – Diagrama de blocos utilizado no <i>SAS Visual Text Analytics</i>	77
Figura 19 – Quantidade de postagens que utilizam a palavra Depressão	78
Figura 20 – Quantidade de postagens que utilizam a palavra Ansiedade	79
Figura 21 – Quantidade de postagens que utilizam a palavra Estresse	79
Figura 22 – Quantidade de postagens por usuários que utilizam as três palavras ...	80
Figura 23 – Quantidade de seguidores e amigos por usuários que utilizam a palavra Depressão	81
Figura 24 – Palavras associadas à palavra Depressão	82
Figura 25 – Palavras associadas à palavra Depressão por postagem dos usuários	82
Figura 26 – Quantidade de postagens que utilizam a palavra Depressão obtidas no <i>SAS VisualText Analytics</i>	83

Figura 27 – Quantidade de postagens que utilizam a palavra Ansiedade obtidas no SAS <i>VisualText Analytics</i>	84
Figura 28 – Quantidade de postagens que utilizam a palavra Estresse obtidas no SAS <i>VisualText Analytics</i>	84
Figura 29 – Quantidade de postagens por usuários que utilizam as três palavras obtidas no SAS <i>VisualText Analytics</i>	85
Figura 30 – Quantidade de seguidores e amigos por usuários que utilizam a palavra Depressão obtida no SAS <i>VisualText Analytics</i>	86
Figura 31 – Palavras associadas à palavra Depressão obtidas no SAS <i>VisualText Analytics</i>	87
Figura 32 – Palavras associadas à palavra Depressão por postagens do usuário obtidas no SAS <i>VisualText Analytics</i>	87

SUMÁRIO

MEMORIAL ACADÊMICO	15
1. INTRODUÇÃO.....	22
1.1 Motivação	24
1.2 Objetivos	26
1.3 Justificativa.....	27
1.4 Metodologia.....	27
1.5 Organização do Trabalho	28
2. FUNDAMENTAÇÃO TEÓRICA	30
2.1 Descoberta de conhecimento em banco de dados	30
2.2 Mineração de dados	31
2.2.1 Técnicas de mineração dados.....	32
2.2.1.1 Redes complexas	36
2.2.1.1.1 Conectividade.....	37
2.2.1.1.2 Ciclos	40
2.2.1.1.3 Distância	41
2.2.1.1.4 Centralidade	42
2.2.1.1.5 K-core.....	44
2.2.2 Tarefas em mineração de dados	45
2.2.2.1 Classificação	46
2.2.2.2 Agrupamento	46
2.2.2.3 Predição	46
2.2.2.4 Redes Bayesianas.....	47
2.2.2.5 Análise discriminante	48
2.2.2.6 Análise exploratória	49
2.3 REDES SOCIAIS.....	51
2.3.1 Redes sociais digitais.....	52
2.3.2 Redes sociais digitais mais utilizadas.....	52
2.3.2.1 Facebook.....	53
2.3.2.2 WhatsApp.....	53
2.3.2.3 FaceBook Messenger	54

2.3.2.4 Youtube	55
2.3.2.5 Instagram.....	56
2.3.2.6 LinkedIn	56
2.3.2.7 Twitter.....	57
2.3.2.8 Pinterest	58
2.3.2.9 Snapchat	58
2.3.2.10 TikTok	60
2.4 TRANSTORNOS EMOCIONAIS	60
2.4.1 Depressão.....	61
2.4.2 Estresse	62
2.4.3 Ansiedade	62
2.5 FERRAMENTAS DE MINERAÇÃO DE DADOS.....	64
2.5.1 SAS.....	64
2.5.2 SAS Enterprise Miner.....	65
2.5.3 SAS Visual Text Analytics	67
2.6 Lei Geral de Proteção aos Dados Pessoais - Lei nº 13.709/2018	68
3. METODOLOGIA.....	69
3.1 Ferramentas utilizadas para a mineração dos Dados	69
3.2 Rede Social utilizada para a coleta dos Dados	69
3.3 Coleta dos Dados	70
3.4 Tratamento dos Dados	71
3.5 Pré-processamento	72
3.6 Mineração dos Dados.....	75
4. Resultados Obtidos	78
4.1 Resultados com a ferramenta SAS Enterprise Miner	78
4.2 Resultados com a ferramenta SAS Visual Text Analytics	83
4.3 Análise dos resultados	88
5. CONCLUSÃO	89
6. PROPOSTAS PARA TRABALHOS FUTUROS	91
REFERÊNCIAS	92

Memorial Acadêmico

Meu nome é Alciano Gustavo Genovez de Oliveira, filho de Sonia Aparecida Genovez, graduada em Contabilidade, que deixou a profissão assim que eu nasci e se tornou “do lar”; e de Alceu de Oliveira, administrador de empresas, pós-graduado, que exerceu a profissão até bem pouco tempo antes de falecer. Sou natural de Lins – SP, nascido em 27 de março de 1980.

1998

Minha carreira acadêmica profissional teve início em 1998, no curso de Bacharel em Engenharia da Computação. Escolhi ser engenheiro por me identificar com eletrônica e informática e tinha muita vontade de unir as duas áreas.

Nessa época, para auxiliar a pagar os estudos, comecei a dar aulas de Manutenção de Computadores e Redes de Computadores em uma franquia de uma escola de informática e idiomas, a “Microlins”, localizada nas cidades de Lins e de Guaíçara. Além disso, consegui uma vaga de estágio remunerado no STH (Suporte Técnico em Hardware) que era um setor interno da UNILINS (Centro universitário de Lins que contemplava a EEL – Escola de Engenharia de Lins). Com a soma dessas duas remunerações, consegui pagar meus estudos.

2003

Em 2003, comecei a luta por uma colocação no mercado de trabalho. O meu primeiro emprego foi como gerente e CPD (Centro de Processamento dos Dados) em um laticínio localizado na cidade de Lins. O valor era menor do que o valor que eu pagava de mensalidade na faculdade; porém, estava empregado e aceitei o serviço.

Entrei em contato com a diretoria da empresa e informei que gostaria de continuar a estudar, mas não forneceram condições para isso; após um ano de frustração por não dar continuidade aos estudos, decidi sair do emprego e trabalhar como autônomo.

Após sair da empresa e dar início à prestação de serviços autônomos, decidi por fundar a Engecomp em setembro de 2003, com o auxílio dos meus pais que eram

sócios da empresa. A proposta era de criar uma empresa no seguimento de tecnologia para prestar serviços de manutenção e suporte em computadores, implantação e manutenção de Redes de Computadores e manutenção de monitores e impressoras.

A Engecomp era gerida pela família. Minha mãe cuidava da contabilidade, meu pai da parte administrativa e da carteira de clientes e eu da parte operacional. Em pouco tempo, consegui contratar funcionários, tendo três contratados, fora os terceirizados, com serviços de impressoras e monitores. A empresa cresceu muito rápido e consegui prestar serviços para empresas como Sabesp, transportadoras, câmara municipal, escritórios de contabilidade e de advocacia.

Nesse período, consegui fazer uma pós-graduação em Administração de Sistemas Operacionais e Servidores Linux na Empresa “Epopéia”, localizada na cidade de Marília – SP, com aulas aos sábados, durante todo o dia. Naquela época, quem conhecia Linux estava no topo do mercado. Dessa forma, a empresa começou a prestar serviços em soluções em Servidores de arquivos, impressão e *Firewall*.

2008

Em 2008 prestei concurso público e me tornei docente da Etec de Lins. Nessa época, eu já havia passado por várias experiências em sala de aula, além da Microlins fui professor de informática básica no Sindicato das Indústrias de Alimentação e Afins de Araçatuba e região mantendo, ainda, a Engecomp atuando no mercado de trabalho corporativo.

Era o momento de tomar uma decisão, se fosse aprovado no concurso público, venderia a empresa e sairia do sindicato para me dedicar exclusivamente à carreira acadêmica. O próximo passo era dar início ao mestrado. Fui aprovado no concurso público para docente na Etec de Lins e comecei a ministrar aulas em março de 2008 atuando como docente nessa instituição desde então. Quanto à Engecomp, um dos funcionários decidiu administrá-la; entretanto, reduziu os serviços prestados e vários clientes foram absorvidos por outras empresas do setor de soluções em TI.

Além da docência, exerci algumas funções dentro da instituição, tais como: responsável pela manutenção dos computadores na Etec de Cafelândia (por aproximadamente dois anos) e coordenador do Curso Técnico em Programação de Computadores, por igual período.

Nessa época, eu conciliava a docência nas Etecs de Lins e Cafelândia. Para conseguir cursar o Mestrado, prescindi as aulas na Etec de Cafelândia, permanecendo na docência somente na Etec de Lins. Com isso, foi possível iniciar a pós-graduação.

2011

Em 2011, iniciei o mestrado de forma efetiva, antes disso, eu havia feito uma disciplina como aluno especial em Engenharia Elétrica na Unesp de Ilha solteira; todavia, devido à minha formação ser em estudos sobre baixa e média tensão e, na Unesp de Ilha Solteira o perfil de interesse ser por alunos com conhecimento em média e alta tensão, tive dificuldades em me adaptar.

Quando descobri o programa de Mestrado em Ciência da Computação, finalmente me identifiquei. Dei início a esses estudos, no segundo semestre de 2011 como aluno especial, cursando a disciplina “Sistemas Biométricos” com o Professor Dr. Nilceu – Coordenador do programa naquela época.

Ainda como aluno especial me interessaram as disciplinas de “Gerencia de Projetos, Teoria e Prática” e “Tópicos Especiais: Modelagem e Simulação de Sistemas a Eventos Discretos”. A cada disciplina cursada, era um desafio vencido diante das dificuldades em conciliar a docência, as viagens para a Unesp de Bauru e de São José do Rio Preto, além de conseguir tempo para os estudos, o que me custou muitas madrugadas em claro.

No primeiro semestre de 2014, tornei-me aluno regular do programa sob orientação do professor Dr. Norian. Conclui os créditos através das disciplinas “Arquitetura de Computadores”, “Análise e Projeto de Algoritmos” e “Estudos Especiais – I”.

Durante esse período, em particular no começo do ano de 2015, meu pai veio a falecer. Foi um momento muito difícil e, com isso, tudo mudou. Assumi a responsabilidade de auxiliar minha avó e minha mãe. Com a perda de familiares próximos, minha mãe e minha avó por parte de pai ficaram sozinhas, mas cada uma tinha a sua casa e a sua vida. Auxiliá-las começou a fazer parte do meu dia a dia.

2016

A defesa da dissertação foi realizada em novembro de 2016, momento em que me tornei Mestre em Ciência da Computação com o desenvolvimento de um sistema de recomendação para classificar alunos por meio da utilização de sete Inteligências Múltiplas propostas na Teoria da Inteligências Múltiplas de Gardner. Esses alunos eram selecionados por perfil, mediante respostas obtidas por meio de aplicação de questionário com perguntas do cotidiano. Com o resultado da classificação, os alunos eram direcionados para ambientes de ensino à distância em que eles provavelmente iriam melhor se adequar. Esse sistema de ensino foi desenvolvido por um aluno de doutorado da Unesp de Ilha Solteira e o meu “papel” foi de contribuir com a ferramenta de classificação para a inserção dos alunos nesse ambiente. Para isso, utilizei um Software programável de classificação conhecido por Weka e o algoritmo era fundamentado na Lógica Fuzzy.

2017

No início de 2017, motivado pela conclusão do mestrado, iniciei o doutorado. Escolhi o programa de pós-graduação oferecido pela FAAC por me identificar com a linha dois “Tecnologias Midiáticas” oferecida pelo programa.

No primeiro semestre de 2017, iniciei os estudos como aluno especial na disciplina “Audiofonia e Cultura Brasileira: da Modinha ao Ciberespaço” oferecida pelo professor Dr. Dino. No segundo semestre de 2017 – após conhecer o professor Dr. Morgado que atua na linha dois do programa de doutorado em Mídia e Tecnologia – manifestei a vontade de trabalhar junto a ele nos grupos de pesquisa gerenciados por sua administração e, ao mesmo tempo, auxiliar no LTIA – Laboratório de Tecnologia da Informação Aplicada.

Dessa maneira, o professor Dr. Morgado orientou-me a cursar as disciplinas oferecidas por ele, para que ambos conhecessem os trabalhos um do outro. Nesse semestre, cursei a disciplina “Inovação e Tecnologias Emergentes”.

2018

No primeiro semestre de 2018, cursei a disciplina “Desenvolvimento de Aplicações Iterativas” também com o professor Dr. Morgado. Nesse período, ele concordou que, caso eu conseguisse uma vaga como aluno regular, que ele seria meu orientador, e decidimos que eu deveria trabalhar com mineração de dados aplicada a algo em que eu poderia definir e propor sob sua orientação.

Como o processo seletivo para aluno regular seria realizado no primeiro semestre de 2019, como forma de continuar ativo no programa, ingressei na disciplina “A Representação do Discurso Tecnológico de Filmes e Séries de Hollywood”, ministrada pelo professor Dr. Marcos Américo (Tuca) coordenador do programa.

2019

Prestei o processo seletivo para aluno regular e, finalmente, depois de muito esforço, consegui a vaga. Nesse mesmo semestre, o professor Morgado e eu decidimos, em comum acordo, trabalhar com a mineração de dados em redes sociais, visando descobrir se jovens adolescentes, que estão cursando o ensino médio, poderiam estar sendo acometidos a alguma síndrome socioemocional, em específico, estresse, depressão ou ansiedade.

Ficou definido que a coleta de dados não deveria ser intrusiva e, sim, de uma forma em que os indivíduos se sentissem à vontade, uma vez que, a maior dificuldade é fazer com que o jovem adolescente possa externar seus sentimentos. A ideia surgiu no período das restrições sociais impostas pela possível chegada de uma pandemia e pelo fato de a Organização Mundial da Saúde (OMS) e em meio aos alertas acerca do aumento no índice de suicídios por parte de jovem adolescentes, na faixa etária dos que estudam no ensino médio.

No ano de 2019, como forma de buscar aprimorar meus conhecimentos na docência, solicitei junto ao professor Dr. Morgado o interesse em participar do estágio de docência nas disciplinas oferecidas por ele na graduação. Com seu deferimento, ministrei junto a ele a disciplina “Ciência de Dados Aplicada”, oferecida aos sábados. Assim, consegui realizar tal etapa nos dois semestres de 2019.

2020

No primeiro semestre de 2020, conclui um ano e meio de estágio em docência na graduação. Com isso minha experiência profissional na área melhorou de forma significativa, sendo que, em nível superior, ela estava atrelada somente a dois anos como professor substituto na Fatec de Lins e eu queria mais. Ainda no primeiro semestre, cursei a disciplina “Ciência de Dados (Data Science)” com o professor Morgado e iniciei o processo de escrita da Tese.

No segundo semestre, continuei a escrita da Tese sempre sob orientação do professor Morgado e cursei a disciplina “Seminário Avançados em Tecnologias Midiáticas” oferecida pela professora Dr. Roberta. Essa disciplina contribuiu muito para meu aprendizado em realizar pesquisas a fim de selecionar referências bibliográficas, como realizar citações e como organizar o conteúdo da Tese que estava ainda desorganizado e sem uma estrutura definida.

2021

No primeiro semestre de 2021, dediquei-me total e integralmente na escrita da Tese. Em meio a tudo isso, desde 2020, o professor Dr. Morgado e eu havíamos iniciado algumas negociações com a Empresa SAS. Apresentamos a proposta do projeto de pesquisa dessa Tese para algumas pessoas de setores administrativos, com o intuito de que a empresa, de repente, demonstrasse interesse no projeto de pesquisa e o financiasse por meio de uma licença de uso do *SAS Enterprise Miner*, ferramenta de mineração de dados de alto desempenho e de alto custo.

Depois de muitas reuniões, por meio de uma funcionária da empresa, chamada Camila, conseguimos o financiamento e direito de uso da ferramenta de mineração de dados SAS Enterprise Miner.

Em meio às ocorrências, durante esses anos de doutorado, desde o ingresso como aluno especial, consegui a publicação de um artigo intitulado “Uma análise da influência da propaganda na cognição das pessoas como parâmetro para a melhoria dos modelos de referência aplicados ao planejamento e organização da propaganda”. Página 238 Referente ao Livro: Políticas Públicas, Modelos de Negócio e Inovação em Ambientes Digitais, Esporte e Mídia. 1ª Edição, Editora UNR, 2018.

Publiquei ainda, em parceria com outro aluno do programa, o artigo intitulado “*Smart Cities: Um estudo sobre a transformação digital de cidades através da Internet das Coisas*”. Página 205 Referente ao Livro: Audiovisual, Cidades, Mobilidade, Cidadania, Jornalismo, Mídia e Tecnologia. 1ª Edição, Editora UNR, 2018.

Participei do “I Simpósio Internacional de Mídia e Tecnologia: diversidade e qualidade na mídia”, realizado na Universidade Estadual Paulista “Júlio de Mesquita Filho”- Unesp, Brasil nos dias 22 e 23 de agosto de 2018.

Participei do Congresso Internacional “Um Novo Tempo na Educação”, realizado na modalidade on-line com sede e transmissão na cidade de Curitiba. O Evento foi realizado nos dias 15 e 16 de julho de 2020, com carga horária de 16 horas.

Publiquei um capítulo de livro em parceria com outro doutorando da FAAC, Thiago Seti Patrício intitulado “Ubiquidade na Era da Informação: um estudo acerca de aplicações hodiernas de Internet das Coisas (IoT)”. Esse capítulo pode ser encontrado no livro Tecnologia e Mídia Digitais na Educação: Conceitos práticos e teóricos, Capítulo 9, página 96, editora Uniedusul no ano de 2021.

2022

Publiquei, também, mais um capítulo de livro em parceria com outro doutorando da FAAC, Thiago Seti Patrício intitulado “O Potencial Didático da Utilização da Realidade Aumentada na Educação”. Esse capítulo pode ser encontrado no livro “Perspectivas da Educação Brasileira no Século XXI”, capítulo 12, página 170, volume 2. Editora UNIEDUSUL, 2022.

Além disso, consegui outras parcerias, publicando com outros doutorandos da FAAC, sendo eles os alunos Thiago Seti Patrício e Anderson Pazin e com os mestrandos Ricardo Amauri Honorato, Rafael Hamamura, intitulado “Utilização de Técnicas de mineração de dados para Análise do Comportamento das Pessoas em Uma Rede Social com o Intuito de Determinar Possíveis Índícios de Síndromes Socioemocionais”. Esse capítulo pode ser encontrado no livro “Ciências Exatas e Inovação: Perspectivas de Desenvolvimento e Aplicabilidade em Sociedade 2”, capítulo 13, página 178, volume 2. Editora e-Publicar, 2022.

1. INTRODUÇÃO

Segundo Laurie Saloman (2019), estudos realizados pela Organização Pan Americana da Saúde, que atua em parceria com Organização Mundial da Saúde (OMS), determinaram que uma grande parcela dos problemas de saúde mental começa aos 14 anos. Problemas como a Depressão e a Ansiedade podem contribuir para que as pessoas desenvolvam quadros clínicos e sintomas frequentes de irritabilidade, de frustração ou episódios de raiva excessivos e de difícil controle, podendo, ainda, levar ao suicídio.

Ainda de acordo com a autora, uma pesquisa divulgada pelo Instituto Nacional de Saúde dos Estados Unidos no ano de 2019, revela que um em cada três adolescentes de 13 a 18 anos deverá sofrer de algum transtorno relacionado à Ansiedade ao passar por essa faixa etária. Vale destacar que os transtornos gerados pela dificuldade em gerir as emoções podem ser profundamente incapacitantes trazendo sérios problemas no futuro dessas pessoas, visto que, provavelmente, distúrbios emocionais contribuiriam para afetar o estado geral de saúde delas.

Associado a isso, a adolescência é a fase do ciclo de vida de um indivíduo considerada como a mais crucial para o aprendizado, o desenvolvimento e a manutenção de grande parte dos hábitos sociais e emocionais mais importantes para a boa saúde mental, (ZACHARIAS, 2020). Sendo assim, devem ser adotados padrões de sono saudáveis, realização regular de exercício físico e o desenvolvimento de habilidades que possibilitam o gerenciamento de dificuldades no futuro.

É esperado, assim, que os pais e/ou responsáveis pela evolução e educação de seus filhos possam proporcionar a eles, o aprendizado de formas de resolução de problemas interpessoais. Esse aprendizado é essencial para gerenciar emoções e conseguir lidar melhor com as adversidades da vida, sobretudo quanto às eventuais frustrações que possam ocorrer. Assim, os ambientes familiares e educacionais são vistos como suporte de proteção ou de risco no que se refere à adequação do comportamento dos jovens ao âmbito social.

Segundo Zacharias (2020), a ausência de tratamento da saúde mental não só durante a juventude, como também na fase adulta, pode gerar um grande impacto no desempenho educacional, pessoal e profissional, aumentando o risco do consumo de álcool ou drogas pelas pessoas, além de, poder gerar comportamento violento.

Dados gerados pela Organização Mundial de Saúde (OMS), ainda de acordo com Zacharias (2020), apontam que a maioria dos transtornos mentais que afeta as pessoas não é diagnosticada e nem tratada, e os sintomas por muitas vezes são prescindidos por amigos e familiares. Essa é uma realidade que desafia os gestores de saúde pública nos países desenvolvidos, em desenvolvimento e subdesenvolvidos.

No Brasil, os números chamam a atenção para a intrínseca relação entre os casos confirmados de distúrbios mentais e a evolução para o suicídio. Segundo a Organização Pan Americana de Saúde, o suicídio já está classificado em terceiro lugar como a principal causa de morte em jovens de 15 a 19 anos.

Assim sendo, a prevenção começa com a detecção dos primeiros sinais e/ou sintomas de alerta de um transtorno mental. Ela pode ser realizada por familiares, em casa, por professores, nas escolas e, ainda, por profissionais na área da saúde. Nas escolas, a atenção deve ser voltada para ajudar os alunos na construção de habilidades para solução de problemas sociais, na adolescência e na vida adulta, ajudando a lidar com os desafios do dia a dia em casa e na escola. Além disso, é importante que seja realizado investimento em criar um apoio psicossocial, oferecido nas escolas e outros ambientes comunitários, dessa forma, o treinamento de profissionais para a detecção e gerenciamento de transtornos de saúde mental também pode ser implementado, aprimorado ou ampliado.

O investimento por parte dos governos e o envolvimento dos setores social, de saúde e de educação em programas abrangentes, integrados e baseados em evidências para a saúde mental das pessoas, é essencial, de acordo com a pesquisadora Laurie Saloman (2019). Ele deve estar vinculado aos programas de conscientização de adolescentes e jovens sobre formas de cuidar da saúde mental e ajudar colegas, pais e professores a apoiar seus amigos, filhos e alunos. Este foi o foco do Dia Mundial da Saúde Mental do ano de 2018.

Todo esse contexto foi pesquisado, entendido e analisado para que esta Tese de doutorado fosse elaborada. A ideia principal desse trabalho é apresentar uma proposta de utilização das técnicas e métodos de mineração de dados, realizada de forma não intrusiva, para detectar síndromes socioemocionais, tais como: depressão, ansiedade e estresse. Para isso são utilizadas duas ferramentas de mineração de dados conhecidas por *SAS Enterprise Miner* e *SAS Visual Text Analytics*, ambas fornecidas pela empresa SAS, que colaborou com esse projeto de pesquisa.

1.1 Motivação

Conforme Papalia et al. (2010), a adolescência é definida como o período de transição do desenvolvimento entre a infância e a vida adulta, marcado por importantes alterações físicas, cognitivas, psicológicas e sociais, todas elas inter-relacionadas.

Devido a diversos fatores e mudanças sociais que vêm ocorrendo com o passar dos anos, os jovens estão, cada vez mais, trocando atividades correspondentes à adolescência, por outras responsabilidades e afazeres. Lazer, diversão e, até mesmo os estudos, estão sendo substituídos por atividades e responsabilidades que deveriam ser atribuídas a eles, em um estágio de suas vidas em que as experiências adquiridas possibilitassem assumir tais responsabilidades e desafios.

Essas pesquisas sobre a adolescência tiveram início com a publicação do primeiro estudo conhecido sobre a adolescência em 1904, intitulado “*Adolescence*”. Nessa obra, segundo Just et al. (2015), a adolescência foi descrita como um período marcado por grande estresse emocional e, no transcorrer do século passado, muitos trabalhos científicos fortaleceram esta visão, enfatizando as dificuldades e os problemas manifestados nesse momento da vida e que persistem na fase adulta, em alguns casos por toda a vida, ou ainda, contribuindo para que a vida dessas pessoas possa ser encerrada de forma precoce.

As causas desse estresse podem vir: da pressão sobre os jovens, para que eles tenham comportamento de pessoas adultas – antes do momento correto – das dificuldades de pessoas adultas e ativas para suprir suas necessidades básicas de sobrevivência, bem como as de seus familiares; ou de dificuldades financeiras, problemas de saúde, entre outros fatores. Dessa forma, o estresse pode contribuir para que transtornos emocionais, tais como, a depressão e a ansiedade surjam nessas pessoas.

Em conformidade com a OMS (organização mundial da Saúde), o índice de suicídios na sociedade, aumentou significativamente nos últimos nove anos. Pesquisas indicam que, entre 2011 e 2020, esse índice teve aumento de 24% principalmente entre jovens e a OMS trabalha em escala global para a redução dessas taxas.

Algumas pessoas, jovens e adultos, por falta de preparo ou por receio de serem discriminadas pela sociedade, acabam se retraindo e, em sua grande maioria, não

demonstram sinais claros e perceptíveis quando estão passando por problemas emocionais. Essas pessoas, pela insegurança, acabam por não externar essa dificuldade, não comentando sobre sua situação emocional com amigos, por receio do preconceito e, ao se sentirem sem condições de superar a situação em que imaginam estar submetidos, acabam recorrendo ao suicídio.

Considerando que a grande maioria das pessoas, principalmente os jovens, está introduzida em redes sociais, a mineração de dados – de forma não intrusiva e realizada a partir dessas redes – pode contribuir para geração de parâmetros, os quais fornecerão indícios para detectar se as pessoas foram acometidas de algum tipo de transtorno emocional, como depressão, estresse e ansiedade.

Consoante a essa linha de pesquisa, Rodrigo et. al. (2012) utilizou uma rede social conhecida por *Microblog Twitter* para obter informações relacionadas às Olimpíadas de 2012. Utilizando-se dos meios já descritos, este trabalho tem como objetivo específico fazer uso dos dados produzidos nessa rede, para obter informações relacionadas às olimpíadas de 2012.

Daniela (2021) apresentou em seu artigo, um estudo sobre a viabilidade da utilização de ferramentas de mineração de dados em redes sociais no intuito de melhorar a prática pedagógica do grupo discente no monitoramento do aluno em um ambiente de aprendizagem colaborativa. Para isso, a autora realizou um estudo sobre esse procedimento aplicado às redes sociais e sua utilização nesse contexto.

Nessa direção, Vieira (2021) também realizou estudo sobre as redes sociais e a desinformação. Para isso, realizou uma análise em um grupo do *Facebook*, utilizando a mineração de dados, objetivando analisar os subtemas mais comunicados no grupo e o que eles revelam sobre as controvérsias em torno das vacinas. Foram coletados dados de um grupo antivacina atuante no *Facebook* utilizando-se essa metodologia, sendo possível obter as temáticas mais faladas naquele espaço entre os anos de 2015 e 2019.

Alguns padrões foram encontrados, revelando, além de outros assuntos em circulação, dois subtemas importantes presentes no grupo: primeiramente, termos evidenciando que a antivacinação seria um assunto intimamente ligado às relações familiares; assim como foi descoberta a recorrência das palavras “sarampo” e “autismo”, ocorrências que podem remeter a controvérsias existentes em torno da vacina tríplice viral, motivadas por desinformações fomentadas desde o final da década de 1990.

Dessa forma, este trabalho tem por motivação apresentar um método de mineração de dados não intrusivo, a partir de dados coletados provenientes de uma rede social denominada por “*Twitter*”. Sendo assim, esta tese descreve todo o procedimento realizado e que foi aplicado para o estudo do comportamento dessas pessoas em uma determinada rede social digital, como também, fornece uma possível resposta para entender de qual forma a mineração de dados pode contribuir para identificar sintomas de depressão, estresse e ansiedade em pessoas que fazem uso dessas redes.

1.2 Objetivos

O objetivo geral desse trabalho é realizar estudos acerca do estado da arte com relação à utilização da mineração de dados e, a partir desses estudos, fornecer, por meio das técnicas escolhidas, um método não intrusivo para obter informações que possam auxiliar na análise do estado emocional das pessoas que utilizam a rede social *Twitter*, investigando se estas estariam sendo acometidas por transtornos emocionais, tais como: depressão estresse e ansiedade.

Os objetivos específicos desse trabalho são:

- a. Utilizar técnicas e ferramentas de mineração de Dados para busca de padrões emocionais em redes sociais digitais;
- b. Conseguir obter informações do comportamento das pessoas sem utilizar métodos intrusivos ao realizar a abordagem do indivíduo;
- c. Pesquisar, no mercado tecnológico, opções de ferramentas computacionais que podem ser utilizadas para a mineração de dados e entender os métodos estatísticos utilizados;
- d. Contribuir com uma das propostas da OMS (Organização Mundial da Saúde), que consiste em fornecer uma ferramenta que possa contribuir com a diminuição da taxa de suicídios entre jovens e adultos, a partir da detecção de perfis suspeitos;
- e. Associar a mineração de dados à área de Psicologia com intuito de contribuir para a saúde socioemocional;
- f. Contribuir para que psicólogos e profissionais da área da Saúde de modo geral, possam determinar se usuários dessas redes sociais

possam estar passando por problemas psicológicos provenientes de Transtornos Emocionais.

1.3 Justificativa

Os estudos necessários para a elaboração dessa tese, devem contribuir para que se possa entender, mesmo que de forma parcial ou preliminar, como utilizar técnicas e métodos de mineração de dados como uma provável opção para atender a demandas da sociedade, podendo ainda, se tornar uma opção para contribuir para que métodos não intrusivos de diagnóstico de síndromes socioemocionais possam ser testados e utilizados em redes sociais digitais.

A aplicação das técnicas e métodos mencionados, bem como os sistemas de recomendação e de inteligência artificial têm ampla empregabilidade. Podem ser utilizados para a indicação de compra de produtos, análise de mercado, tomadas de decisões em linhas de produção, logística de transportes, entre outras finalidades.

Porém, pouco foi encontrado, quando se considera a utilização da mineração de dados em diagnósticos de Síndromes Socioemocionais. No entanto, a ferramenta utilizada para isso foi fornecida por uma empresa que desenvolve estes sistemas de mineração e afins, justamente por seus diretores, terem interesse nos resultados obtidos nesse trabalho e pela temática ser de interesse para a empresa em uma exploração futura.

Os resultados obtidos provenientes da aplicação da metodologia proposta nesse trabalho, contribuem para fomentar o desenvolvimento de novos projetos que façam uso dos resultados e métodos utilizados, contribuindo com a área da saúde, em específico no tratamento de síndromes socioemocionais.

1.4 Metodologia

Para que os objetivos deste trabalho fossem atingidos, foi necessário que algumas etapas fossem seguidas em uma determinada ordem, são elas:

- a. A elaboração de uma pesquisa bibliográfica sistematizada sobre conceitos de Mineração de Dados.
- b. A elaboração de uma pesquisa bibliográfica sistematizada sobre os conceitos de redes sociais.

- c. A elaboração de uma pesquisa bibliográfica sistematizada sobre os conceitos de Transtornos Emocionais.
- d. Escolha da rede social digital para obtenção da base de dados.
- e. Realização da extração dos dados mediante escolha de ferramenta adequada.
- f. Realização do tratamento da base de dados, para que essa pudesse ser acessada pelas ferramentas de mineração de dados.
- g. Realização do pré-processamento da base de dados, definido percentuais dessa base para teste, treinamento e validação dos Dados, garantindo a integridade dos Dados para posterior mineração.
- h. Realização do processo de mineração da base de dados, obtendo resultados para análise.
- i. Realização da análise dos resultados obtidos, validando a proposta apresentada neste trabalho.

1.5 Organização do trabalho

O presente trabalho foi desenvolvido com a seguinte estrutura:

Capítulo 1 – Introdução – Nesse capítulo é apresentada a introdução, a motivação, os objetivos subdivididos em geral e específicos, a justificativa, a metodologia por fim um item que descreve como está organizado esse trabalho.

Capítulo 2 – Fundamentação teórica – Nesse capítulo, são apresentadas as informações que norteiam e compõem a fundamentação teórica. São expostos conceitos referentes à Descoberta de Conhecimento em Banco de Dados (KDD), mineração de dados, redes sociais, pesquisa bibliográfica sobre transtornos emocionais, a fim de que consigam ser entendidos os sintomas e o perfil comportamental das pessoas que podem estar sofrendo de algum transtorno emocional, em específico: a ansiedade, o estresse e a depressão. Por fim, são apresentadas as ferramentas utilizadas para a mineração de dados e a justificativa da escolha dessas ferramentas.

Capítulo 3 – metodologia – Nesse capítulo é descrita a metodologia utilizada para a realização dos testes. Inicialmente, é definida a rede social digital e ser utilizada para gerar a base de dados, é realizada a coleta dos Dados, preparação dos Dados, divisão da base de dados em percentuais para treinamento, validação e de mineração. Em seguida, são definidos os parâmetros para mineração dos Dados, que servem de

norte para a busca por padrões comportamentais que servem de indícios de que indivíduos estão passando por transtornos emocionais, nomeadamente – depressão, estresse e ansiedade. Ao final de todas essas etapas e, ainda dentro desse capítulo, são realizados testes e por conseguinte, resultados são obtidos.

Capítulo 4 – Resultados Obtidos – Nesse capítulo são apresentados e discutidos os resultados obtidos pelo processo de mineração de dados. Por fim, é realizada, também, a análise comparativa entre os resultados obtidos pelas duas ferramentas de Mineração utilizadas.

A conclusão e as propostas para trabalhos futuros são apresentadas logo após o quarto capítulo.

2. FUNDAMENTAÇÃO TEÓRICA

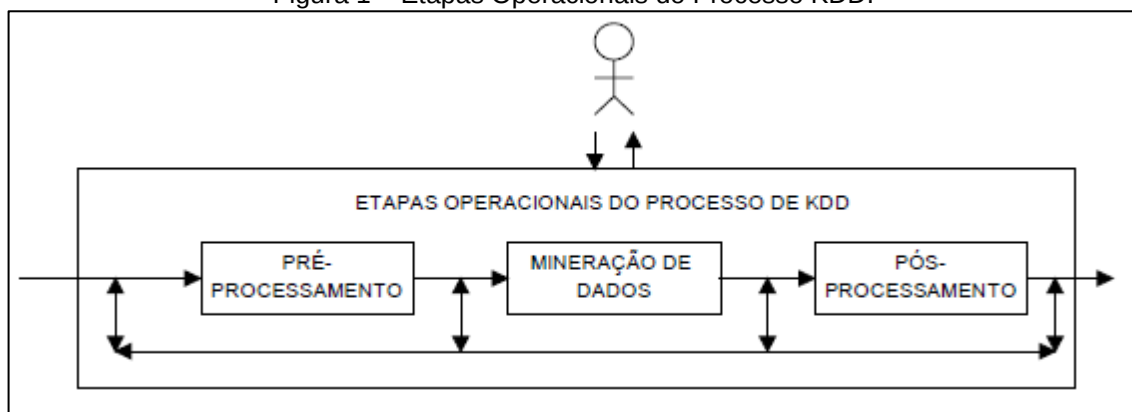
Nesse capítulo são apresentados os conceitos de Descoberta de Conhecimento em Banco de Dados (KDD), de mineração de dados e de redes sociais; além de pesquisa bibliográfica sobre transtornos emocionais, com o intuito de que consigam ser entendidos os sintomas e o perfil comportamental de pessoas que possam estar sofrendo de depressão, estresse e ansiedade. Por fim, é apresentada a ferramenta utilizada para a mineração de Dados e a justificativa da escolha por essa ferramenta.

2.1 DESCOBERTA DO CONHECIMENTO EM BANCO DE DADOS

A descoberta do Conhecimento em Banco de Dados, também conhecida por *KNOWLEDGE DISCOVERY DATABASE* (KDD), é o processo que tem como principal objetivo extrair conhecimento de uma ampla base de dados. Segundo Fayyad et al. (1996), a extração de conhecimento, a partir de uma base de dados, consiste em um processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nesses dados.

De acordo com Goldschmidt e Passos (2005), o processo de Descoberta de Conhecimento em um banco de dados pode ser definido como um processo composto por várias etapas operacionais. Por meio da figura 1, é possível observar a configuração resumida das etapas operacionais realizadas neste processo:

Figura 1 – Etapas Operacionais do Processo KDD.



Fonte: BRACHMAN, R. J. e ANAND, T. (1996).

Conforme expõe Brachman e Anand (1996), a primeira etapa é o pré-processamento em que são ensinadas as funções de compreensão, a organização e o tratamento dos dados, com a finalidade de prepará-los para a etapa seguinte, que consiste na mineração de dados.

Na etapa de mineração é executada uma busca por informações e conhecimentos importantes no ambiente da aplicação do Processo de Descoberta de Conhecimento. Por último é realizado o pós-processamento, que engloba o tratamento das informações e do conhecimento obtidos com a etapa anterior, com o objetivo de proporcionar a avaliação do benefício do conhecimento descoberto.

2.2 MINERAÇÃO DE DADOS

A mineração de dados, também conhecida por *Data Mining*, é uma área de estudo que contribui para o processo de Descoberta de Conhecimento em Bancos de Dados e pode ainda ser utilizada e estudada separadamente. Ela possui diversas técnicas e, por conseguinte, é capaz de realizar diversas tarefas.

Segundo Parthasarathy (2007), essa é uma área de estudo que surgiu no final dos anos 80 como uma nova disciplina da Computação. Ela tem se desenvolvido e ganhado espaço no tratamento de diversos problemas, em que é necessária a busca por padrões por meio do cruzamento de informações que, em sua maioria, estão dentro de grandes volumes de dados, sendo estes agrupados em um servidor ou distribuídos por meio da computação em nuvem.

Para isso, combina-se técnicas de diferentes áreas, como: estatística, inteligência artificial, aprendizagem de máquina e tecnologia de gerenciamento de base de dados para lidar com questões teóricas e práticas da extração de conhecimento, padrões e relacionamentos em grandes volumes de dados, Conforme Sumathi e Sivanandam (2006).

Ainda nesse sentido, Fayyad et al. (1996) acreditam que a mineração de dados não é um processo trivial, ele consiste na habilidade de identificar, nos dados, os padrões válidos, potencialmente úteis e compreensíveis, envolvendo métodos estatísticos, ferramentas de visualização e técnicas de inteligência artificial.

2.2.1. Técnicas de mineração de dados

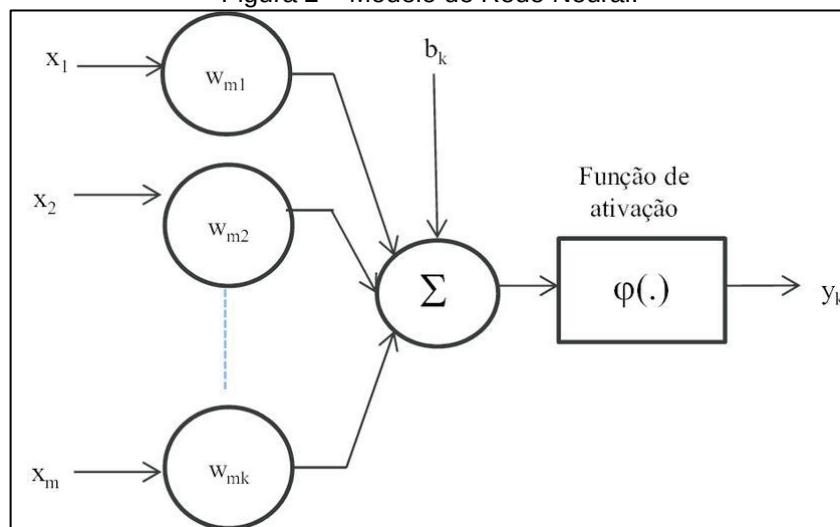
Consoante o expõe de Rodrigues (2000), existem diversas técnicas de mineração de dados que, ao serem aplicadas, devem ser capazes de produzir bons resultados, transformando dados em conhecimento útil e favorecendo as práticas baseadas em evidências. São várias as técnicas existentes; entretanto, são citadas as mais utilizadas: Rede Neural Artificial, Árvore de Decisão, Algoritmos Genéticos (AGs), Lógica Nebulosa (Lógica *Fuzzy*), Análise Estatística e Redes Complexas. Os parágrafos a seguir descrevem sucintamente os cinco primeiros métodos. Para esse trabalho, foram escolhidas as Redes Complexas, que são explicadas de forma mais detalhada na seção 2.2.1.1.

Segundo Vale (2008), a Rede Neural Artificial, por exemplo, é uma técnica computacional que faz uso de um modelo matemático, inspirado no funcionamento do cérebro humano. Ela é utilizada para reconhecimento de imagens e sons, com capacidade de aprendizado, generalização, associação e abstração, constituído por sistemas computacionais paralelos e distribuídos em unidades de processamento.

Essas redes têm contribuído com sucesso para modelar relações envolvendo séries temporais complexas em várias áreas do conhecimento. A grande vantagem sobre as demais técnicas convencionais é que, conforme Abbott e Lee (2006), elas não requerem informações detalhadas sobre os processos do sistema a ser modelado, sendo este descrito explicitamente na forma matemática, além de serem robustas e terem uma alta taxa de acurácia preditiva. Por meio de repetidas apresentações dos dados à rede, elas aprendem padrões, procuram por relacionamentos e podem, ainda, construir modelos automaticamente.

Ainda sobre isso, de acordo com Pereira et al. (2008), as Redes Neurais relacionam pesos sinápticos às conexões entre neurônios; esses pesos funcionam como as terminações nervosas de entrada em um neurônio. Eles mudam por meio de algoritmos de aprendizagem na medida em que novas informações ou novas observações são incorporadas à rede. O modelo de rede neural espelha os pesos sinápticos que possibilitam, depois de ajustes regulares de uma função de ativação, a implantação de uma variável de saída ou de resposta a partir de dados de entrada. Por meio da figura 2 é possível observar um exemplo disso:

Figura 2 – Modelo de Rede Neural.



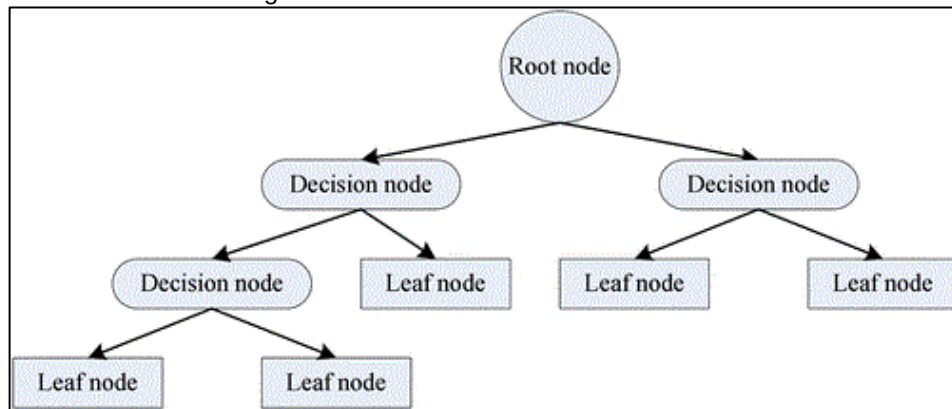
Fonte: HAYKIN, 2001, p.36.

Árvore de decisão é uma técnica de mineração de dados que utiliza um modelo representado graficamente por nós e por galhos, parecida com uma árvore, mas no sentido invertido. Em concordância com Meira et al. (2008), essa técnica também é conhecida pelo nome de árvore de classificação ou de regressão, caso a variável dependente seja categórica ou numérica, respectivamente. O modelo de conhecimento que tem em cada nó interno da árvore representa uma decisão sobre uma variável, determinando como esta será alocada na árvore, passando por uma série de galhos (nós filhos) até ser disposta em algum ponto específico.

Segundo Meira et al. (2008), existem duas finalidades quando se utiliza uma árvore de decisão. Uma é produzir um modelo de predição eficaz; e a outra é descobrir a estrutura preditiva de um determinado problema.

Através da utilização dela, pode-se compreender quais variáveis e suas interações podem conduzir ao fenômeno estudado. Por fim, em um mesmo estudo é possível que as duas finalidades possam aparecer juntas. Algumas pesquisas recentes fazem uso de técnicas de indução na árvore de decisão para prever e obter conhecimento. Por meio da figura 3 é possível visualizar um modelo do que dissemos:

Figura 3 – Modelo de Árvore de Decisão.

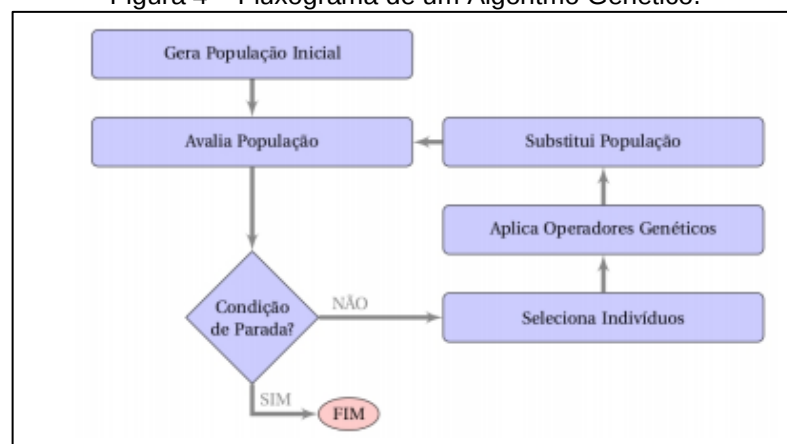


Fonte: SYACHRANI; JEONG; JUNG, 2012, p.636.

Os algoritmos genéticos, de acordo com Goldschmidt e Passos (2005), podem ser utilizados para formular estratégias de otimização inspiradas nos princípios observados na evolução natural e na genética para solução de problemas. Eles fazem uso de operadores de seleção, de cruzamento e de mutação; dessa forma, o algoritmo vai se reproduzindo para desenvolver sucessivas gerações de solução. Com a evolução deste, somente as soluções com maior poder de previsão sobrevivem, até convergirem numa solução ideal.

Conforme salienta Costa et al. (2013), por meio da figura 4, é possível observar um fluxograma que representa a estrutura de um algoritmo genético, formado por um ciclo que define as gerações de indivíduos em que são efetuados os elementos básicos dessa técnica, tais como: a função de avaliação, a seleção de indivíduos a integrar a nova população, os operadores genéticos e a troca da antiga população pela nova que foi gerada, criando indivíduos mais aptos.

Figura 4 – Fluxograma de um Algoritmo Genético.



Fonte: COSTA et al., 2013, p.15.

Segundo Han e Kamber (2006), outra técnica de mineração de dados muito utilizada é a Lógica Nebulosa ou Lógica *Fuzzy*, sendo uma teoria matemática que permite realizar uma modelagem do modo aproximado de raciocínio, imitando a habilidade humana de tomar decisões, em ambientes de incertezas e imprecisão. Com isso, pode-se construir sistemas inteligentes de controle e de suporte à decisão.

Consoante a patente de Han e Kamber (2006), a Lógica Nebulosa pode ser utilizada de várias formas; no entanto, existem duas delas que são mais recorrentes como: representar a extensão da lógica clássica para uma mais flexível, com objetivo de formalizar conceitos imprecisos, e aplicar conjuntos *Fuzzy* a diversas teorias e tecnologias para processar informações imprecisas como, por exemplo, em processos de tomada de decisão.

Conforme expõe Lopes (2010), quando se aplica a teoria clássica dos conjuntos *Fuzzy*, um elemento pode pertencer, ou não, a um conjunto, sendo essa pertinência representado por 1 (pertence) ou 0 (não pertence) aos elementos de um conjunto universo. O problema dessa abordagem é que muitas vezes essa definição binária não é suficiente para descrever um elemento em várias situações do mundo real. Nesse caso, a teoria mencionada acima estende a Teoria Clássica de Conjuntos e proporciona que um elemento possa pertencer a mais de um conjunto com diferentes graus de pertinência para cada um deles. Este grau pode variar de 0 (zero) a 1 (um) para cada conjunto.

O processo de tomada de decisão utilizado pela lógica *Fuzzy* recebe o nome de Inferência. Nesse processo, um elemento pode ser classificado em mais de um grupo conforme regras de tomada de decisão que utilizam proposições condicionais (se...então...Senão) e conectivos (“e” ou “ou”). Sendo assim, todo elemento é submetido a todas as regras de tomada de decisão e, conforme os parâmetros utilizados, ele pode ser classificado como pertencente a um ou mais grupos de elementos. O percentual do quanto cada elemento pertence a cada grupo pode ser definido conforme a aplicação de regras de inferência *Fuzzy*.

Segundo Cardoso e Machado (2008), a análise estatística é uma das técnicas mais tradicionais nos processos de mineração de dados e, quando utilizada corretamente, pode proporcionar modelos para análise e interpretação de dados. O princípio estatístico consiste em fornecer uma maneira pela qual se pode estimar a probabilidade de um evento a partir de dois tipos de conhecimento. Dentre os diversos

modelos estatísticos, os mais utilizados são: redes bayesianas, análise discriminante e análise exploratória de dados, porém, existem diversos outros.

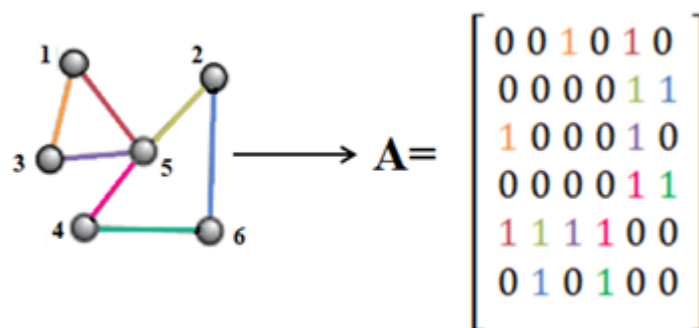
Dentre os modelos estatísticos mais utilizados, as redes bayesianas emergiram em anos recentes como uma poderosa tecnologia e, se aplicadas corretamente, podem produzir representações gráficas de distribuições probabilísticas derivadas da contagem da ocorrência dos dados num determinado conjunto, representando um relacionamento de variáveis, conforme nos dizem Abbott e Lee (2006).

2.2.1.1 Redes Complexas

A teoria das redes complexas tem por objetivo representar sistemas complexos, utilizando para isso, grafos que por sua vez, são constituídos por um conjunto de vértices (nós) que são interligados por arestas (*links*) sendo que a existência deles deve ser relacionada a algum tipo de iteração.

Segundo Albert e Barabási (2002) em termos computacionais, redes complexas podem ser representadas por meio de matrizes de adjacência. Por exemplo, considere uma matriz (G), com $n \times n$ elementos A_{ij} , cujos valores são 1 ($A_{ij} = A_{ji} = 1$), caso os nós i e j estiverem conectados, e 0 caso contrário. Por meio da figura 5 é possível visualizar a representação do grafo de uma Rede Complexa por meio de uma matriz.

Figura 5 – matriz de representação do grafo de uma rede complexa

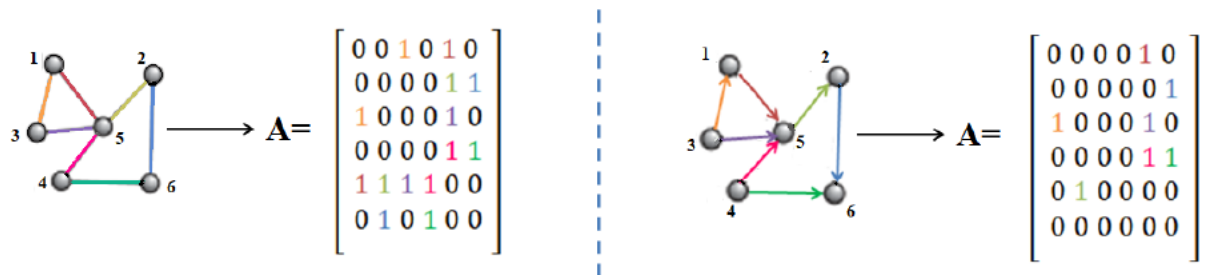


Fonte: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-07032019-102825/publico/CarolineLourencoAlves_revisada.pdf

Outro aspecto que deve ser considerado e que é relevante ser entendido, para que se possa aprofundar nesse quesito, é que existem dois de redes complexas, são elas: as redes dirigidas e as não dirigidas.

Ao analisar dois grafos diferentes, é possível entender e diferenciar esses dois tipos. Por meio da figura 6, é possível observar um que representa uma rede complexa não dirigida (grafo da esquerda) e outro que representa uma rede complexa dirigida (grafo da direita).

Figura 6 – Exemplo de grafo para redes complexas dirigidas e não dirigidas



Fonte: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-07032019-102825/publico/CarolineLourencoAlves_revisada.pdf

Em resumo, o que pode ser observado e que diferencia um do outro, é que o grafo não dirigido possui arestas sem indicação de direção e o outro possui em suas arestas uma “seta”, que indica a direção que deve ser seguida para mapear o grafo. Por conseguinte, a representação computacional em forma de matriz é diferente para cada um dos dois tipos.

Com relação a representação das redes complexas por meio de grafos, é importante que diversas medidas sejam analisadas. Essas medidas são fundamentais para caracterizar, classificar e modelar as redes e podem ser diferentes para cada um dos dois tipos de redes complexas.

2.2.1.1.1 Conectividade

A conectividade pode ser definida também como sendo o grau. Este, por sua vez, corresponde ao número de conexões de um vértice i , onde i representa a cardinalidade de sua vizinhança. Em termos da matriz de adjacência, o valor para redes complexas não dirigidas pode ser obtido pela equação 1:

$$K_i = \sum_{j=1}^N (A_{ij}) \quad (1)$$

Para o cálculo da conectividade para redes complexas dirigidas, deve-se considerar o número de arestas que saem do nó K_i^{out} , e o número de arestas incidentes ao nó, K_i^{in} . Nesse caso, o grau total, K_i , pode ser obtido pela equação 2.

$$k_i = k_i^{out} + k_i^{in} \quad (2)$$

A partir da equação 1 é viável calcular a conectividade média global, definida por K , cujo valor pode ser obtido pela equação 3.

$$K = \frac{1}{N} \sum_{i=1}^N (k_i) \quad (3)$$

É possível obter a distribuição dos graus P_k , que representa a probabilidade de um vértice escolhido aleatoriamente ter um grau K . Considerando uma rede com N nós, a distribuição de graus P_k pode ser obtida pela equação 4, onde N_k , é um número de nós com grau K .

$$P_k = \frac{N_k}{N} \quad (4)$$

Além disso, é possível, ainda, medir o quão conectada é a rede, essa medida pode ser obtida pela divisão do segundo momento da distribuição do grau pela média desta. Pode ser calculada a entropia (incerteza de uma variável aleatória) da distribuição do grau H , definida pela equação 5.

$$H = - \sum_i (P k_i) \log(P k_i) \quad (5)$$

Uma das mais importantes propriedades quando se trata de redes complexas é que a distribuição de graus deve seguir as Leis de Potência. Uma forma eficiente de quantificá-la é por meio da medida de “retidão”. Ela é determinada pelo cálculo do coeficiente de Pearson dos pontos da distribuição de graus na escala logarítmica. Considerando um conjunto de N pontos com coordenadas x e y , o coeficiente de Pearson, r_{xy} , pode ser calculado utilizando a equação 6.

$$r_{xy} = \frac{\sum_{i=1}^N (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sum_{i=1}^N (x_i - \langle x \rangle)^2 \sum_{i=1}^N (y_i - \langle y \rangle)^2} \quad (6)$$

Onde, x_i e y_i são valores do i -ésimo elemento dos vetores x e y , de tamanho n , e $\langle x \rangle$ e $\langle y \rangle$ representam a média de x e y . Para que se possa calcular o coeficiente de Pearson, basta substituir x pelo logaritmo da conectividade k e y pelo logaritmo da probabilidade $P(k)$. O valor de r_{xy} pode variar entre -1 e 1. Segundo Boas (2008), caso o valor seja -1, indica a presença de uma rede complexa com distribuição livre de escala, pois o coeficiente de Pearson tem valor mínimo quando duas variáveis são completamente anti-correlacionadas.

Segundo Barabási (2016), outra forma de utilização do coeficiente de Pearson é para calcular o coeficiente de correlação da conectividade (graus) de ambos os lados da aresta. Esse cálculo é denominado de coeficiente de assortatividade, que varia de -1 a 1. Para os valores positivos, pode-se dizer que a rede complexa é assortativa e, para os valores negativos, ela é definida como sendo uma rede complexa dissortativa. Caso o valor seja igual a 0 (zero) não existe correlação.

Segundo Barabási (2016), em redes assortativas, os *hubs* têm tendência de se conectar a outros *hubs*, isso resulta em um maior grau médio de seus vizinhos mais próximos. Enquanto em redes complexas dissortativas, os *hubs* tendem a se conectar a nós com baixos graus.

Outra medida que pode ser relacionada ao grau é a conectividade média entre os vizinhos, ela é representada por $k_{nn}(i)$. Sendo assim, deve ser considerado que, para um nó i , a conectividade média pode ser calculada pela equação 7.

$$k_{nn}(i) = \frac{1}{k_i} \sum_j^N a_{ij} k_j \quad (7)$$

2.2.1.1.2 Ciclos

Segundo Schank e Wagner (2004), os ciclos são formados por três vértices cuja medição é realizada pelo coeficiente de aglomeração. Essa medida foi introduzida aos tipos de medição aplicados ao contexto de redes complexas. Essa medição pode ser interpretada como sendo a probabilidade de dois vizinhos de um vértice estarem conectados. Uma maneira de calcular essa medida é utilizando a fórmula da transitividade definida pela equação 8.

$$C = \frac{3N_{\Delta}}{3N_3} \quad (8)$$

Onde, $3N_{\Delta}$ é o número de triângulos presentes na rede e $3N_3$ é o número de tuplas conectadas. O valor correspondente a C deve ser sempre um valor que esteja dentro do intervalo: $0 < C < 1$. Existe ainda o coeficiente de aglomeração local que pode ser calculado por meio da utilização da equação 9, que calcula essa média para cada vértice i .

$$C(i) = \frac{3N_{\Delta}(i)}{3N_3(i)} \quad (9)$$

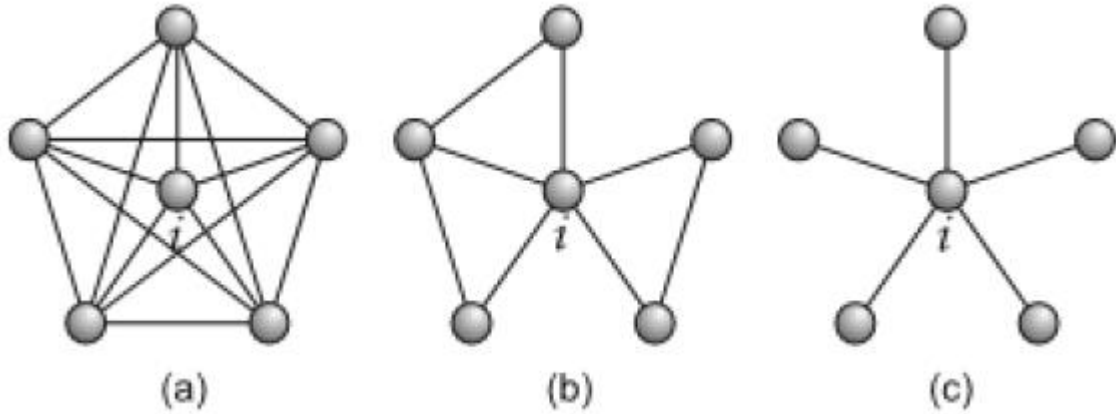
Onde, $3N_{\Delta}(i)$ é a quantidade de triângulos formados envolvendo o vértice i e $3N_3(i)$ é o número de trios conectados e que possuem o vértice i como sendo o vértice central. Utilizando o coeficiente de aglomeração local é possível calcular o coeficiente de aglomeração médio, cuja medida de uma rede complexa não dirigida e sem peso é definida pela equação 10.

$$C = \frac{1}{N} \sum_{i=1}^N C(i) \quad (10)$$

Por meio da figura 7, é possível visualizar três redes complexas com coeficientes de aglomeração distintos. Para figura 7a, todos os vértices estão

conectados. Nesse caso o valor do coeficiente de aglomeração é máximo e igual a 1 ($C=1$). Para a figura 7b, o coeficiente de aglomeração tem valor igual a $C = \frac{3}{10}$ e por fim, para a figura 7c, o valor do coeficiente de aglomeração é igual a zero ($C=0$), isso devido ao fato de não haver nenhum triângulo formado entre os vértices.

Figura 7 – Redes complexas com coeficientes de aglomeração distintos



Fonte: Rodrigues, 2007.

2.2.1.1.3 Distância

As medidas relacionadas à distância em redes complexas, remetem a possibilidade de calcular o diâmetro da rede, que por sua vez pode ser definido como sendo a maior distância entre qualquer par de vértices. Por outro lado, o menor caminho (caminho geodésico) entre eles i e j , d_{ij} é aquele cujo comprimento é o menor de todos os caminhos possíveis entre esses vértices.

A representação do valor do menor caminho entre os vértices i e j é dada por uma matriz de distância D , cujos elementos, d_{ij} , expressam o menor caminho entre estes. O valor máximo $d_{max_{ij}} = \max_{i,j}(d_{ij})$ é chamado de diâmetro da rede. Por conseguinte, a média entre os valores dos elementos matriciais é denominada de caminho característico da rede, representando o menor caminho médio da rede, o qual pode ser calculado por meio da utilização da equação 11.

$$L = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij} \quad (11)$$

É importante ressaltar que, essa medida diverge quando há mais de um componente conexo. Assim, com o intuito de eliminar essa limitação da medida de L , é necessário utilizar uma medida denominada de eficiência da rede. Por meio da equação 12 é possível obter esse valor.

$$E = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}} \quad (12)$$

É importante ter ciência de que E é diferente de L . Isso se deve ao fato de que, E considera o inverso da média harmônica ao invés da média aritmética. Segundo Arruda (2013), a medida de E , também pode ser utilizada como um indicador da capacidade de tráfego na rede. Isso porque, essas medidas estão diretamente relacionadas ao transporte e à comunicação na rede. A analogia é a seguinte: quanto menor o caminho, mais rápido é o transporte de informação pela rede e com isso, maior é a medida de eficiência.

2.2.1.1.4 Centralidade

Segundo Arruda (2013), a partir das medidas das distâncias, é possível expressar o quanto um vértice é importante e central. Uma das medidas de centralidade é denominada de *betweenness centrality* (centralidade de intermediação). Essa medida é definida como sendo a fração de menores caminhos de pares de vértices que passam por um vértice i . Utilizando a equação 13, é possível determinar o valor correspondente a centralidade de intermediação b_i .

$$b_i = \frac{\sum_{s < t} g_i^{(st)} / n_{st}}{(1/2)n(n-1)} \quad (13)$$

Onde $g_i^{(st)}$, é o número de caminhos geodésicos do vértice s ao vértice t passando pelo vértice i . n_{st} é o número total de caminhos geodésicos de s a t e n é o número total de vértices da rede complexa. Existe uma medida denominada de

dinâmica do ponto central (*CDP*) que representa uma medida global relacionada ao grau de intermediação e que pode ser definida pela equação (14).

$$CDP = \frac{1}{N-1} \sum_i (B_{max} - B_i) \quad (14)$$

Onde B_{max} é o maior valor do grau de intermediação da rede complexa e B_i é o valor da centralidade de intermediação de determinado vértice i e n é o total de caminhos geodésicos. O valor da *CPD* deve variar de 0 a 1, sendo que para o valor 0 (zero) as redes se encontram totalmente desconectadas e para o valor 1 (um) indica que existe um vértice central em que todos os caminhos passam.

Outra medida de centralidade é o *closeness centrality* (*Cl*) ou centralidade de proximidade. Por meio da equação (15) é possível obter o valor dela.

$$Cl(i) = \frac{n-1}{\sum_{j=1}^n d_{ij}} \quad (15)$$

No caso acima, d_{ij} é a distância de i a j e a soma é baseada na distância de i a todos os $(n-1)$ pontos. A medida da centralidade de proximidade $Cl(i)$, pode ser definida como sendo o inverso da média da distância de i a todos os demais pontos.

Segundo Newman (2010), existe, ainda, outra medida de centralidade denominada de *eigenvector centrality* (centralidade autovetor), ela se fundamenta no conceito de que a importância de um vértice é aumentada por ter conexões com outros. Dessa forma, ao invés de pontuar um ponto para cada vizinho, a centralidade de autovetor atribui a cada vértice uma pontuação proporcional a soma da pontuação de seus vizinhos. Matematicamente, a centralidade de autovetor de um vértice i , x_i , é proporcional a soma das centralidades de seus vizinhos e pode ser obtida por meio da equação 16.

$$x_i = k_1^{-1} \sum_j A_{ij} x_j \quad (16)$$

Considerando a equação 16, tem-se que A_{ij} são os elementos da matriz de adjacência A e k_1 é o seu maior autovalor. Existe ainda, a medida denominada de *pageRank* que tem por finalidade determinar a medida de importância em páginas da Internet, por meio de um *ranking* para cada página, baseada em um grafo da *World Wide Web*. É importante ressaltar que esse recurso é muito utilizado pela empresa Google. Matematicamente, o *Page Rank* $P_R(i)$ de um vértice i pode ser obtido pela equação 17.

$$P_R(i) = \frac{q}{N} + (1 - q) \sum_j A_{ij} \frac{P_R(j)}{k_{out,J}} \quad (17)$$

Onde A_{ij} são os elementos da matriz de adjacência, $k_{out,J}$ é o grau de saída do vértice, N é o número de páginas da Internet e q é o *damping factor*. Este indica a probabilidade de um determinado usuário parar de navegar na Internet seguindo o caminho dos *links* da página em que ele se encontra navegando e começar uma nova navegação em outra página inserindo outro endereço, segundo nos aponta Cestari (2017).

2.2.1.1.5 K-core

Considerando que existem estruturas de grafos muito grandes, é importante levar em consideração que estas por sua vez, podem se tornar difíceis de serem manipuladas, dessa forma, uma possibilidade de facilitar o processo de análise é segmentá-las em partes menores. Uma das medidas de segmentação hierárquica é o *K-core*.

O *K-core* é definido como o subconjunto máximo de vértices em que cada um deles esteja conectado com pelo menos k outros no subconjunto. Considerando que G é um grafo. Se H é um subgrafo de G , e denotando $\delta(H)$ como sendo o grau mínimo de H . Pode-se dizer que, cada ponto de H é, então, adjacente no mínimo $\delta(H)$ a outros pontos de H . Se H é o subgrafo de conexão máxima (induzida) de G , com $\delta(H) \geq k$, H é o *K-core* de G .

A escolha por utilizar redes complexas se deve, portanto, a dois fatores. Primeiro porque ela possibilita análise e mineração textual. O segundo fator é pela complexidade dos dados a serem minerados nessa Tese. Com isso, essa técnica

possibilita a leitura dos dados que são representados na forma textual, filtrando e separando o que não se pretende analisar e associando o que se pretende buscar, quantificando-o. De forma que é possível visualizar o resultado graficamente, partindo da passagem de parâmetros que são responsáveis por nortear a mineração.

2.2.2. Tarefas em Mineração de Dados

Nesse contexto, quando se fala em “tarefa”, o que se deve entender é que ela é um tipo de mineração de dados com um propósito particular, a partir da qual existem diversas ou até dezenas de implementações distintas por meio de vários algoritmos. Estes são divididos pelas tarefas levando em consideração o objetivo da implementação; ou seja, os algoritmos de uma mesma tarefa possuem a mesma finalidade.

Devido ao fato de que a mineração de dados vem sendo desenvolvida para que possa ser aplicada a diferentes fins, as tarefas também estão sendo diversificadas cada vez mais. O propósito disso é poder oferecer várias formas para extrair diferentes tipos de conhecimento. Dessa maneira, é necessário que antes do início do processo de mineração, seja realizada a escolha da tarefa e do algoritmo a serem utilizados, para que os resultados possam estar de acordo com o que foi proposto.

Como forma de facilitar o entendimento e a separação entre técnica e tarefa, pode-se dizer que a técnica é utilizada com o intuito de determinar como conduzir o pensamento; assim como a ação, para atingir o objetivo com eficiência; ou seja, ela é um plano que contém os procedimentos para que se possa obter as informações desejadas. A tarefa apresenta como se deve agir para utilizar a técnica escolhida, sendo que a implementação da tarefa deve ser capaz de possibilitar que o objetivo seja atingido imediatamente.

Após estudo bibliográfico, é possível determinar que existem diversas tarefas utilizadas para a Mineração de Dados e que todas elas descrevem o tipo de Mineração utilizada. Por exemplo, a utilização de uma técnica requer a escolha de uma tarefa que, em harmonia, são capazes de explicar qual objetivo da aplicação da mineração para solucionar um determinado problema. Com a conclusão dos estudos, é possível obter informações suficientes para apresentar algumas das principais tarefas utilizadas nesse processo.

2.2.2.1 Classificação

Segundo Michie et al. (1994), na tarefa de classificação, existem dois tipos de atributo no conjunto de dados, um deles é do tipo de atributo “previsor” e o outro é do tipo “atributo-alvo”. Para os valores que diferem deste, existe uma classe que faz referência a um grupo categórico relacionado a um conjunto predefinido. O objetivo da tarefa de classificação é descobrir uma função que relacione um conjunto de registros com determinado conjunto de classes, assim que for descoberta essa função é possível aplicá-la a novos registros. Dessa forma, é possível prever a qual classe esse registro pertence. Dentre as diversas técnicas que podem fazer uso da tarefa de classificação, pode-se destacar as rede neurais, os algoritmos genéticos e a lógica nebulosa.

2.2.2.2 Agrupamento

De acordo com Fayyad et al. (1996), a tarefa de agrupamento ou *Clusterização* compreende um processo de partição dos elementos de um Banco de Dados em subconjuntos ou *clusters*. Esse processo é realizado de forma que os registros semelhantes fiquem agrupados diferenciando-se dos registros dos outros subconjuntos. Diferentemente da tarefa de classificação, não existem classes pré-definidas, os elementos são reunidos baseados na similaridade entre eles, sendo a *Clusterização* responsável por verificar as bases de dados. Os algoritmos de *k-Modes*, *k-Means*, *k-Prototypes* entre outros são usados na implementação dessa tarefa.

2.2.2.3 Predição

Conforme Fayyad et al. (1996), a tarefa Preditiva tem por objetivo prever o valor de um determinado atributo ou variável, levando em consideração os valores de outros atributos. O que deve ser predito é comumente conhecido como a variável preditiva, dependente ou alvo, enquanto os atributos usados para fazer a predição são conhecidos com as variáveis preditoras, independentes ou explicativas. De modo mais simples, pode-se dizer que a predição utiliza de uma ou um conjunto de variáveis para prever outras variáveis ou valores desconhecidos.

Na tarefa de predição, a meta é desenvolver modelos que façam inferência sobre aspectos específicos dos dados (variáveis preditivas) por meio da análise e associação dos diversos aspectos encontrados neles (variáveis preditoras).

Existem dois benefícios que estão relacionados à utilização da tarefa de predição e à mineração de dados, segundo Baker (2011). Primeiro, os métodos de predição podem ser utilizados para estudar quais aspectos de um modelo são importantes. Esta estratégia pode, por exemplo, ser utilizada em pesquisas que tentam, de forma direta, prever os benefícios educacionais de determinadas técnicas e ferramentas para um conjunto de estudantes, isso sem considerar os fatores intermediários.

Ainda de acordo com Baker et al. (2008), o segundo benefício de utilizar métodos de predição consiste em auxiliar a prever o valor das variáveis utilizadas em um modelo. O intuito de utilizar essa abordagem é verificar quais dados são mais importantes, pois, analisar todos os dados, de um grande banco de dados, para gerar um modelo, é inviável, seja do ponto de vista financeiro ou referente ao tempo. Dessa forma, o modelo pode ser construído utilizando parte deles e, assim, ser aplicado para trabalhar com dados mais extensos. Esse tipo de técnica pode, por exemplo, auxiliar no desenvolvimento e uso de atividades instrucionais, uma vez que se consegue estimar os benefícios educacionais antes mesmo da atividade ser aplicada aos alunos.

2.2.2.4 Redes Bayesianas

Consoante as vertentes de Bobbio et al. (2001), é importante dizer que as redes Bayesianas – também conhecidas por redes casuais, rede de crença ou gráficos de dependência probabilística – surgiram na década de 80 e têm sido aplicadas em uma grande variedade de atividades do mundo real. Algumas aplicações se estendem a diversas áreas, tais como: finanças, saúde, desenvolvimento de jogos, mineração de dados, entre outras.

Conforme enaltece Neapolitan (2004), essas redes surgiram em um contexto no qual há um grande número de variáveis, tendo por objetivo verificar qual a influência probabilística não direta de uma variável para as demais.

Essa teoria combina, portanto, Segundo Ben-Gal (2007), princípios de teoria de grafos, teoria de probabilidades, ciência da computação e estatística. Além disso,

pode ser considerada como uma representação visual e informativa da tabela de probabilidade conjunta de todas as variáveis que envolvem o domínio do problema.

2.2.2.5 Análise Discriminante

Quando o assunto é mineração de dados, a análise discriminante também pode ser considerada como uma tarefa. Quando utilizada para esse fim, ela deve ser capaz de fornecer uma técnica estatística multivariada utilizada para discriminar e classificar dados.

De acordo com Khattree e Naik (2000), a análise discriminante é uma técnica da estatística multivariada que estuda a separação de objetos de uma população em duas ou mais classes. A discriminação ou separação é a primeira etapa, sendo essa a parte exploratória e de análise que consiste em procurar características capazes de serem utilizadas para alocar objetos em diferentes grupos previamente definidos.

Conforme Johnson e Wichern (1999), essa segunda etapa é a classificação ou alocação, ela pode ser definida como um conjunto de regras que serão usadas para alocar novos objetos. Contudo, é importante ressaltar que a função que separa objetos pode também servir para alocar e, regras que alocam objetos podem ser usadas para separar. Normalmente, discriminação e classificação se sobrepõem na análise e a distinção entre separação e alocação quase sempre é confusa.

O problema da discriminação entre dois ou mais grupos, visando posterior classificação, foi inicialmente abordado em 1936. Um exemplo disso, de acordo com Regazzi (2000), consiste em obter funções matemáticas capazes de classificar um indivíduo X em uma de várias populações π_i , ($i=1, 2, \dots, g$), com base em medidas de um número p de características, buscando minimizar a probabilidade de má classificação, isto é, reduzir a probabilidade de classificar erroneamente um indivíduo em uma população π_i , quando realmente pertence à população π_j , ($i \neq j$) $i, j=1, 2, \dots, g$.

2.2.2.6 Análise exploratória

A tarefa de análise exploratória dos dados – além de ser uma opção muito utilizada na etapa inicial de qualquer tratamento estatístico – pode ser utilizada para que se possa conhecer a estrutura dos dados disponíveis, a fim de que o cientista tenha uma visão mais profunda da natureza destes, pode identificar abordagens mais apropriadas, considerando as limitações que uma base de dados possa conter, conforme nos explica Diggle et al. (1996).

Assim sendo, a utilização dessa tarefa pode dispensar a necessidade da criação de hipóteses e/ou modelos. Ela compreende duas etapas: o cálculo das estatísticas básicas e a representação gráfica dos dados. A primeira delas consiste em desenvolver um sumário estatístico que quantifica a locação e a variabilidade dos dados por medidas de tendência central, de dispersão e de distribuição.

A segunda consiste na representação gráfica dos dados que visa identificar padrões e relacionamentos entre eles. A execução da tarefa de análise exploratória deve ser realizada logo após o manuseio dos dados, de modo a colocá-los em tabelas compatíveis com o método computacional a ser utilizado para o processamento deles.

Por meio da figura 8 é possível observar a associação de algumas das técnicas utilizadas e suas respectivas tarefas:

Figura 8 – Quadro de associação entre técnicas e tarefas utilizadas em Mineração de Dados

Principais Técnicas e suas tarefas		
Técnica	Descrição	Tarefa
Rede Neural Artificial	Fornecer modelos inspirados na fisiologia do cérebro. O conhecimento obtido é proveniente de um mapa das conexões neurais e dos pesos dessas conexões	Classificação Agrupamento Predição
Árvore de Decisão	Fornecer modelos fundamentados em grafos que possuem estágios de decisão conhecidos por “nós” e nas separações das classes e subconjuntos. A organização dos dados é realizada de forma hierárquica.	Classificação Predição
Algoritmos Genéticos (AGs)	Fornecer modelos genéricos de busca e otimização que são inspirados na Teoria da Evolução. Em cada nova geração, soluções melhores têm mais chances de serem descobertas.	Classificação Agrupamento
Lógica Nebulosa (Lógica Fuzzy)	Fornecer modelos para análises que necessitam de alto grau de abstração, além de, possibilitar análises com graus de incerteza.	Classificação Agrupamento
Análise Estatística	Fornecer modelos para análise e interpretação de dados. O princípio estatístico consiste em fornecer uma maneira pela qual se pode estimar a probabilidade de um evento a partir de dois tipos de conhecimento.	Redes Bayesianas Análise Discriminante Análise Exploratória
Redes Complexas	Fornecer modelos fundamentados em grafos que por sua vez, são constituídos por um conjunto de vértices (nós) que são interligados por arestas (<i>links</i>) sendo que a existência destes, deve ser relacionada a algum tipo de iteração. A análise das medidas do grafo é, fundamental para caracterizar, classificar e modelar as redes. Pode ser representada computacionalmente por meio de uma matriz.	Classificação Predição

Fonte: Elaborada pelo próprio autor, 2022.

2.3 REDES SOCIAIS

O termo, “redes sociais” difere do termo “rede”; ou seja, cada um tem sua própria definição. Considerando que o primeiro é um termo com definição própria e única, e que o segundo pode ser ramificado para rede de computadores, redes de comunicação de dados, rede elétrica, entre outros, é necessário deixar explícito que existe uma separação entre eles. Esta tese refere-se, de forma única, à utilização das “redes sociais”, assim, o termo “rede” e suas ramificações não serão abordados.

Segundo Tomaél e Marteleto (2006), uma rede é uma representação formal de atores e suas relações, e uma rede social refere-se a um conjunto de pessoas, organizações ou outras entidades sociais conectadas por relacionamentos, motivados pela amizade e por relações de trabalho ou compartilhamento de informações e, por meio dessas ligações, vão constituindo e reconstruindo a estrutura social.

Uma rede social é, portanto, em concordância com Silva e Ferreira (2007), um conjunto de pessoas, empresas ou qualquer outra entidade socialmente criada, interligadas por um conjunto de relações que podem ser de amizade, familiares, comerciais, sexuais, entre outros. Dentro dela é possível, ainda, compartilhar crenças, informação, poder, conhecimento, prestígio, entre outros.

Existem, assim, diversas formas de comunicação que podem ser realizadas dentro das redes sociais, pode-se ter, por exemplo, se formar uma comunidade de pessoas com deficiência auditiva e a comunicação ser realizada por meio da linguagem de Libras. As pessoas podem, se comunicar pela fala, por gestos, por expressões corporais ou faciais, entre outras formas.

Para realização desta pesquisa, é escolhida uma rede social digital em que a comunicação ocorre por meio de dispositivos eletrônicos, conectados à Internet e o meio de comunicação é a postagem de mensagens de texto, que podem ser visualizadas pelos demais membros dessa rede social conforme configurações realizadas pelas pessoas que a utilizam.

2.3.1 Redes sociais digitais

Uma das formas de participação em uma rede social é via ferramentas digitais, partindo desse princípio, é possível associar a Internet a um canal de comunicação que interliga pessoas, organizações, associações, empresas ou outras entidades sociais, oferecendo a eles uma forma de comunicação instantânea que pode ser capaz de interligar o mundo todo. A tendência de crescimento desse canal acompanha o desenvolvimento da população.

Segundo Torres (2009), é necessário saber quais são as ferramentas disponíveis na Internet e utilizá-las a nosso favor. Um exemplo desse movimento são as redes sociais digitais que se caracterizam por permitir a criação e o compartilhamento de informações e conteúdos pelas pessoas e para as pessoas, nas quais o consumidor é, ao mesmo tempo, produtor e consumidor da informação.

Para o uso e participação de uma pessoa de forma efetiva nesse processo de comunicação, basta que ela tenha um dispositivo eletrônico com acesso à Internet. São exemplos de dispositivos eletrônicos que podem ser utilizados para essa finalidade: computadores, *notebooks*, *tablets*, aparelhos celulares, televisores, *netbooks*, entre outros.

Ainda nesse sentido, existem milhares de sites e aplicativos de relacionamento que devem ser avaliados dentro do seu contexto, pois cada um deles é direcionado a um determinado público, focado em relações dos mais variados tipos. Portanto, a escolha de qual rede social usar também é importante para interação e comunicação, de acordo com Torres (2009).

2.3.2 Redes Sociais Digitais mais Utilizadas

Existem diversas redes sociais digitais utilizadas por pessoas e empresas pelo mundo todo. São citadas nesta tese as dez mais utilizadas, levando em consideração a quantidade de usuários no Brasil. Essas informações são analisadas e servem para nortear a escolha de qual a utilizada para a realização dos testes.

2.3.2.1 Facebook

O *Facebook* foi criado em 4 de outubro de 2012, por Mark Zuckerberg e por seus amigos da faculdade, Eduardo Saverin, Dustin Moskovitz e Chris Hughes. A criação do site foi inicialmente limitada pelos fundadores aos estudantes da Universidade Harvard, mas foi expandida para outras faculdades na área de Boston, da *Ivy League* e da Universidade *Stanford*.

Atualmente possui dois bilhões de usuários espalhados pelo mundo todo, sendo uma ferramenta de comunicação e informação, promotora de gostos, difusora de consumos e criadora de hábitos de relacionamentos.

Mais de 140 milhões de empresas fazem uso do *Facebook* para se comunicar com os clientes. Por dia, são compartilhadas mais de 100 bilhões de mensagens com o objetivo principal de permitir que as pessoas possam se comunicar mesmo estando distantes.¹

A empresa está sempre testando, resolvendo problemas e trabalhando para conectar pessoas do mundo todo, preocupando-se com a diversidade dos usuários que utilizam os seus serviços todos os dias, tendo a preocupação de tornar a tecnologia acessível a todos e, baseando seu modelo de negócios em publicidade on-line, para que seu acesso possa ser gratuito.

2.3.2.2 WhatsApp

O *WhatsApp* foi criado em 03 de maio de 2009 por Jan Koum e Brian Acton que, juntos, passaram quase 20 anos na empresa *Yahoo*. O *WhatsApp* juntou-se ao *Facebook* em 2014; entretanto, continua operando como um aplicativo independente e com o foco direcionado em construir um serviço de mensagens que seja rápido e que funcione em qualquer lugar do mundo².

Assim, mais de dois bilhões de pessoas, em mais de 180 países, usam o aplicativo para manter contato com amigos e familiares a qualquer hora ou lugar.

¹ As informações nas quais nos baseamos para elaboração das discussões desta seção estão disponíveis no endereço eletrônico: (<https://about.fb.com/br/company-info/>) e constam, também, nas referências, ao final deste trabalho.

² As informações nas quais nos baseamos para elaboração das discussões desta seção estão disponíveis no endereço eletrônico: (<https://www.whatsapp.com/about/>) e constam, também, nas referências, ao final deste trabalho.

Apesar de ser gratuito, ele está sujeito à cobrança de valores pela gestão dos dados e oferece um serviço de mensagens e chamadas simples, seguro e confiável para celulares em todo o mundo. O seu nome é um trocadilho com a frase em inglês "*What's Up*" que significa, em gíria do português, "e aí".

O serviço surgiu como uma alternativa ao sistema de SMS e, pouco tempo depois, passou a possibilitar o envio e recebimento de diversos arquivos de mídia, tais como: fotos, vídeos, documentos e localização, além de textos e chamadas de voz.

Segundo a política de privacidade da empresa, as mensagens e chamadas estão protegidas com a criptografia de ponta a ponta, o que significa que terceiros, incluindo o *WhatsApp*, não podem lê-las e/ou ouvi-las. A ideia principal é possibilitar que as pessoas se comuniquem sem barreiras.

2.3.2.3 Facebook messenger

O *Facebook Messenger* foi criado em 9 de agosto de 2011. Ele é um mensageiro instantâneo e, também é um aplicativo que fornece comunicação por texto e por vídeo. Uma vez que é baseado no recurso de bate-papo *Web*, ele tem como principal função, conectar os seus usuários. Isso é realizado por meio de sincronização dele com a plataforma *Facebook*.

Cada usuário do *Facebook Messenger* é responsável por garantir que qualquer uso dele esteja em conformidade com a legislação aplicável, incluindo as leis sobre manutenção de registros e privacidade de informações educacionais. Muitas pessoas utilizam esse aplicativo para disponibilizar conteúdos de estudo, porém, ele não é uma instituição de ensino

O *Facebook Messenger* pode possibilitar também que, os sujeitos possam se conectar com diversas comunidades de forma digital e em tempo real, mesmo quando não é possível estar junto. Para isso, é disponibilizado aos usuários o recurso de ligações de voz e de vídeo que permitem até oito participantes³.

As empresas podem enviar mensagens que têm por finalidade, ajudá-las a se manter conectadas com os clientes quando a interação pessoalmente não é possível, fazendo uso de ferramentas de mensagens gratuitas e, além disso, sendo capazes de

³ As informações nas quais nos baseamos para elaboração das discussões desta seção estão disponíveis no endereço eletrônico: (<https://www.messenger.com/features>) e constam, também, nas referências, ao final deste trabalho.

fornecer informações importantes aos clientes, ao configurar respostas automáticas para perguntas frequentes, sobre o fechamento de lojas, horário de atendimento, entre outras.

2.3.2.4 YouTube

Segundo Dantas (2020), o *Youtube* foi criado em fevereiro de 2005, por Chad Hurley e Steve Chen, dois funcionários de uma empresa de tecnologia situada em São Francisco, EUA. O site surgiu em virtude do inconveniente que era compartilhar arquivos de vídeo, já que estes eram muito grandes, o que dificultava seu envio por e-mail.

Segundo Dantas (2020), a palavra é uma junção de dois termos da língua inglesa, são eles: “*you*”, que significa “você” e “*tube*”, que é uma gíria para uma palavra que se aproxima de “televisão”, na língua portuguesa. Traduzindo, de uma forma que a junção das palavras tenha sentido, pode-se dizer que o termo significa “televisão feita por você”. Essa é justamente a principal função dessa rede social digital, permitir que os usuários carreguem, assistam e compartilhem vídeos em formato digital.

Segundo Dantas (2020), por meio *Youtube*, os sujeitos podem disponibilizar seus próprios vídeos na Internet e, estes podem ser visualizados por qualquer pessoa no mundo inteiro. A plataforma utiliza o formato *Macromedia Flash* para reproduzir os conteúdos, além de permitir que usuários coloquem os vídeos em seus *blogs* e *sites* pessoais. Todo o seu potencial foi reconhecido pela revista americana “*Time*”, que elegeu o *site* como a melhor invenção de 2006.

Devido ao grande sucesso do *Youtube*, a Google anunciou a compra do *site* em outubro de 2006. A quantia paga foi de US\$1,65 bilhão de dólares. Dessa forma, foi possível unificar os serviços do site da *Google* com o compartilhamento de vídeos. Ainda segundo Dantas (2020), estima-se que diariamente cerca de vinte mil novos vídeos são carregados e trinta milhões são assistidos no *Youtube*.

2.3.2.5 *Instagram*

O *Instagram* foi criado em 06 de outubro de 2010 pelo americano Kevin Systrom e pelo brasileiro Mike Krieger. Poucos meses depois, essa rede social digital se tornou um dos aplicativos mais promissores da *App Store*.

Um ano após sua criação, ele possuía dez milhões de usuários, sendo que o serviço estava disponível apenas para proprietários de *iPhones* e *iPads*. Em 2012, o *Facebook* comprou o *Instagram* por cerca de 1 bilhão de dólares e logo após a compra, o disponibilizou para dispositivos *Android*.⁴

O *Instagram* permite o compartilhamento de fotos e vídeos, bem como a integração com outros aplicativos. Entre as suas funcionalidades estão: a aplicação de filtros, o *Boomerang*, os *Stories*, além das gravações e transmissões de vídeos ao vivo. Atualmente, ele também é um dos principais veículos para a publicidade de empresas de todo o mundo.

A plataforma possui, até o presente momento, cerca de 800 milhões de usuários e, para que seja possível atender toda essa demanda de utilização, por parte dos usuários, a empresa está se baseando em valores como simplicidade, criatividade e privacidade.

No Brasil, o *Instagram* é uma das redes sociais digitais mais acessadas. Desde 2015, os brasileiros ocupam lugar de destaque nas estatísticas do aplicativo. Entre os sujeitos que o utilizam a participação dos brasileiros é maior do que a média global.

2.3.2.6 *LinkedIn*

O *LinkedIn* começou na sala do co-fundador Reid Hoffman em 2002 e foi oficialmente lançado em 5 de maio de 2003. A empresa está sediada no Vale do Silício, nos Estados Unidos, e possui escritórios em todo o mundo. Jeff Weiner é o chefe executivo e a equipe de gestão inclui executivos experientes, vindos de empresas como: *Yahoo!*, *Google*, *Microsoft*, *PayPal* e *Electronic Arts*.

O *LinkedIn* possui um modelo de negócios diversificado, no qual a receita provém de assinaturas, vendas de publicidade e de soluções de recrutamento. Em

⁴ As informações nas quais nos baseamos para elaboração das discussões desta seção estão disponíveis no endereço eletrônico: (<https://canaltech.com.br/empresa/instagram/>) e constam, também nas referências, ao final deste trabalho.

dezembro de 2016, a *Microsoft* adquiriu o *LinkedIn*, unindo, assim, o maior serviço de computação em nuvem profissional do mundo, com a maior rede profissional do mundo⁵

O propósito da rede é conectar os profissionais do mundo para torná-los mais produtivos e bem-sucedidos. Com mais de 675 milhões de membros em todo o mundo, incluindo executivos de todas as empresas da *Fortune 500* (uma classificação das 500 maiores corporações em todo o mundo, medida pela receita das empresas), o *LinkedIn* pode ser considerado como sendo a maior rede social digital profissional do mundo.

2.3.2.7 *Twitter*

O *Twitter* foi lançado em 21 de março de 2006, em São Francisco, Califórnia, Estados Unidos. Os fundadores são: Jack Dorsey, Biz Stone, Noah Glass, Evan Williams. Ele é uma rede social digital e um servidor para *microblogging*, que permite aos usuários enviar e receber atualizações pessoais de outros contatos, por meio de um *website* do serviço, por SMS e por *softwares* específicos de gerenciamento⁶

Dessa maneira, ele acabou se tornando um serviço por meio do qual amigos, familiares e colegas de trabalho podem se comunicar e se manter conectados, trocando mensagens rápidas e frequentes. Os usuários publicam *Tweets*, sendo que, um *Tweet* é qualquer mensagem publicada no *Twitter*. Eles podem conter fotos, vídeos, *links* e texto. Essas mensagens são publicadas no perfil do usuário e enviadas a seus seguidores, podendo ser encontradas por meio dos mecanismos de busca da própria plataforma⁷.

Para que se possa utilizar essa rede social digital é necessário um dispositivo eletrônico que pode ser um celular, um computador ou qualquer outro que possua conexão com a Internet ou que possibilite o envio de mensagens via SMS.

⁵ As informações nas quais nos baseamos para elaboração das discussões desta seção estão disponíveis no endereço eletrônico: (https://about.linkedin.com/pt-br?trk=homepage-basic_directory_aboutUrl) e constam, também nas referências, ao final deste trabalho.

⁶ As informações nas quais nos baseamos para elaboração das discussões desta seção estão disponíveis no endereço eletrônico: (<https://about.twitter.com/pt.html>) e constam, também nas referências, ao final deste trabalho.

⁷ As informações nas quais nos baseamos para elaboração das discussões desta seção estão disponíveis no endereço eletrônico: (<https://help.twitter.com/pt/new-user-faq>) e constam, também nas referências, ao final deste trabalho.

2.3.2.8 *Pinterest*

Pinterest foi lançado em janeiro de 2010, com sede em San Francisco, California, Estados Unidos e seus fundadores foram: Ben Silbermann, Paul Sciarra e Evan Sharp. Ele é uma rede social digital de compartilhamento de fotos. A ideia principal é disponibilizar uma espécie de álbum de inspirações, onde se possa compartilhar e gerenciar imagens temáticas, como por exemplo, imagens de jogos, de hobbies, de roupas, de perfumes, entre outros⁸

Atualmente, o *Pinterest* conta com mais de 200 milhões de usuários em todo o mundo. Além disso, a plataforma é responsável pelo armazenamento de 100 bilhões de ideias. Segundo dados da empresa, cerca de 2 bilhões de pesquisas são feitas mensalmente na rede, sendo 300 milhões destas, pesquisas por imagem.⁹ A principal proposta é fornecer, portanto, uma plataforma digital com acesso via Internet para ajudar os usuários a fazer o que gostam. Para isso, segue diretrizes para nortear o trabalho desenvolvido.

Entre elas estão a segurança, respeito ao direito de propriedade intelectual, privacidade e qualidade. Como proposta de expansão para os negócios, a empresa criou o *Pinterest Labs*, que reúne pesquisadores, cientistas e engenheiros para o desenvolvimento de aprendizado de máquina e inteligência artificial.

2.3.2.9 *Snapchat*

Snapchat foi lançado em 08 de julho de 2011 e os criadores e desenvolvedores foram: Evan Spiegel, Bobby Murphy e Reggie Brown, estudantes da Universidade *Stanford*. Ele é um aplicativo de mensagens com base em imagens¹⁰ que pertence a empresa Snap Inc., uma empresa de câmeras e cuja proposta foi a utilização de uma rede social digital, aprimorando a maneira como as pessoas vivem e se comunicam.

⁸ As informações nas quais nos baseamos para elaboração das discussões desta seção estão disponíveis no endereço eletrônico: (<https://about.pinterest.com/pt-br>) e constam, também nas referências, ao final deste trabalho.

⁹ As informações nas quais nos baseamos para elaboração das discussões desta seção estão disponíveis no endereço eletrônico: (<https://canaltech.com.br/empresa/pinterest/>) e constam, também nas referências, ao final deste trabalho.

¹⁰ As informações nas quais nos baseamos para elaboração das discussões desta seção estão disponíveis no endereço eletrônico: (<https://canaltech.com.br/redes-sociais/guia-do-snapchat-para-iniciantes/>) e constam, também, nas referências, ao final deste trabalho.

A rede é, assim, considerada por seus desenvolvedores como uma forma de contribuir com o progresso da humanidade incentivando as pessoas para se expressarem, viverem os momentos, aprenderem sobre o mundo e se divertirem juntas

Dessa forma, pode-se dizer que O *Snapchat* é um novo tipo de câmera usada por milhões de pessoas todos os dias para manter contato com amigos, se expressar, explorar o mundo, além de tirar algumas fotos, tornando-se um dos aplicativos mais comentados de 2015. Sua proposta diferenciada agradou muitos usuários, principalmente os mais jovens e, conseqüentemente, se tornou muito utilizado entre celebridades de todo o mundo¹¹

No entanto, apesar da crescente popularidade, muitas pessoas ainda não entenderam bem como o *Snapchat* funciona, esse fato está relacionado à proposta, ou seja, oferecer aos usuários, uma rede social digital diferente do Facebook, Twitter, entre outras.

Em resumo, ele é um aplicativo para dispositivos móveis que permite aos usuários enviar vídeos curtos e imagens que são excluídas automaticamente poucos segundos depois de serem visualizadas por seus contatos. Nele, quem quiser criar uma espécie de histórico precisa utilizar um recurso chamado "Minha História", que deixa as imagens e vídeos postados disponíveis para visualização durante 24 horas.

O grande diferencial é que o aplicativo possui um grau de espontaneidade muito maior do que outras redes sociais, como *Instagram*, por exemplo, na qual as pessoas tendem a postar apenas fotos de alta resolução, com muitos filtros, maquiagem, cenários e ângulos perfeitos. No *Snapchat*, as pessoas compartilham diversos momentos aleatórios do seu dia-a-dia, sem muita preparação, o que deixa tudo com formato de diário pessoal, mas que pode ser compartilhado com os mais de 200 milhões de usuários.

A maioria das pessoas que utiliza esse aplicativo são sujeitos famosos e seus fãs, que se divertem acompanhando momentos da intimidade dos seus ídolos, os quais postam diversos vídeos em situações que, vão desde os bastidores de um show, até uma escovação dental matinal, passando por jantares em restaurantes disputados, treinos em academias e viagens para locais paradisíacos.

¹¹ As informações nas quais nos baseamos para elaboração das discussões desta seção estão disponíveis no endereço eletrônico: (<https://www.snap.com/pt-BR/>) e constam, também, nas referências, ao final deste trabalho.

Alguns artistas também optam por compartilhar conteúdo oficial em primeira mão por meio da plataforma. A cantora Madonna, por exemplo, usou o *Snapchat* para lançar o clipe do seu *single* "*Living for Love*". Algumas marcas também já aderiram à rede social digital, como *Burberry*, que apresentou *looks* completos de sua coleção de verão 2016.

2.3.2.10 TikTok

O TikTok é considerado uma rede social ou aplicativo de mídia utilizado para criar e compartilhar vídeos curtos. Além desse nome ela é conhecida também por Douyin e, anteriormente, era chamada de Musical.ly dentro da China, sendo propriedade da companhia de tecnologia chinesa ByteDance.

Ele foi lançado como Douyin na China em setembro de 2016, e introduzido no mercado internacional como musical.ly um ano depois, porém, em novembro de 2017 essa empresa acabou comprando o Musical.ly. Hoje o TikTok é considerado uma plataforma de vídeos curtos, líder de acessos na Ásia, nos Estados Unidos e em outras partes do mundo. O aplicativo ganhou popularidade e se tornou o mais baixado nos Estados Unidos, em outubro de 2018. A partir desse mesmo ano, esteve disponível em mais de 150 mercados e em 75 idiomas. Ele permite que os usuários criem vídeos curtos de até 60 segundos e, em ainda julho de 2018, contava com mais de 500 milhões de usuários em todo o mundo.

2.4 TRANSTORNOS EMOCIONAIS

Segundo Estevam *et al.* (2022), Transtorno Emocional pode ser considerado como uma modificação da saúde mental de uma pessoa, tendo como principal característica gerar uma queda no bem-estar dos indivíduos. O termo transtorno emocional é utilizado para representar uma ou diversas alterações que ocorrem no funcionamento da mente dos sujeitos, afetando o seu humor, raciocínio, comportamentos, atitudes, bem como o desempenho, tanto na vida pessoal, quanto na profissional.

Sendo assim, o transtorno emocional vem sendo tratado com seriedade pela área da saúde, transformando-se em uma das principais causas de incapacitação e afastamento do trabalho no Brasil, nos últimos anos. Os dados que comprovam isso

são da Organização Mundial da Saúde (OMS), que após análise estatística, determinou que só no Brasil em 2019, 12 milhões de pessoas sofreram de depressão e 18,6 milhões tiveram ansiedade. Isso fez com que a preocupação do Governo Federal e dos Governos Estaduais, em oferecer métodos de diagnóstico e de tratamento, aumentasse muito em relação ao que foi investido nos anos anteriores a 2019.

São apresentados três Transtornos Emocionais, sendo esses, os que estão associados ao tema desta tese, são eles: depressão, estresse e ansiedade. Esses três transtornos, bem como seus sintomas, são utilizados para realizar os testes apresentados nesta pesquisa.

2.4.1 Depressão

Segundo Katon (2003), a depressão é uma síndrome psiquiátrica altamente prevalente na população em geral e quando tratada de forma inadequada, com sub doses de medicamentos e manutenção de sintomas residuais pode, ~~vir a~~ comprometer a evolução clínica dos pacientes em outros tratamentos.

De acordo com Duarte e Rego (2007), ela expressa-se em uma ampla variedade de perturbações físicas e funcionais. Sendo assim, é considerada como um dos mais importantes problemas de saúde pública. Os sintomas podem influenciar na forma de pensar, em como lidar com as atividades diárias, além de poder afetar o sono, a alimentação, e o desenvolvimento de atividades no trabalho. Os sintomas mais comuns são: tristeza, ansiedade, sentimentos de desesperança, pessimismo, irritabilidade, fadiga, dificuldade de concentração, perda de apetite, entre outros.

É necessário destacar, portanto, que a maioria dos autores na área dos transtornos depressivos na infância e na adolescência, cita que os sintomas variam com a idade, enfatizando a importância do processo de maturação das diferentes fases do desenvolvimento dos sinais e comportamentos depressivos, existindo uma caracterização sintomatológica predominante por faixa etária, conforme nos explica Versiani et al. (2000).

2.4.2 Estresse

Segundo Houaiss et al. (2001), o termo estresse é utilizado para definir o estado gerado pela percepção de estímulos que provocam excitação emocional e que, ao perturbarem a homeostasia (processo de regulação pelo qual um organismo mantém-se em constante equilíbrio), podem dar início a um processo de adaptação caracterizado, entre outras alterações, pelo aumento da liberação de adrenalina, produzindo diversas manifestações sistêmicas, tais como, distúrbios fisiológicos e psicológicos.

Consoante o pressuposto da OMS – Organização Mundial da Saúde, a resposta ao estresse é resultado da interação entre as características da pessoa e as demandas do meio, ou seja, as discrepâncias entre o meio externo e interno e a percepção do indivíduo quanto a sua capacidade de resposta.

De acordo com a OMS – Organização Mundial da Saúde, a resposta às discrepâncias entre o meio externo e interno compreende aspectos cognitivos, comportamentais e fisiológicos, visando propiciar uma melhor percepção da situação e de suas demandas, assim como um processamento mais rápido da informação disponível, possibilitando busca de soluções, selecionando condutas adequadas e preparando o organismo para agir de maneira rápida e vigorosa. A sobreposição destes três níveis (fisiológico, cognitivo e comportamental) é eficaz até certo limite e, uma vez ultrapassado, poderá se desencadear um efeito desorganizador.

Conforme a OMS – Organização Mundial da Saúde, diferentes situações estressantes ocorrem ao longo dos anos, e as respostas a elas variam entre os indivíduos na sua forma de apresentação, podendo ocasionar manifestações psicopatológicas diversas como sintomas inespecíficos de depressão ou ansiedade, ou transtornos psiquiátricos definidos, como por exemplo, o Transtorno de Estresse Pós-Traumático (TEPT).

2.4.3 Ansiedade

Segundo Allen et al. (1995), a ansiedade pode ser considerada como um sentimento vago e desagradável, causando de medo e apreensão e caracterizando-se como tensão ou desconforto derivado de antecipação de perigo, de algo desconhecido ou estranho.

Além disso, de acordo com Hirshfeld et al. (1999), ela passa a ser reconhecida como patológica quando é exagerada e desproporcional em relação ao estímulo, ou qualitativamente diversa do que se observa como norma naquela faixa etária, interferindo na qualidade de vida, no conforto emocional ou o desempenho diário do indivíduo.

Essas reações exageradas ao estímulo ansiogênico se desenvolvem, mais comumente, em indivíduos com uma predisposição neurobiológica herdada. De acordo Rosen e Schilkin (1998), a maneira prática de se diferenciar ansiedade normal de ansiedade patológica é basicamente avaliar se a reação é de curta duração, autolimitada e relacionada ao estímulo do momento ou não. Os transtornos ansiosos são quadros clínicos em que os sintomas da Ansiedade são primários, ou seja, não são derivados de outras condições psiquiátricas como: depressões, psicoses, transtornos do desenvolvimento, transtorno hiperativo, entre outros.

Em concordância com Bernstein et al. (1996), pode-se dizer que, transtornos ansiosos podem também, surgir de outras síndromes sendo frequentes em outros transtornos psiquiátricos. Pode haver ainda, casos em que vários transtornos estão presentes ao mesmo tempo e não se consegue identificar o que é primário e o que não é, sendo mais correto determinar que o paciente apresenta mais de um diagnóstico coexistente (comorbidade).

Segundo as vertentes de Bernstein et al. (1996), esses transtornos ansiosos são os quadros psiquiátricos mais comuns tanto em crianças quanto em adultos, com uma prevalência estimada durante o período de vida de 9% e 15% respectivamente. A distribuição entre os sexos é de modo geral equivalente, exceto fobias específicas, transtorno de estresse pós-traumático e transtorno de pânico com predominância do sexo feminino. De uma maneira geral, os transtornos ansiosos na infância e na adolescência apresentam um curso crônico, embora flutuante ou episódico, se não tratados, Segundo nos diz Sylvester (2000).

2.5 FERRAMENTAS DE MINERAÇÃO DE DADOS

Devido à intensificação na realização de pesquisas em mineração de dados, diversos algoritmos e ferramentas têm surgido para essa finalidade. Com isso, passou a existir uma grande oferta de ferramentas para isso, em ambiente livre e gratuito, com código fonte aberto. Entre elas destacam-se: *Weka*, *Mahout*, *Orange Data Mining*, *Rapid Miner*, *Tanagra*, *Keel*, *R Studio*, entre outras.

Em contrapartida existem as aplicações proprietárias, que por sua vez, possuem várias soluções e algoritmos de mineração incorporados em ferramentas de *business intelligence*, que são desenvolvidas por diversas empresas com foco em soluções que utilizam a Tecnologia da Informação, como, por exemplo: Oracle, Microsoft, SAS, entre outras.

Mesmo diante de diversos caminhos a serem seguidos, a escolha pela ferramenta mais adequada e que pode gerar resultados consistentes, deve ser fundamentada na identificação de qual estratégia algorítmica melhor se adequa ao contexto, problema e questão.

Para a realização dos testes, foi firmada parceria entre Unesp de Bauru e a Empresa SAS. Essa parceria estabelece que a Empresa SAS deve fornecer as duas ferramentas de Mineração de Dados a serem utilizadas para o processo de mineração, são elas: *SAS Enterprise Miner* e *SAS Visual Text Analytics*. A oferta dessas ferramentas vai ao encontro da análise e da proposta apresentada por essa tese junto à diretoria da empresa SAS, que demonstrou interesse nesse tipo de pesquisa, solicitando em troca do direito de uso gratuito de duas soluções SAS, que ela possa ter acesso aos resultados obtidos para os testes realizados.

2.5.1 SAS

A sigla SAS significava “Sistema de Análise Estatística” e teve início na Universidade Estadual da Carolina do Norte como um projeto para analisar uma pesquisa agrícola. Como a demanda por esse tipo de *software* cresceu, uma empresa com o nome do próprio sistema foi fundada em 1976, para ajudar a todos os tipos de

clientes, desde empresas farmacêuticas e bancos, até entidades acadêmicas e governamentais¹²

O SAS é um sistema integrado de aplicações para a análise de dados, que consiste em: recuperar dados, gerenciar arquivos, realizar análises estatísticas, acessar Bancos de Dados, gerar gráficos, gerar de relatórios, entre outras funcionalidades. De forma resumida o sistema trabalha com quatro ações básicas sobre: acessar, manipular, analisar e apresentar dados.

Os *Softwares* desenvolvidos pela empresa, em sua grande maioria, oferecem suporte para diversos sistemas operacionais disponíveis no mercado, sendo os programas e arquivos, portáteis para qualquer um desses ambientes. De forma geral, a portabilidade e disponibilidade é compatível com qualquer tipo de Sistema Operacional, mesmo que por meio de virtualização.

Atualmente, eles são utilizados em mais de 70 mil locais de 136 países, incluindo 93 das 100 maiores empresas da lista 2013 *Fortune Global 500® list* e, no Brasil, isso não é diferente, onde grandes empresas de diversos segmentos desde financeiras a órgãos governamentais, contam com as soluções dessa empresa para terem maiores e melhores informações sobre o seu negócio.

Cada vez mais profissionais das áreas de *marketing*, finanças e tecnologia, têm se aprofundado no conhecimento desta poderosa ferramenta. Muitas universidades brasileiras conceituadas, usam o SAS na grade curricular, para deixar os alunos ainda mais preparados para o mercado. Cursos e treinamentos de capacitação são oferecidos no mercado para que o profissional possa sempre se reciclar. Dessa forma, os *Softwares* desenvolvidos pela instituição vêm ganhando cada vez mais espaço no mercado.

2.5.2 SAS Enterprise Miner

Uma das ferramentas escolhidas para a realização dos testes é a SAS *Enterprise Miner*, cuja escolha vai ao encontro da proposta de pesquisa apresentada. Após diversas reuniões entre a Unesp de Bauru e a empresa, em que foi discutida,

¹² As informações nas quais nos baseamos para elaboração das discussões desta seção estão disponíveis no endereço eletrônico: (https://www.sas.com/pt_br/company-information-old-before-redesign/profile) e constam, também, nas referências, ao final deste trabalho.

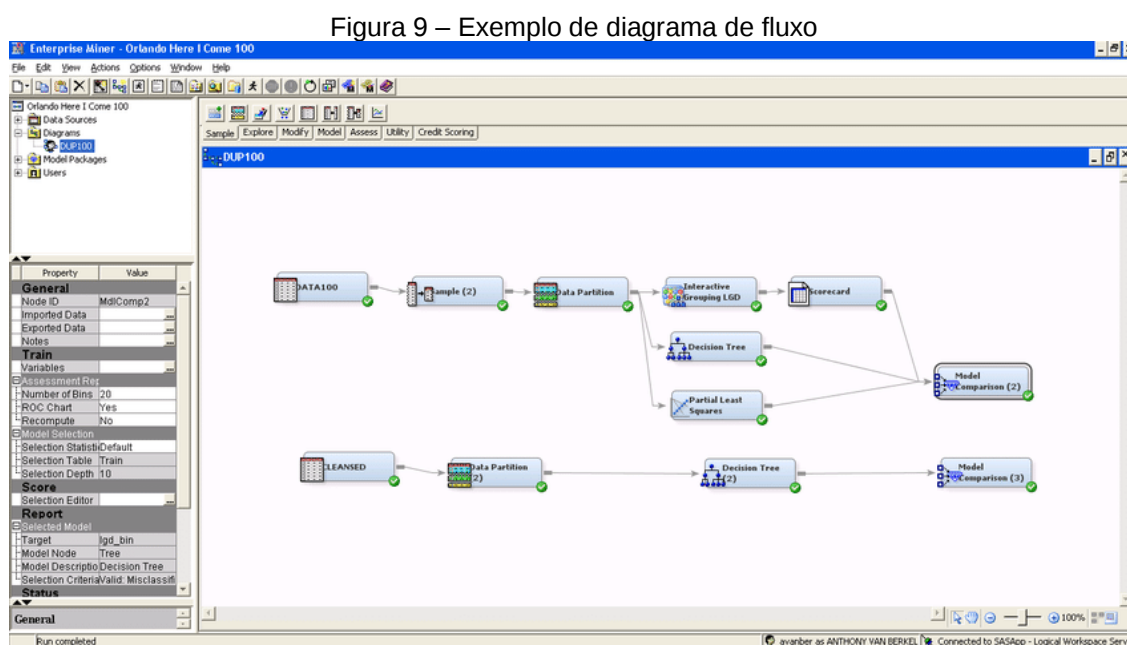
por diversas vezes, essa proposta de pesquisa, os profissionais técnicos decidiram por indicar essa ferramenta para a realização dos testes.

O *SAS Enterprise Miner*, de forma resumida, é indicado quando se deseja criar modelos analíticos precisos, com alto nível de confiança nos resultados obtidos. Ele é muito utilizado quando se deseja minerar, realizar comparações e tabular resultados em grandes volumes de dados.

Ele simplifica o processo de mineração para que seja possível criar modelos analíticos preditivos e descritivos precisos em grandes volumes de Dados e pode, ainda, ser utilizado para detectar fraudes, minimizar riscos, antecipar demandas de recursos, reduzir o tempo de inatividade de ativos, aumentar taxas de resposta para campanhas de *marketing*, entre outras possibilidades¹³.

Ele oferece recursos avançados de análise preditiva e mineração de dados que possibilitam a análise de dados complexos, encontrar informações úteis e agir com confiança para a tomada de decisões baseadas em fatos.

Por meio da figura 9 é possível observar um diagrama de fluxo para um modelo de mineração de dados:



Fonte: https://www.sas.com/content/dam/SAS/en_us/doc/factsheet/sas-enterprise-miner-101369.pdf

¹³ As informações nas quais nos baseamos para elaboração das discussões desta seção estão disponíveis no endereço eletrônico: (https://www.sas.com/content/dam/SAS/en_us/doc/factsheet/sas-enterprise-miner-101369) e constam, também, nas referências, ao final deste trabalho.

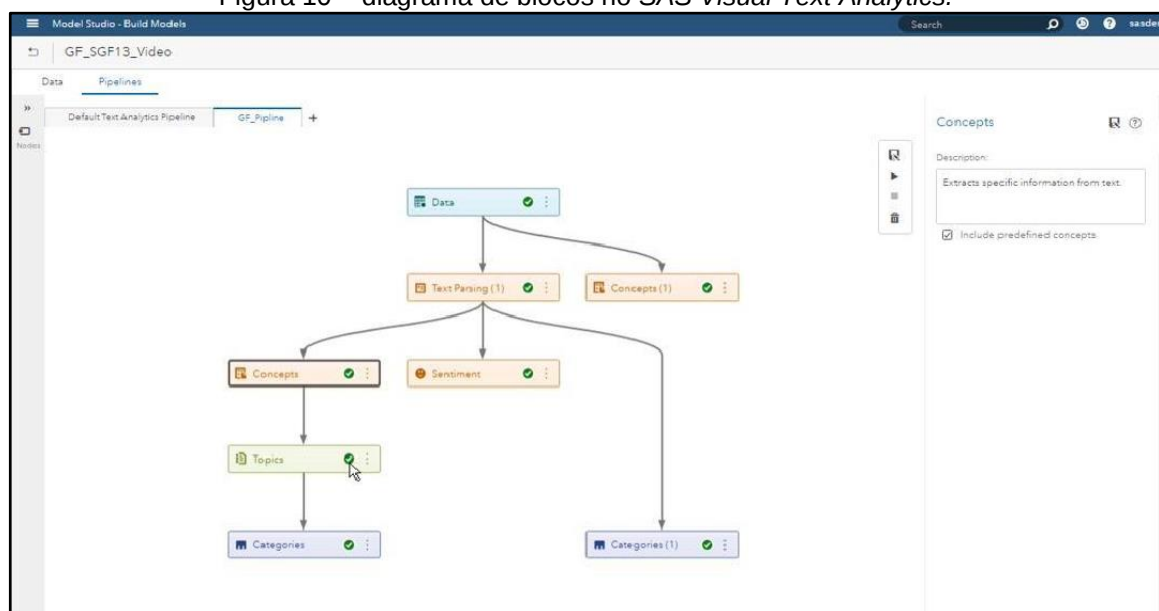
2.5.3 SAS Visual Text Analytics

O SAS *Visual Text Analytics* é outra ferramenta indicada pela empresa para a realização do processo de mineração dos dados. A ideia de utilizar duas ferramentas distintas funciona na mineração dos dados por meio de análise estatística, sendo esta, realizada utilizando o *SAS Enterprise Miner* e pela aplicação do conceito de redes complexas, que nesse caso, é utilizada através do *SAS Visual Text Analytics*.

Esta é uma ferramenta de análise de texto baseada na Web que utiliza um contexto para fornecer soluções abrangentes para o desafio de identificar e categorizar dados textuais importantes. Por meio da dela é possível criar diagramas de bloco, que tem por finalidade representar modelos estatísticos de mineração de dados com base em documentos.

Ela combina, ainda, o fluxo de programação visual do *SAS Enterprise Miner* com os métodos linguísticos baseados em regras de categorização e extração de conceitos. Dessa forma, é possível identificar os principais dados textuais em coleções de documentos, por meio da aplicação de conceitos e modelos de categorização, podendo ainda, remover dados textuais sem sentido. Por meio da figura 10 é possível visualizar a interface do *SAS Visual Text Analytics* com um modelo de mineração de dados representado por um diagrama de blocos:

Figura 10 – diagrama de blocos no SAS Visual Text Analytics.



Fonte: <https://video.sas.com/detail/video/5812813159001/creating-concept-rules-using-textual-elements-in-sas%C2%AE-visual-text-analytics>

2.6 Lei Geral de Proteção aos Dados Pessoais - Lei nº 13.709/2018

Por se tratar da análise de informações disponibilizadas por meio de postagens realizadas em uma rede social, via mídia digital, existe a preocupação com a privacidade das pessoas e que seus direitos sejam preservados e respeitados.

Dessa forma, as informações coletadas são obtidas por meio de processo de extração em uma base de dados proveniente de postagens de perfil público, na rede social denominada *Twitter*.

Considerando todas as redes sociais digitais apresentadas nesse capítulo, foi realizada tentativa de extração de dados no *Facebook* e *Instagram*. Devido ao um vazamento na base de dados do *Facebook* em 2021 e as alterações realizadas na Lei Geral de Proteção dos Dados Pessoais não foi possível obter os dados devido ao bloqueio do acesso às bases de dados nessas redes sociais digitais, mesmo em perfis públicos.

O mesmo aconteceu com diversas redes sociais digitais. Já no caso do *Twitter* é possível realizar a coleta dos dados por meio da utilização da ferramenta de extração de dados conhecida por *Facepager*.

Além disso, é preciso se dizer, ainda, que algumas redes sociais foram descartadas após estudo e análise realizados em função do que é postado nelas em comparação ao que se pretende obter e analisar no âmbito dessa tese.

3. METODOLOGIA

Nesse capítulo é descrito todo o processo utilizado para validar a proposta dessa tese, que consiste em verificar se a mineração de dados pode contribuir para, determinar se as pessoas que frequentam a rede social digital *Twitter* estão acometidas por alguma síndrome socioemocional, em específico depressão, estresse e ansiedade. Todo o procedimento para que se possa atingir esse objetivo está descrito na ordem em que cada etapa é realizada.

Inicialmente, faz-se necessário definir as duas ferramentas utilizadas para a mineração dos dados e, em seguida, é possível definir as demais etapas que são: a escolha da rede social digital utilizada para obter os dados, a definição do processo de coleta, o tratamento dos dados obtidos, a mineração destes, os resultados gerados para cada processo de Mineração realizado e, por fim, a análise dos resultados apresentados no quarto capítulo.

3.1 Ferramentas Utilizadas para a Mineração dos Dados

Diante disso, é definido que as ferramentas utilizadas para a Mineração de Dados são o *SAS Enterprise Miner* e o *SAS Visual Text Analytics*. Essa escolha está vinculada a uma parceria estabelecida entre a Unesp de Bauru, via LTia (Laboratório de Tecnologia da Informação Aplicada) e a Empresa SAS "*Statistical Analysis System*".

A SAS é uma empresa pioneira em *Business Intelligence* e de uma família de softwares gerenciadores de bancos de dados comercializados por ela, a qual se chamava *SAS Institute* até 2001, tendo seu nome alterado para SAS, somente.

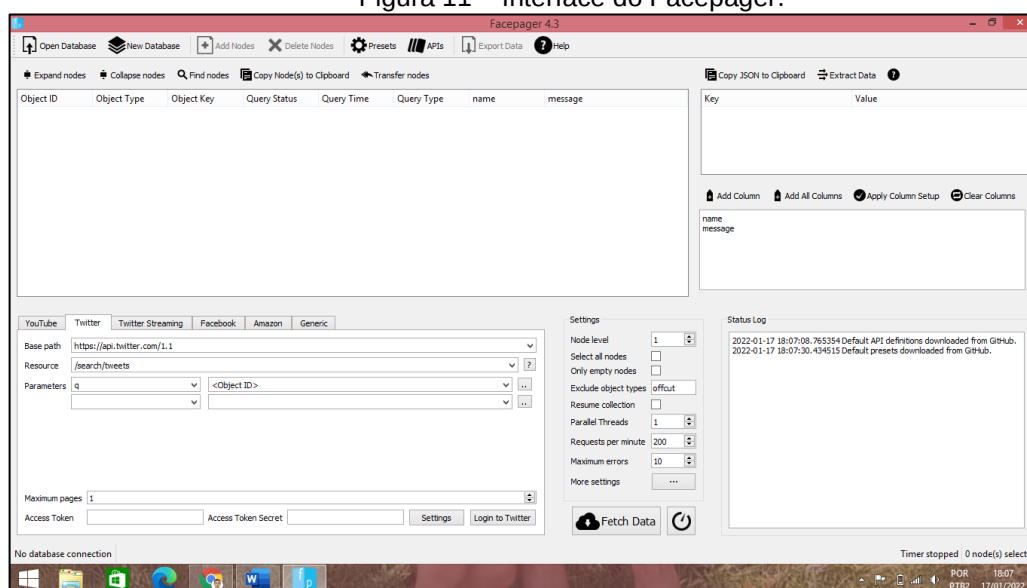
3.2 Rede Social Utilizada para Coleta dos Dados

A escolha pela rede social *Twitter* fundamenta-se na Lei Geral de Proteção dos Dados (LGPD). Após a LGPD entrar em vigor e, sofrer algumas alterações, que começaram a ser postas em prática em 14 de agosto de 2020, diversas outras redes sociais digitais deixaram de permitir o acesso aos dados, ainda que estes sejam de perfil público. Isso se tornou uma dificuldade com relação à obtenção de uma base de

dados de postagens realizadas dentro de uma rede social digital, a partir das quais essa análise está respaldada, através de parâmetros legais.

No caso do *Twitter*, no entanto, existe uma API conhecida pelo nome de *Facepager* que conforme as diretrizes da LGPD, permite o *download* das postagens de perfil público. Para conseguir fazer o *download* das postagens é necessário que se tenha uma conta de usuário na plataforma e toda vez que um acesso a ela é realizado via *Facepager*, fica registrado o usuário que solicitou acesso ao *Download* das postagens. Por meio da figura 11 é possível visualizar a interface principal do *Facepager*:

Figura 11 – Interface do Facepager.



Fonte: Elaborada pelo próprio autor, 2022.

O maior problema em acessar as outras redes sociais digitais que foram pesquisadas é que, ou não é permitido o acesso aos dados delas, ou não é permitido utilizar como parâmetros de busca uma ou mais de uma das palavras-chave como: depressão, estresse e ansiedade.

3.3 Coleta dos Dados

Com o objetivo de verificar uma periodicidade nas publicações, a base de dados foi gerada a partir de *downloads* realizados em datas distintas e de escolha aleatória. A cada processo de *download*, o *Facepager* gera uma base de dados das postagens feitas durante as últimas vinte e quatro horas a contar da data e hora da

solicitação de busca. As coletas de Dados foram realizadas nos dias 25 de novembro de 2021, 02 de dezembro de 2021, 05 de dezembro de 2021 e 08 de dezembro de 2021 sempre por volta das 17 horas.

No total, considerando as quatro datas, foi gerada uma base com 109.093 registros. Os registros gerados são organizados de forma automática pelo *Facepager* em um arquivo de formato CSV com leitura realizada via *Microsoft Excel*. Por meio da figura 12, é possível visualizar o cabeçalho da Planilha do Excel com os títulos de algumas das colunas e alguns de seus respectivos dados. Os nomes dos usuários foram suprimidos para evitar uma possível identificação.

Figura 12 – Base de dados em formato CSV

Data_pesquisa	Data_hora_postagem	texto	sigla_idioma	origem_postagem	nome_usuario	apelido_usuario	localizacao_usuario
2021-11-25 13:37:13.3	Thu Nov 25 16:35:04 +0000 2021	RT @caiozits: Venom 2 é uma farofada tão pt		<a href="http://twitter.com/ci		hermesnkiwis	ela/dela/she/her
2021-11-25 13:37:13.3	Thu Nov 25 16:35:04 +0000 2021	Que calor que sono que depressão que pr pt		<a href="http://twitter.com/ci		intaenataluke	
2021-11-25 13:37:13.3	Thu Nov 25 16:35:02 +0000 2021	@shysmimo você acabou de me deixar co pt		<a href="http://twitter.com/ci		sourtsw	Baby, i'm dancing in the e
2021-11-25 13:37:13.3	Thu Nov 25 16:34:59 +0000 2021	q depressão eh essa pt		<a href="http://twitter.com/ci		wfreitaz7	Manaus, Brasil
2021-11-25 13:37:13.3	Thu Nov 25 16:34:58 +0000 2021	RT @lavi_rocha: Cada dia mais vendo meu pt		<a href="http://twitter.com/ci		savethekittycat	Fire Nation
2021-11-25 13:37:13.3	Thu Nov 25 16:34:57 +0000 2021	ainda com um pouco de depressao mas co pt		<a href="http://twitter.com/ci		gihmctr	Brasília
2021-11-25 13:37:13.3	Thu Nov 25 16:34:55 +0000 2021	SAIU O TRAILER DE MSHD NUNCA TIVE DEF pt		<a href="http://twitter.com/ci		angstyangore	
2021-11-25 13:37:13.3	Thu Nov 25 16:34:52 +0000 2021	RT @amauariitigji: depois que alguém falc pt		<a href="https://mobile.twittl		pedroquemia	santa cruzes, rio de janoc
2021-11-25 13:37:13.3	Thu Nov 25 16:34:52 +0000 2021	apaga isso tem gente q passa mal tem gen pt		<a href="http://twitter.com/ci		heybribz	wfp
2021-11-25 13:37:13.3	Thu Nov 25 16:34:48 +0000 2021	Quando eu tô triste e só colocar mtg no ta pt		<a href="http://twitter.com/ci		zaidan0601	Betim, Brasil
2021-11-25 13:37:13.3	Thu Nov 25 16:34:43 +0000 2021	RT @frogtales : To com depressao fiz a pr pt		<a href="http://twitter.com/ci		smtvpostagem	
2021-11-25 13:37:13.3	Thu Nov 25 16:34:39 +0000 2021	@erenwebsite @ITADORICAT manooooo pt		<a href="http://twitter.com/ci		chuuyapeitos	조조 (ele/dele) ; que les
2021-11-25 13:37:13.3	Thu Nov 25 16:34:38 +0000 2021	Depressão :) https://t.co/WILCT2jKAN pt		<a href="http://twitter.com/ci		ElisuaAmiguinha	Porto Seguro, Brasil
2021-11-25 13:37:13.3	Thu Nov 25 16:34:34 +0000 2021	zhongli e xiao back-to-back e vamos de de pt		<a href="http://twitter.com/ci		peekabyul	marcela razão do meu vi
2021-11-25 13:37:13.3	Thu Nov 25 16:34:34 +0000 2021	comi um ovo hj de manhã, to até agora sei pt		<a href="http://twitter.com/ci		monicafelii	
2021-11-25 13:37:13.3	Thu Nov 25 16:34:33 +0000 2021	RT @wooyospace: — fim da thread. espero pt		<a href="http://twitter.com/ci		/w00s4ns	모모 (s.her) ; que les a

Fonte: Elaborada pelo próprio autor, 2022.

3.4 Tratamento dos Dados

Para realizar a mineração dos dados utilizando o *SAS Enterprise Miner* e o *SAS Visual Text Analytics* é necessário que a base de dados, esteja em um formato específico e compatível com eles. Para realizar a conversão da base de dados do formato CSV para o formato *SAS Data Set*, utiliza-se outra ferramenta SAS, nesse caso o *SAS Studio*.

Essa ferramenta possibilita via programação em linguagem própria dos sistemas SAS, converter os dados que estavam no formato CSV para o formato *SAS Data Set*. A integridade deles precisou ser verificada e validada. Por meio da figura 13, é possível visualizar o cabeçalho da planilha que representa a base de dados gerada pelo *SAS Studio* com os títulos de algumas das colunas e alguns de seus respectivos Dados:

Figura 13 – Amostra da base de dados no formato SAS Data Set

Table:	TESTE_1.BASE_DADOS_V1	View:	Column names	     Filter: (none)
Total rows: 109093 Total columns: 13				  Rows 1-100  
	Data_pesquisa	Data_hora_postagem	texto	
1	2021-11-25 13:37:13.335526	Thu Nov 25 16:35:04 +0000 2021	RT @caiozits: Venom 2 é uma farofada tão go	
2	2021-11-25 13:37:13.335526	Thu Nov 25 16:35:04 +0000 2021	Que calor que sono que depressão que preç	
3	2021-11-25 13:37:13.335526	Thu Nov 25 16:35:02 +0000 2021	@shysmimo você acabou de me deixar com	
4	2021-11-25 13:37:13.335526	Thu Nov 25 16:34:59 +0000 2021	q depressão eh essa	
5	2021-11-25 13:37:13.335526	Thu Nov 25 16:34:58 +0000 2021	RT @lavi_rocha: Cada dia mais vendo meus a	
6	2021-11-25 13:37:13.335526	Thu Nov 25 16:34:57 +0000 2021	ainda com um pouco de depressao mas corr	

Fonte: Elaborada pelo autor, 2022.

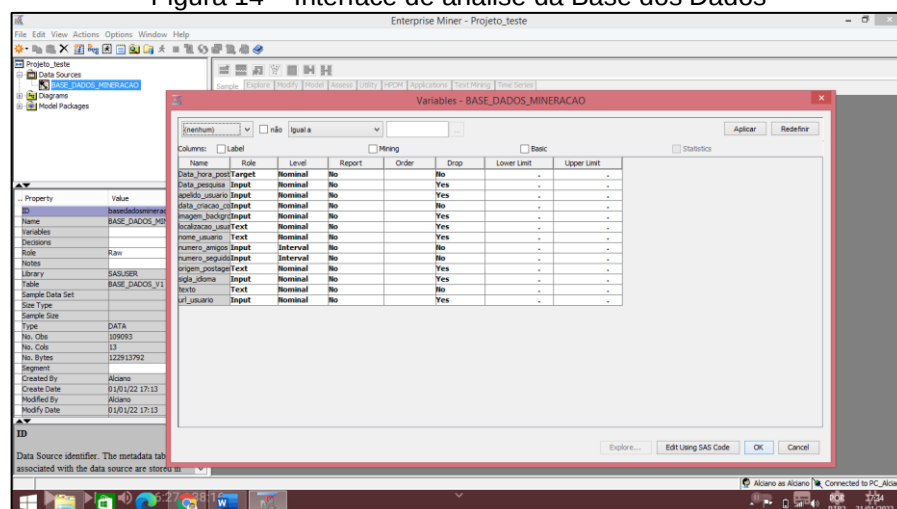
3.5 Pré-Processamento

O processo de mineração tem estágios intermediários, nos quais é necessário realizar o pré-processamento dos dados antes do processo de mineração dos dados em si.

Nessa primeira etapa é possível ter uma análise dos dados e/ou prepará-los para serem minerados. Considerando-se que isso deve ser realizado dentro do SAS Enterprise Miner e dentro do SAS Visual Text Analytics, além de que o pré-processamento é idêntico, em ambas as ferramentas, assim, essa etapa é descrita de forma única e equivale a ambas.

A primeira etapa, então, consiste em selecionar a base de dados a ser minerada. Por meio da figura 14 é possível visualizar a base de dados, os atributos e algumas opções de configuração de acesso e utilização dos valores armazenados:

Figura 14 – Interface de análise da Base dos Dados



Fonte: Elaborada pelo autor, 2022.

Por meio dessa interface, é possível determinar qual é a variável (coluna da tabela) que é utilizada como parâmetro para a resposta da mineração, nesse caso ficou definido que a variável “Texto” é a referência para essa resposta, informando quantas vezes as palavras estresse, depressão e ansiedade estão sendo utilizadas nas postagens, em quais mensagens e em quantas vezes elas aparecem associadas a outras palavras-chaves, selecionadas para a mineração, e associando-as, ainda, aos usuários.

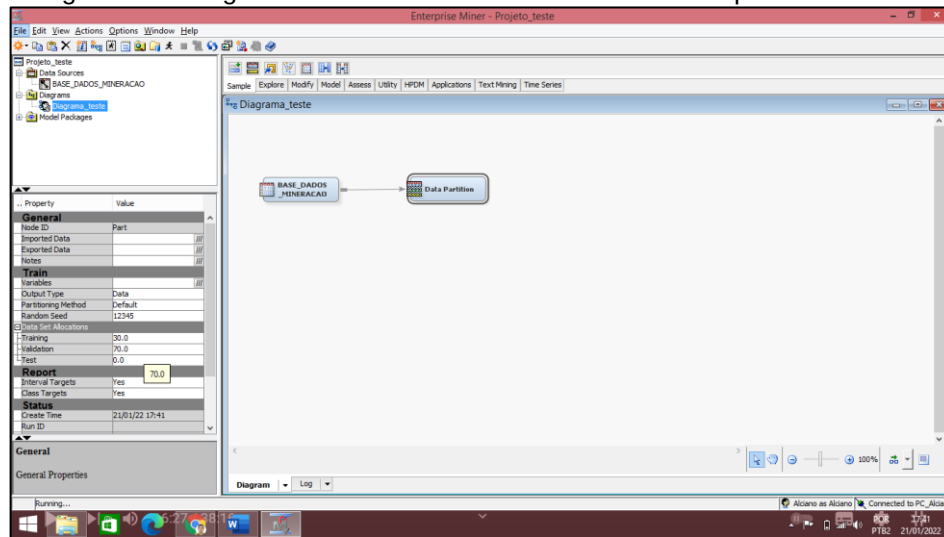
É possível também, excluir variáveis (colunas da tabela) que, por definição do cientista de dados, não são utilizadas para essa mineração em específico. Para isso, na coluna *Drop*, basta alterar a palavra “No” para “Yês”, excluindo, assim, a coluna desse processo de mineração em específico. A próxima etapa consiste em montar o estágio inicial do diagrama de blocos que deve compor o processo de mineração dos dados.

Esse estágio inicial é responsável por possibilitar o pré-processamento. O primeiro bloco do diagrama deve ser a base de dados. Ao inserir o bloco que corresponde à base de dados no início do diagrama, é necessário inserir em seguida o bloco *Data Partition*, que possibilita particionar a base dividindo-a em duas partes, uma para o treinamento e outra para a Mineração.

Dessa forma, a base é particionada em setenta por cento dos Dados para Mineração e trinta por cento para treinamento do Sistema. Por meio da figura 15 é possível visualizar a interface de criação do diagrama de blocos, a base de dados inserida e o bloco *Data Partition* utilizado para o particionamento dela.

Nas opções de configuração à direita na imagem é possível visualizar o percentual de divisão da base:

Figura 15 – Diagrama de blocos com a base de dados e o particionamento



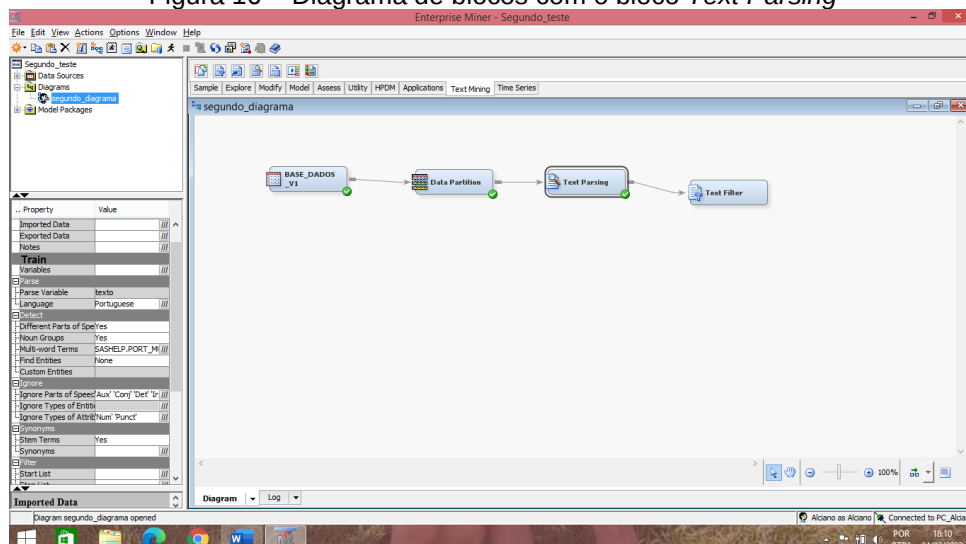
Fonte: Elaborado pelo autor, 2022.

Após essa etapa é necessário determinar a forma utilizada para filtrar as mensagens de texto, para isso, é adicionado um bloco denominado “*Text Parsing*”. Ele é um bloco que possui técnicas avançadas, que permitem dividir documentos em termos como: palavras, frases, sinais, pontuação, entre outros; além disso, possibilita que isso seja realizado em vários idiomas.

É possível, também, extrair endereços, números de telefone, nomes, entre outros. Nesse caso, ele é utilizado para que ao analisar todas as postagens, são identificadas as palavras estresse, depressão e ansiedade.

Por meio da figura 16 é possível visualizar o nó *Text Parsing* inserido no diagrama de blocos:

Figura 16 – Diagrama de blocos com o bloco *Text Parsing*



Fonte: Elaborado pelo autor, 2022.

Ao término dessa etapa o processo de pré-processamento está definido. Para a etapa seguinte, é possível, portanto, minerar os dados, adicionando mais um bloco ao diagrama.

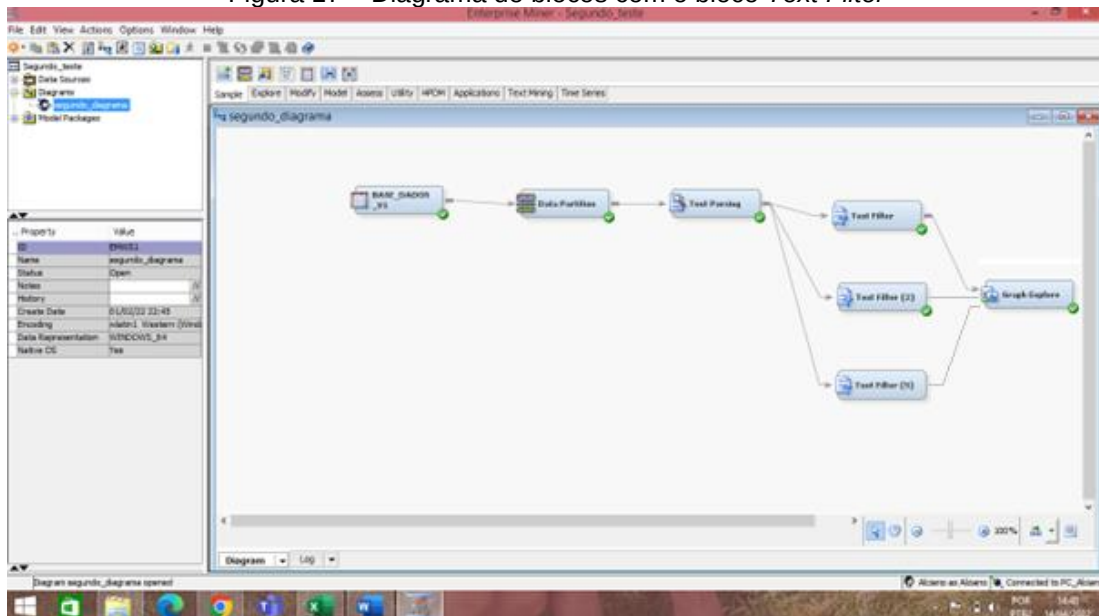
3.6 Mineração dos dados

O bloco responsável pela Mineração dos Dados no *SAS Enterprise Miner* é o *Text Filter*. Ele possibilita reduzir o número total de termos, selecionando somente o que foi pré-definido, atuando da seguinte forma: palavras comuns ou infrequentes podem não ser úteis para análise e podem ser filtradas. É possível explorar uma coleção de documentos analisados no bloco *Text Parsing* e obter informações que possibilitem agrupar coleções de documentos, com base nos atributos de um documento ou no conteúdo do documento. Após sua execução é possível obter o resultado da mineração.

O ajuste nos parâmetros desse bloco permite que sejam considerados e selecionados, sinônimos ou outros aspectos que são relevantes para o processo. Conforme orientações da equipe de suporte da Empresa SAS, considerando a base de dados e o que se pretende minerar, é indicado utilizar a configuração padrão para cada bloco utilizado no processo. Ao total, utiliza-se três, um para filtrar pela palavra “estresse”, um para filtrar pela palavra “ansiedade” e outro para filtrar pela palavra “depressão”.

Por meio da figura 17 é possível observar o diagrama de blocos utilizado para o processo de mineração como um todo no *SAS Enterprise Miner*. Incluindo a inserção da base de dados, pré-processamento e mineração:

Figura 17 – Diagrama de blocos com o bloco *Text Filter*



Fonte: Elaborado pelo autor, 2022.

O diagrama é executado bloco a bloco e os resultados são apresentados e discutidos no capítulo 4. O que difere o processo de mineração realizado no SAS *Enterprise Miner* para o SAS *Visual Text Analytics* é que na primeira ferramenta citada, o procedimento é realizado por meio de análise estatística, já na segunda, através da aplicação dos conceitos de redes complexas.

Comparando o diagrama de blocos do SAS *Enterprise Miner* com o SAS *Visual Text Analytics*, é possível perceber que o bloco “Base de Dados” e o “Data partition” são equivalentes ao bloco “Dados”. O “Text Parsing” é o mesmo em ambos os diagramas, já o “Text Filter” equivale aos blocos “Sentiments” e “Topics”. É importante ressaltar ainda, que nessa última etapa, ocorre o processo de mineração em ambas as ferramentas, porém, de forma diferente.

O “Text Filter” na ferramenta SAS *Enterprise Miner* realiza a mineração dos dados por meio de análise estatística. Em “Sentiments” e “Topics” no SAS *Visual Text Analytics* é realizada a mineração de dados, utilizando como referência o conceito de redes complexas.

Por meio da figura 18 é possível visualizar o diagrama de blocos utilizado no SAS *text Analytics*:

Figura 18 – Diagrama de blocos utilizado no SAS Visual Text Analytics



Fonte: Elaborada pelo autor, 2023.

Conforme dito anteriormente, as ferramentas disponibilizadas para os testes geram, como resultado dos processos de mineração, pequenas bases de dados que são lidas somente pelas ferramentas SAS. Para gerar os resultados graficamente, é necessário converter os resultados obtidos em planilhas CSV e, a partir delas, são gerados os gráficos.

No capítulo 4, são apresentados e discutidos os resultados obtidos em ambas as ferramentas, com seus respectivos processos de mineração. Os parâmetros de busca utilizados são os mesmos para as duas ferramentas.

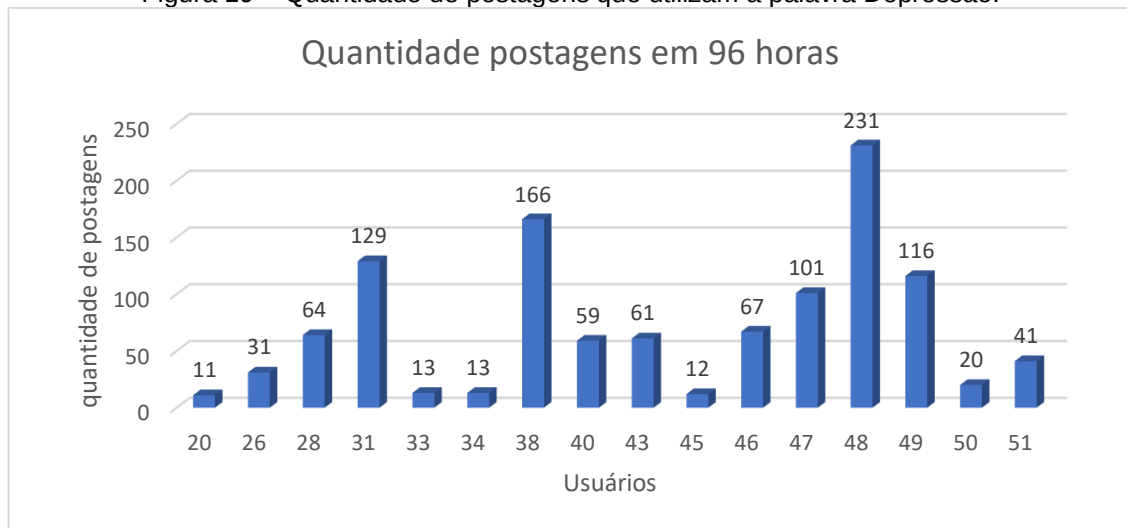
4. RESULTADOS OBTIDOS

Nesse capítulo são apresentados os resultados obtidos para cada processo de mineração, de forma individual. Antes de mais nada, é importante ressaltar que os dados foram coletados em quatro dias diferentes e, que a cada coleta, foram obtidos os registros de postagens de perfil público, utilizando o idioma português para as últimas vinte e quatro horas, totalizando 96 horas de coleta de dados.

4.1 Resultados com a ferramenta *SAS Enterprise Miner*

Para uma primeira análise, é realizada busca por quantidade de postagens por usuário que contêm a palavra “depressão”. Por meio da figura 19 é possível observar que, para um total de 96 horas, alguns usuários têm uma quantidade de postagens que utiliza a palavra depressão de forma bem acentuada. Outros, com quantidades menores, são desconsiderados.

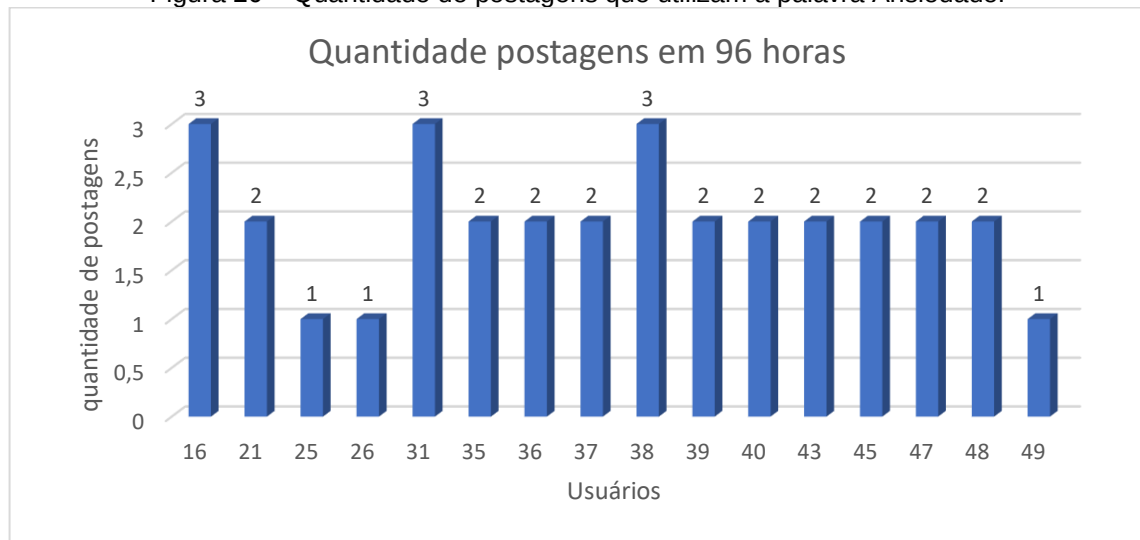
Figura 19 – Quantidade de postagens que utilizam a palavra Depressão.



Fonte: Elaborado pelo autor, 2022.

Para uma segunda análise, realiza-se uma busca por quantidade de postagens por usuário que contêm a palavra “ansiedade”. Por meio da figura 20 é possível observar, para um total de 96 horas, as postagens que utilizam a palavra mencionada. Os demais usuários não utilizaram essa palavra em suas postagens.

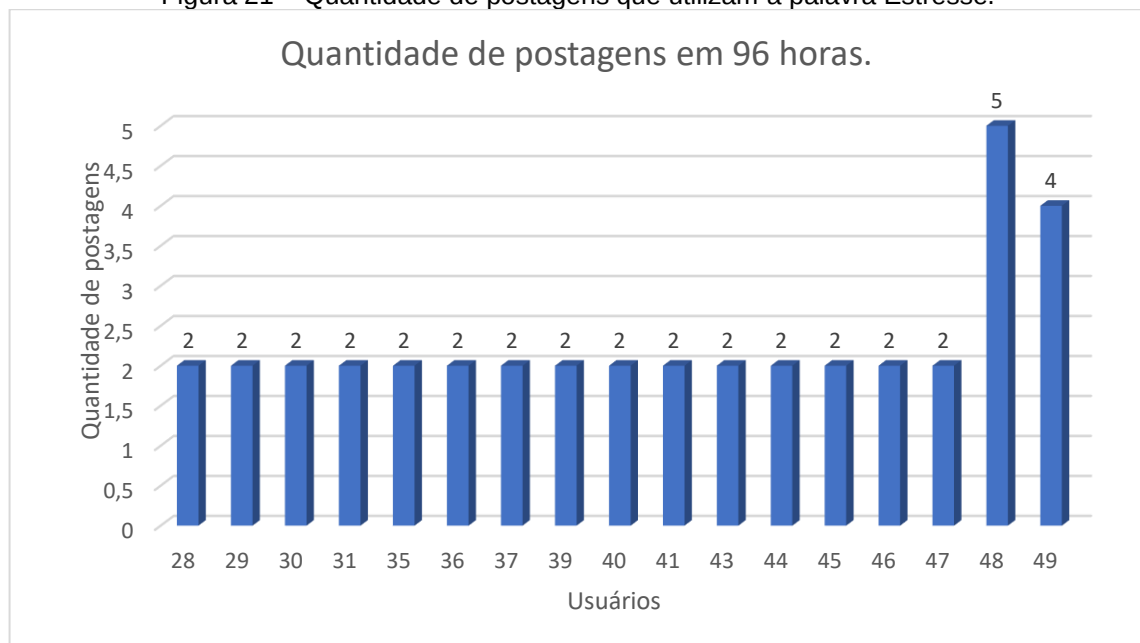
Figura 20 – Quantidade de postagens que utilizam a palavra Ansiedade.



Fonte: Elaborado pelo autor, 2022.

Em uma terceira análise, é realizada uma busca por quantidade de postagens, por usuário, que contêm a palavra “estresse”. Por meio da figura 21 é possível observar, para um total de 96 horas, as postagens que utilizam a palavra mencionada. Além desses usuários, alguns realizaram uma única postagem ou não realizaram postagens utilizando esse termo.

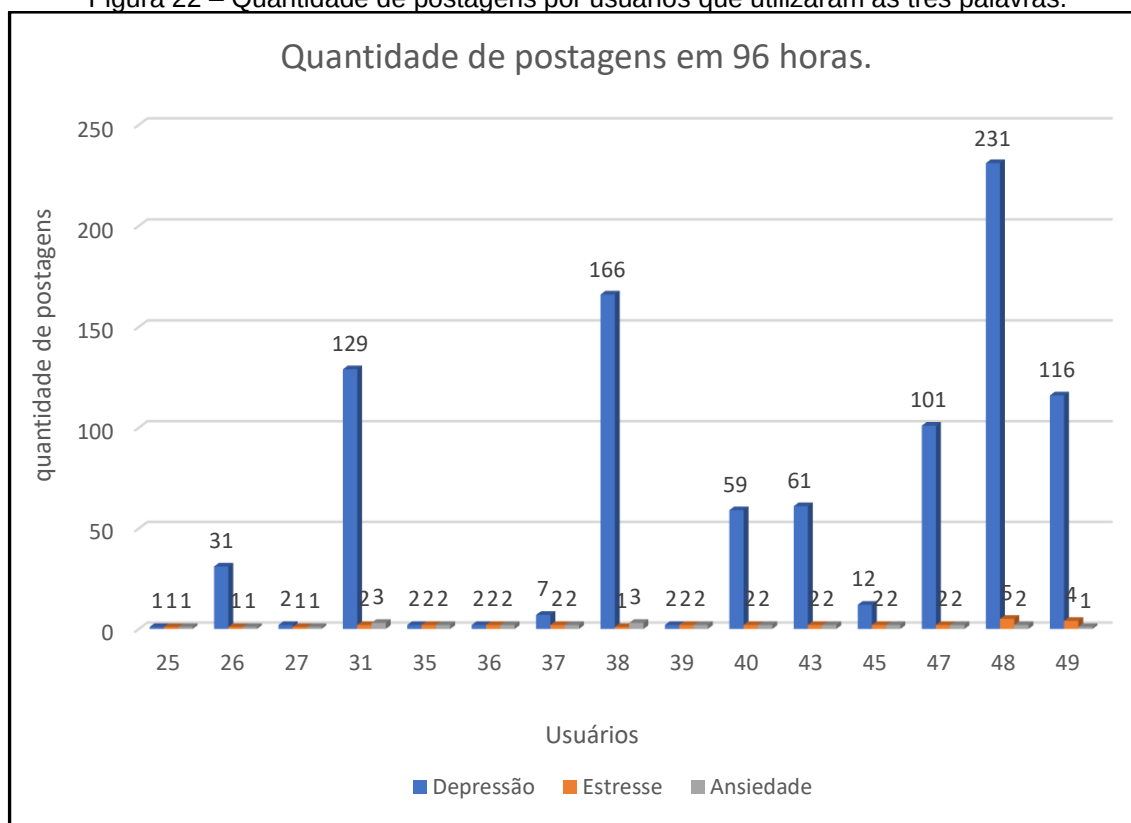
Figura 21 – Quantidade de postagens que utilizam a palavra Estresse.



Fonte: Elaborado pelo autor, 2022.

Em outra análise, é realizada uma busca pela quantidade de vezes em que as três palavras são utilizadas por um único usuário em suas postagens. Por meio da figura 22 é possível observar, para um total de 96 horas, as postagens e a quantidade de vezes que as palavras são utilizadas. Isso somente para os usuários que, durante as 96 horas de coleta de dados, fizeram uso das três palavras, não necessariamente juntas, em uma única postagem.

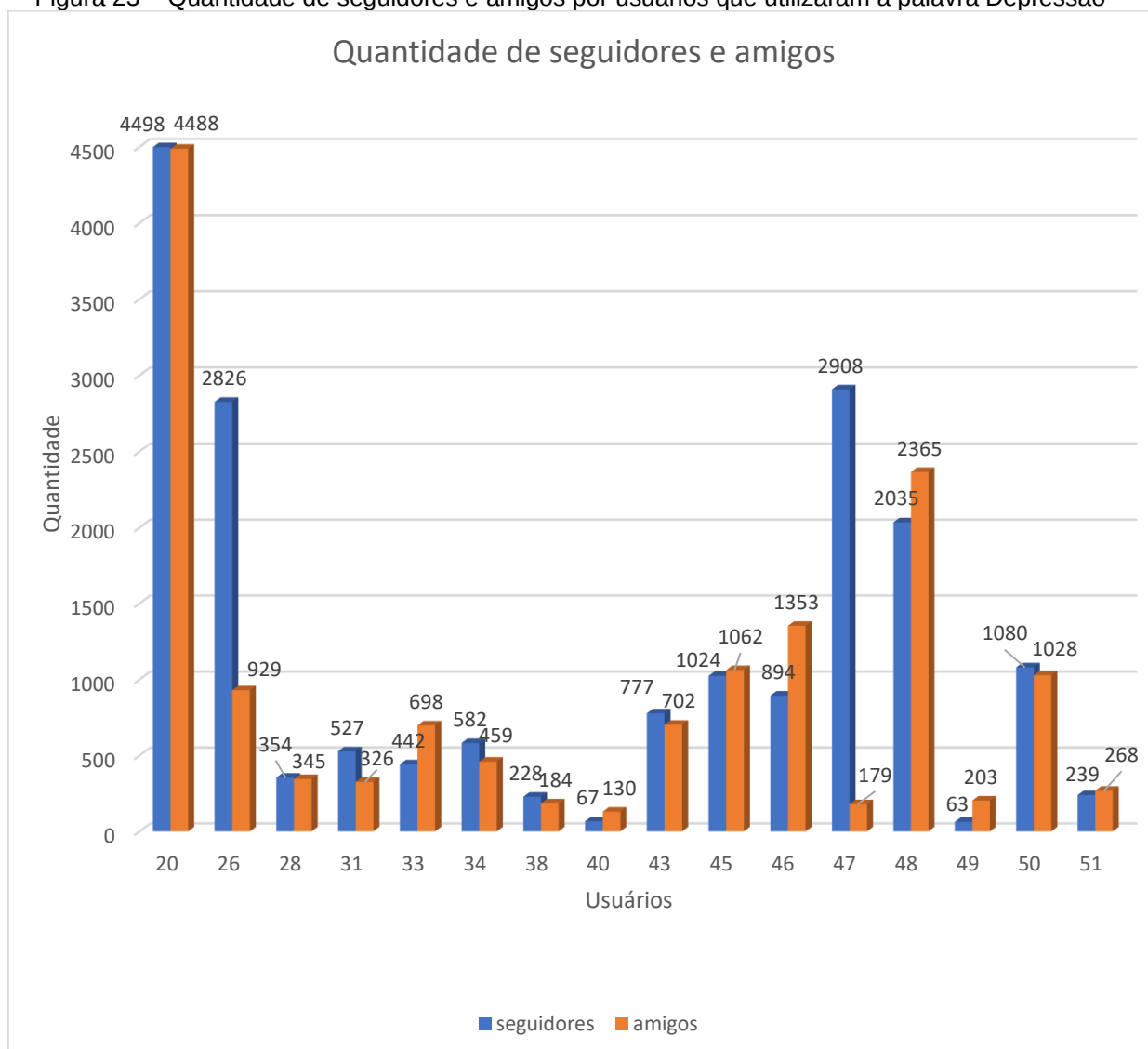
Figura 22 – Quantidade de postagens por usuários que utilizaram as três palavras.



Fonte: Elaborado pelo autor, 2022.

Por meio da figura 23 é possível, também, visualizar os resultados da associação dos usuários que mais realizaram postagens utilizando a palavra “depressão”, bem como suas respectivas quantidades de seguidores e amigos. Com isso pretende-se ver o quanto a mensagem postada é disseminada, ou seja, sempre que ela é publicada, os seguidores e amigos têm acesso a ela. Porém não é possível ver quanto deles visualizaram ou comentaram.

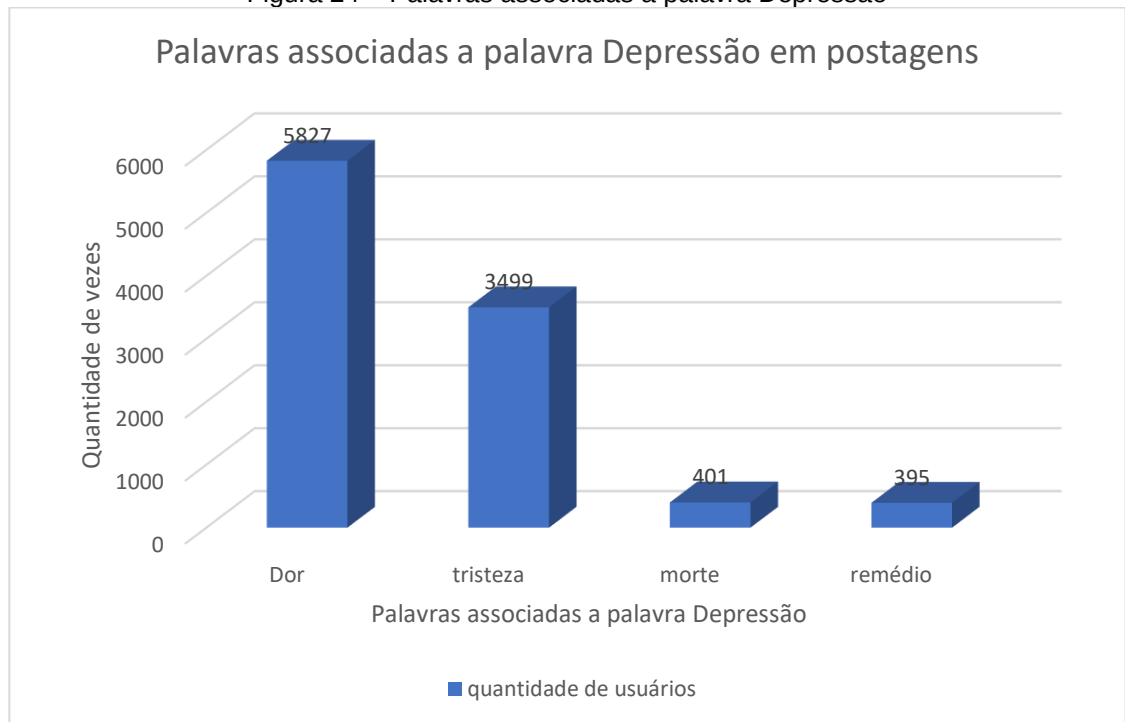
Figura 23 – Quantidade de seguidores e amigos por usuários que utilizaram a palavra Depressão



Fonte: Elaborada pelo autor, 2022.

Por meio da figura 24 é possível, ainda, visualizar o gráfico gerado pela associação de palavras-chave, associadas à palavra depressão. Nesse caso tem-se o total de postagens que utilizam esse termo associado a palavras como: dor, tristeza, morte e remédio. Foram desconsideradas todas as ocorrências dessas palavras em anúncios ou *links* de propaganda.

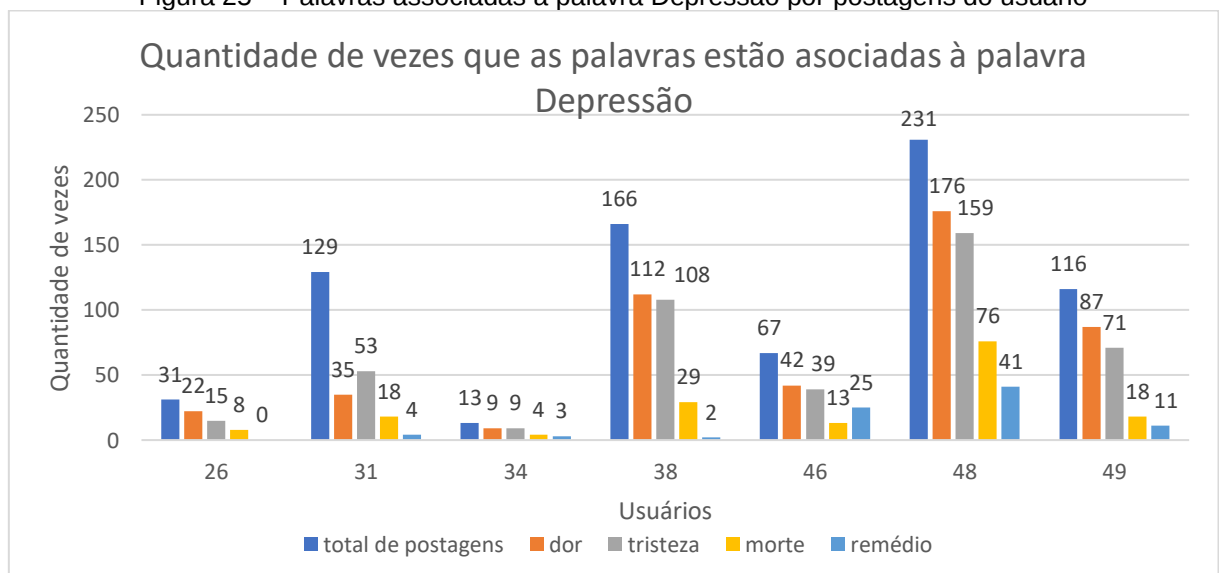
Figura 24 – Palavras associadas à palavra Depressão



Fonte: Elaborado pelo autor, 2022.

Outra análise também é realizada a fim de determinar a quantidade de vezes que um mesmo usuário utiliza a palavra “depressão”, associada às palavras “dor”, tristeza, morte ou remédio na mesma postagem, isso de forma proporcional ao total de postagens realizadas. São selecionados os usuários que mais fizeram uso das palavras “dor”, “tristeza”, “morte” e “remédio” associadas a postagens que têm a palavra “depressão”. Por meio da figura 25 é possível visualizar os resultados obtidos.

Figura 25 – Palavras associadas à palavra Depressão por postagens do usuário



Fonte: Elaborado pelo autor, 2023.

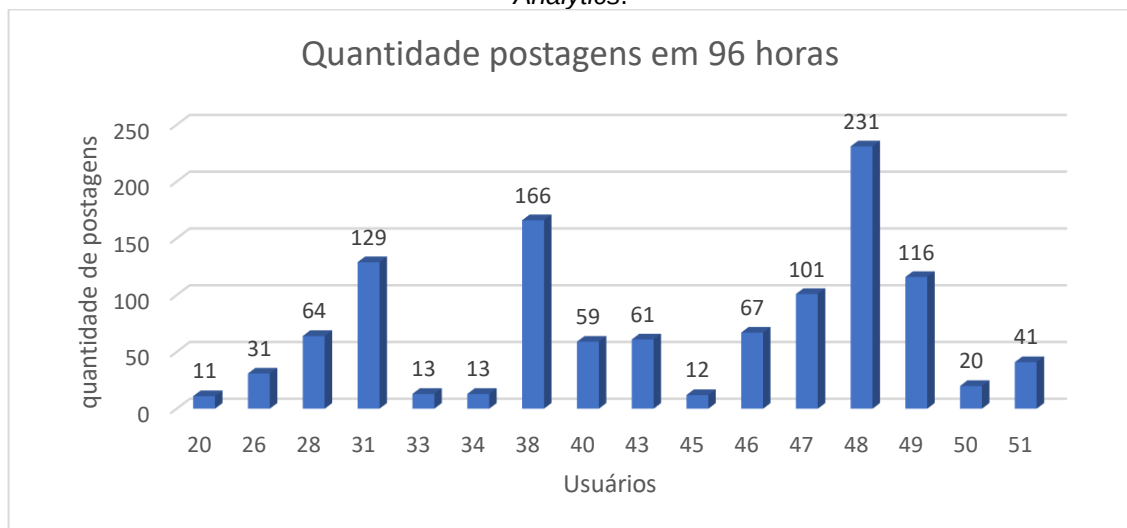
4.2 Resultados com a ferramenta *SAS Visual Text Analytics*

Com relação aos testes realizados com a ferramenta *SAS Visual Text Analytics*, é importante ressaltar que são os mesmos realizados com a ferramenta *SAS Enterprise Miner*.

Tanto neste primeiro teste quando nos demais, conforme dito anteriormente, o que os difere é que para os testes realizados no item 4.1 desta tese, o processo de mineração está fundamentado em análises estatísticas. Já para os realizados no item 4.2, o processo de mineração é o mesmo, porém, fundamentado na utilização de técnicas que utilizam os conceitos de redes complexas.

Para uma primeira análise, é realizada busca por quantidade de postagens por usuário que contêm a palavra depressão em suas postagens. Por meio da figura 26 é possível observar que, para um total de 96 horas, alguns usuários têm quantidade de postagens que utilizam a palavra “depressão” de forma bem acentuada. Outros, com quantidades menores de postagens foram desconsiderados.

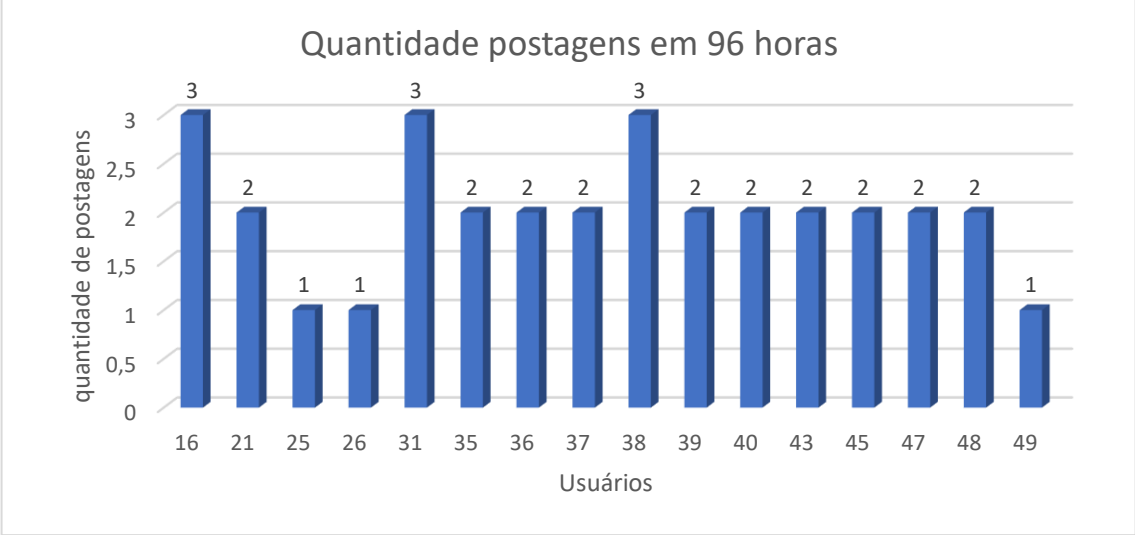
Figura 26 – Quantidade de postagens que utilizam a palavra Depressão obtida no *SAS Visual Text Analytics*.



Fonte: Elaborado pelo autor, 2023.

Para uma segunda análise, é realizada uma busca por quantidade de postagens, por usuário, que contivesse a palavra “ansiedade” em suas postagens. Por meio da figura 27 é possível observar, para um total de 96 horas, as postagens por usuários que utilizam o termo mencionado. Os demais usuários não utilizaram essa palavra em suas postagens.

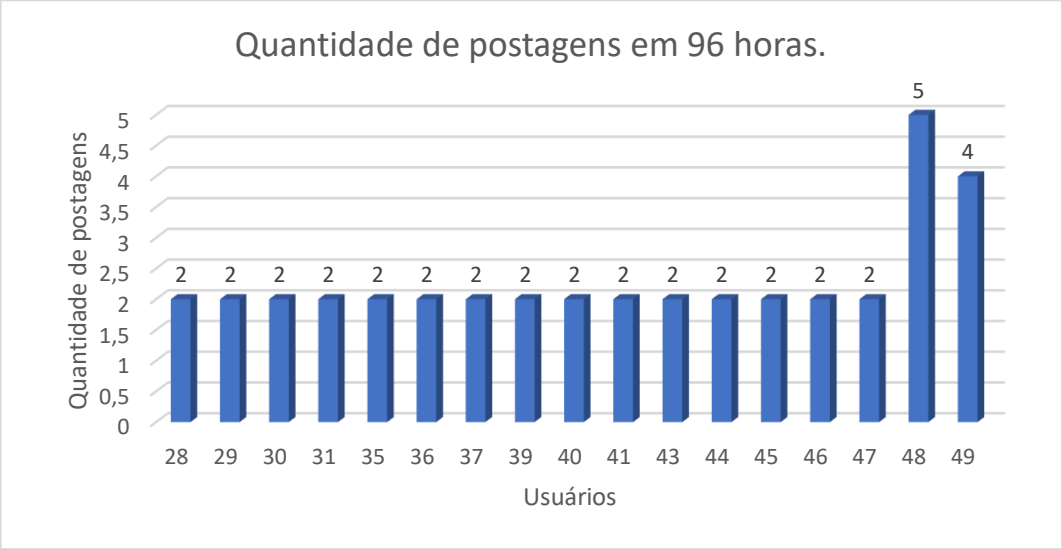
Figura 27 – Quantidade de postagens que utilizam a palavra Ansiedade obtida no SAS Visual Text Analytics.



Fonte: Elaborado pelo autor, 2023.

Em uma terceira análise, fez-se uma busca por quantidade de postagens, por usuário, que contêm a palavra “estresse”. Por meio da figura 28 é possível observar, para um total de 96 horas, as postagens por usuários que utilizam o termo supracitado. Além destes, alguns realizaram uma única postagem ou não realizaram postagens utilizando essa palavra.

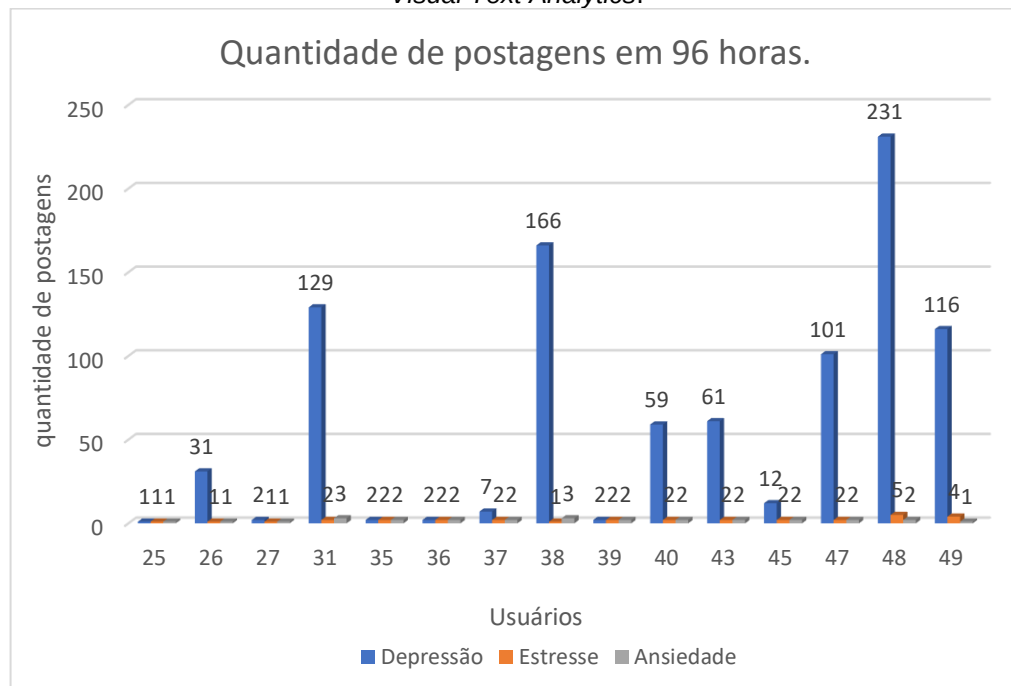
Figura 28 – Quantidade de postagens que utilizam a palavra Estresse obtida no SAS Visual Text Analytics.



Fonte: Elaborado pelo autor, 2023.

Em outra análise, busca-se a quantidade de vezes em que as três palavras são utilizadas por um único usuário em suas postagens. Por meio da figura 29 é possível observar, para um total de 96 horas, as postagens por usuários e as vezes nas quais os termos mencionados são utilizados nas postagens. Isso somente para os usuários que, durante as 96 horas de coleta de dados, fizeram uso das três palavras, não necessariamente juntas em uma única postagem.

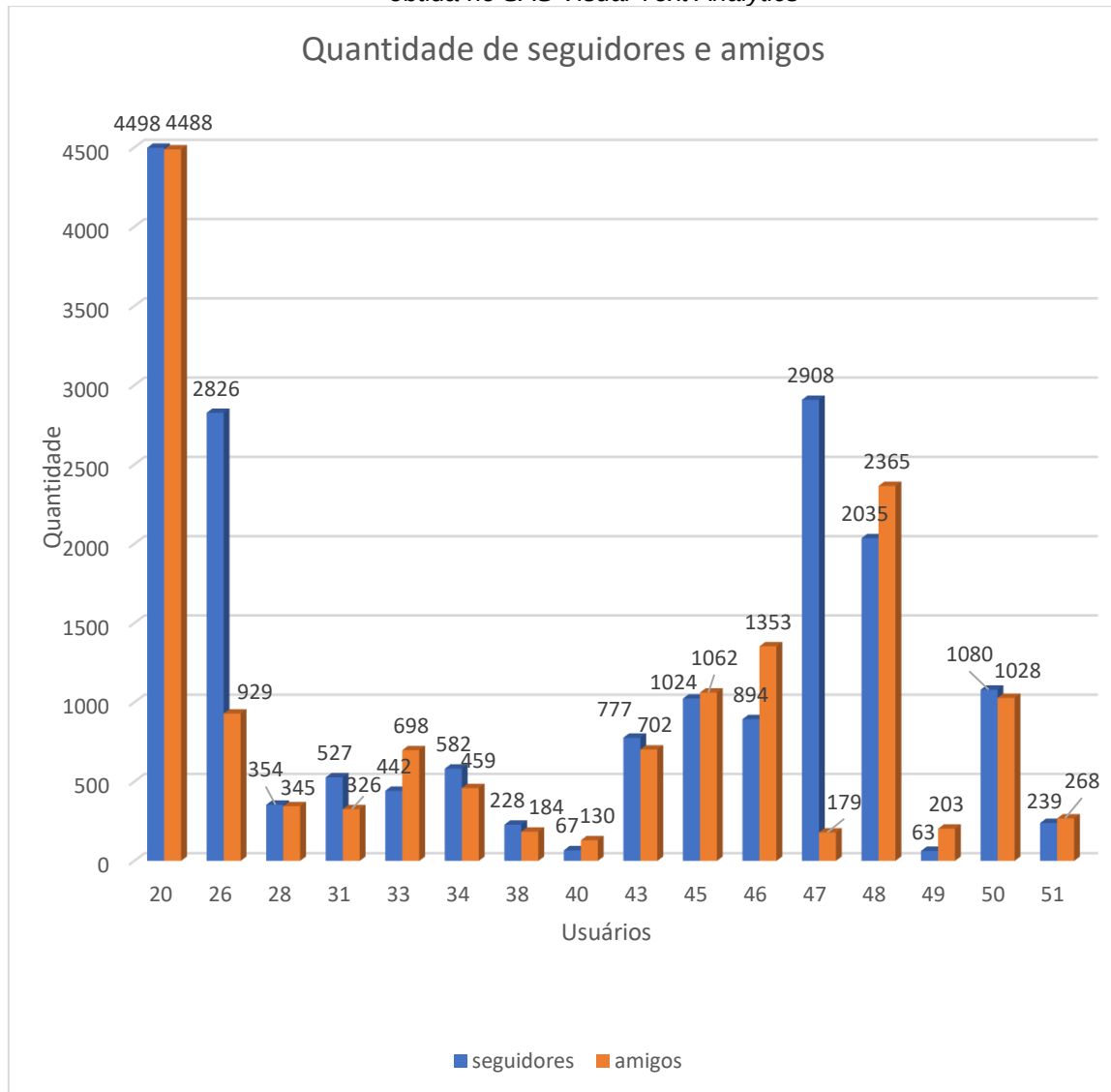
Figura 29 – Quantidade de postagens por usuários que utilizaram as três palavras obtida no SAS Visual Text Analytics.



Fonte: Elaborado pelo autor, 2023.

Por meio da figura 30 é possível visualizar os resultados da associação dos usuários que mais realizaram postagens, utilizando o termo “depressão”, bem como pode-se ver a disseminação deste termo, a partir da quantidade de seguidores e amigos. Com isso pretende-se ver o quanto a mensagem postada tem amplitude, ou seja, sempre que ela é publicada, os seguidores e amigos têm acesso a ela. Porém não é possível ver quanto deles visualizaram ou comentaram.

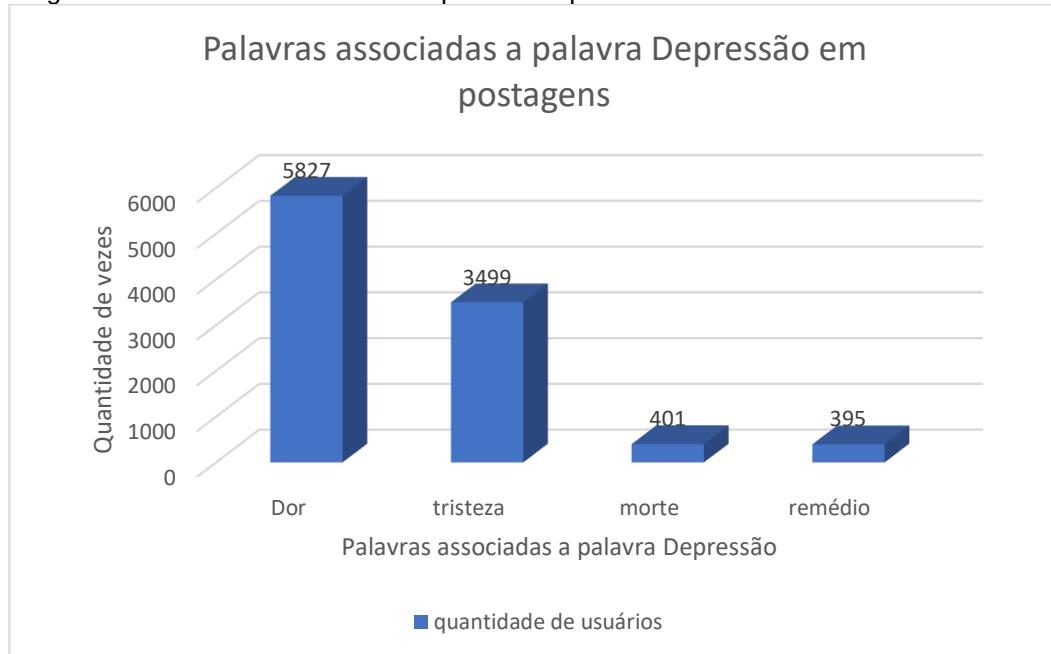
Figura 30 – Quantidade de seguidores e amigos por usuários que utilizaram a palavra Depressão obtida no SAS Visual Text Analytics



Fonte: Elaborada pelo autor, 2023.

Por meio da figura 31 é possível visualizar o gráfico gerado pela associação de palavras-chave associadas à palavra “depressão”. Nesse caso tem-se o total de postagens que utilizam a palavra associada às palavras: “dor”, “tristeza”, “morte” e “remédio”. São desconsideradas todas as ocorrências dessas palavras em anúncios ou *links* de propaganda.

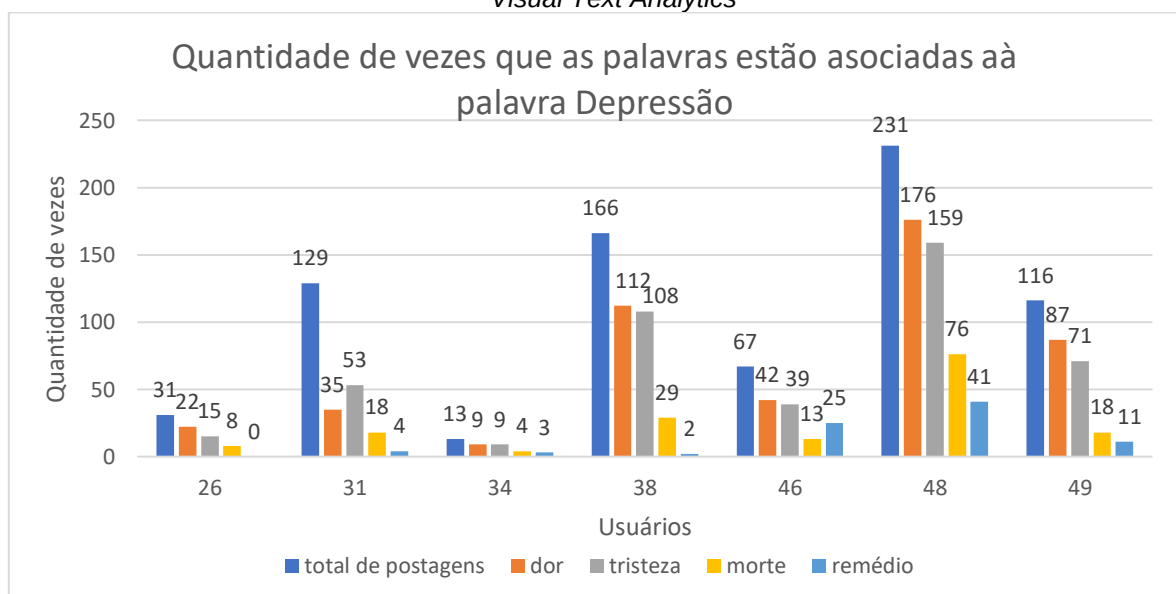
Figura 31 – Palavras associadas à palavra Depressão obtidas no SAS Visual Text Analytics



Fonte: Elaborado pelo autor, 2022.

Realizou-se, ainda, outra análise, a fim de determinar a quantidade de vezes que um mesmo usuário utiliza a palavra “depressão”, associada às palavras: “dor”, “tristeza”, “morte” ou “remédio” na mesma postagem. Isso de forma proporcional ao total de postagens realizadas. São selecionados os usuários que mais fizeram uso dos termos mencionados associados a postagens que têm a palavra “depressão”. Por meio da figura 32 é possível visualizar os resultados obtidos:

Figura 32 – Palavras associadas à palavra Depressão por postagens do usuário obtidas no SAS Visual Text Analytics



Fonte: Elaborado pelo autor, 2023.

4.3 Análise dos resultados

É possível observar que, de posse dos resultados obtidos no processo de mineração de dados por análise estatística, o mesmo processo, fundamentado no conceito de redes complexas, tem resultados iguais. Assim, pode-se dizer que não foram observadas alterações em nenhuma das análises.

Um resultado que também pode ser considerado expressivo é que alguns usuários precisam ser acompanhados por um especialista, de acordo com a quantidade de postagens realizadas por um mesmo usuário em 96 horas. Alguns deles têm a quantidade de postagens utilizando a palavra “depressão”, muito além do considerado normal, obtendo uma quantidade de postagens considerada excessiva servindo de alerta.

Algo que chama a atenção é que foram coletadas 109.093 postagens que têm pelo menos umas das três palavras: “estresse”, “ansiedade” e “depressão” e todas as mensagens, têm a presença da palavra “depressão”.

A palavra “ansiedade”, nesses casos, é a menos utilizada, e “estresse” aparece com um pouco mais de frequência, porém, nada que é inicialmente relevante para chamar a atenção na análise dos resultados da mineração.

Já no que diz respeito à palavra “depressão”, ela é a que mais chama a atenção pela popularidade de sua utilização e por sua repetição, por meio de alguns usuários. A partir dessa análise, são verificadas as quantidades dos seus seguidores e amigos que mais utilizaram a palavra em suas postagens, para que fosse possível determinar a quantidade de pessoas que receberam essas mensagens.

Foi realizada uma última análise a fim de determinar a quantidade de vezes que um mesmo usuário utiliza a palavra “depressão”, associada às palavras: “dor”, “tristeza”, “morte” ou “remédio” na mesma postagem, isso de forma proporcional ao total de postagens realizadas. Entre as palavras-chave mencionadas, é importante se dizer que “dor” e “tristeza”, são utilizadas com mais frequência em postagens que envolvem a palavra “depressão”.

5. CONCLUSÃO

No presente trabalho são propostos dois processos de mineração de dados por meio de ferramentas computacionais. Uma delas faz uso de métodos estatísticos e a outra utiliza o conceito de redes complexas. Ambas foram utilizadas para realizar a mineração proveniente de uma base de dados, originária das postagens em perfis públicos, de uma rede social digital (*Twitter*).

O intuito é determinar se é possível a utilização desse processo para obter informações que servem de apontamento para que se possa descobrir se alguns usuários dessa rede social, podem estar passando por problemas psicológicos, devido a transtornos socioemocionais.

Os resultados obtidos fornecem indícios de que é possível realizar essa análise. Diversos parâmetros foram utilizados nos processos de mineração de dados em ambas as ferramentas e, o cruzamento das informações convergiu para alguns resultados, os quais apontam que alguns usuários, em um período de 96 horas, apresentam padrões comportamentais fora do se considera “normal” – tendo em vista a maioria das pessoas que fazem uso dessa rede social digital – tanto no que diz respeito aos termos utilizados quanto no que tange à quantidade de postagens.

Devido a Lei Geral de proteção dos Dados pessoais, os usuários foram identificados por número e não por nome ou apelido, garantindo o anonimato. Todas as 109.093 postagens que foram selecionadas para compor a base de dados e que são originárias de uma busca em que foram selecionadas postagens com pelo menos uma das três palavras-chave: “depressão”, “estresse”, “ansiedade” contêm a palavra “depressão”. Ficando expresso, portanto, que nenhuma das 109.093 postagens têm o uso das Palavras-chave: “estresse” e/ou “ansiedade” sem o uso do termo “depressão”. Esse fator foi considerado para a definição dos parâmetros utilizados no processo de mineração.

Um exemplo do que é mencionado acima se manifesta quando se vê que, o usuário 48 realizou 231 postagens, o usuário 38 realizou 166 postagens e o usuário 31 realizou 129 postagens em um intervalo de 96 horas. Se for considerado que uma pessoa precisa ter intervalos de sono, que essa pessoa deveria estar trabalhando ou estudando ou, ainda, que ela tenha afazeres do dia a dia, essa quantidade postagens é preocupante, devido ao uso da palavra “depressão” em todas elas.

Verificou-se, então, que alguns usuários com o perfil selecionado podem estar, de alguma forma, influenciando ou sendo influenciados por demais usuários da rede social digital, isso se deve ao fato de que, alguns possuem uma quantidade significativa de seguidores e amigos, como é o caso do usuário 20, que tem 4.498 seguidores e 4488 amigos e do usuário 48, que possui 2035 seguidores e 2365 amigos. Isso, contudo, pode ser considerado um apontamento para outros estudos. O que foi considerado é a quantidade de pessoas que pertencem a um mesmo grupo de amigos, terem acesso às postagens, mas não o impacto causado pelas mesmas.

Outro fator que contribui para determinar que alguns usuários, podem sim, estar acometidos de alguma síndrome socioemocional, é que alguns deles fazem uso das palavras-chave: “dor”, “tristeza”, “morte” e “remédio” na mesma postagem em que a palavra-chave “depressão” é utilizada. Por exemplo, o usuário 26, em um total de 31 postagens que possui a palavra-chave “depressão”, utilizou 22 vezes o termo “dor” e, 15 vezes o léxico “tristeza”. O usuário 48, em um total de 231 postagens, utilizou 176 vezes a palavra dor e, 159 vezes, “tristeza”.

A incidência dos termos “morte” e “remédio” aparecem associadas à palavra-chave “depressão”, principalmente em postagens dos usuários 46 e 49. Para o usuário 46, no total de 67 postagens, “morte” é utilizada 13 vezes e “remédio” é utilizada 25 vezes. Para o usuário 49, no total de 116 postagens, “morte” é utilizada 18 vezes e “remédio” é utilizada 11 vezes.

Por fim, é possível concluir que o processo de mineração de dados, quando utilizado para analisar postagens públicas em redes sociais digitais, contribui para a detecção, de forma não intrusiva, de perfis de pessoas que podem estar acometidas de síndromes socioemocionais tais como depressão, estresse e ansiedade.

Cabe ressaltar, porém, que esses estudos apontaram uma maior facilidade em obter indícios relacionados à depressão. Com relação ao estresse e a ansiedade, mais estudos se fazem necessários, mas, entende-se que ajustes no processo de mineração e na escolha por outras palavras-chave podem remeter a outros resultados, mais conclusivos.

6. PROPOSTAS PARA TRABALHOS FUTUROS

Diante de tudo que foi abordado nesta tese, fica como proposta para trabalhos futuros aplicar outras técnicas de mineração de dados e comparar os resultados com as técnicas apresentadas, para verificar se por meio da utilização de outros processos os resultados podem ser melhorados.

É possível também, realizar pesquisas por outras palavras-chave que associadas às síndromes propostas neste trabalho, possam fornecer informações que, ao serem analisadas, indiquem possíveis problemas socioemocionais, dos quais os usuários de redes sociais possam estar acometidos.

Outra proposta para trabalhos futuros, é a realização de um estudo mais aprofundado dos algoritmos utilizados para a mineração de dados, gerando um algoritmo próprio, sendo esse, fundamentado a uma ou mais técnicas de mineração, que permitam uma continuação dos estudos nesta linha de pesquisa, aplicando a mineração de dados à área da psicologia. Principalmente, quando o assunto é estudar os problemas socioemocionais que têm, em seu índice, o aumento no número de pessoas acometidas de tais síndromes e que podem, por sua vez, externar isso em redes sociais.

Referências

ABBOTT, P. A.; LEE, S. M. **Data mining and knowledge discovery**. In: **Saba VK, McCormick KA. Essentials of nursing informatics**. 4 ed. New York: McGraw-Hill Medical Pub. Division, 2006.

ALBERT, R.; BARABÁSI, A. Statistical mechanics of complex networks. **Reviews of modern physics**, v. 74, n. 1, 2002, p. 47-52.

ALLEN A.J.; LEONARD, H.; SWEDO, S. E.; Current knowledge of medications for the treatment of childhood anxiety disorders. **J. Am Acad Child Adores Psychiatry**, v. 34, [s.n.], 1995, p. 976-986.

APA. **DSM-IV-TR: manual diagnóstico e estatístico de transtornos mentais**. 4. ed. Porto Alegre: Artmed, 2002.

ARRUDA, G. F. **Mineração de dados em redes complexas: estrutura e dinâmica**. Tese (Doutorado) – Universidade de São Paulo, 2013, 105 p.

BAKER, R. S. Mineração de dados educacionais: Oportunidades para o brasil. **Revista Brasileira de Informática na Educação**, v. 19, n.2, 2011, p. 3-13.

BAKER, R. S.; CORBETT, A. T.; ROLL, I.; KOEDINGER, K. R. Developing a generalizable detector of when students game the system. **User Modeling and User-Adapted Interaction**, v. 8, n. 3, 2008, p. 287–314.

BEN-GAL, I. Bayesian Networks. **Encyclopedia of Statistics in Quality and Reliability**. New York: John Wiley & Sons LTDA, 2007.

BERNSTEIN, G. A.; BORCHARDT, C. M.; Prewire A. R. Anxiety disorders in children and adolescents: a review of the past 10 years. **J. Am Acha Child Adolesc Psychiatry**, v. 35, [s.n.], 1996, p.1110-1119.

BESSAS, Alex. O índice de suicídio entre adolescentes cresce 24 vezes em nove anos. **O tempo**. Disponível em: <<https://www.otempo.com.br/interessa/indice-de-suicidio-entre-adolescentes-cresce-24-em-nove-anos-1.2173731>>. Acesso em: 25 mai. 2020.

BOAS, P. R. V. **Efeito da amostragem nas propriedades topológicas de redes complexas**. Tese (Doutorado) – Universidade de São Paulo, 2008, 120 p.

BOBBIO, A.; PORTINALE, L.; MINICHINO, M.; CIANCAMERLA, E. Improving the Analysis of Dependable Systems by Mapping Fault Trees into Bayesian Networks. **Reliability Engineering & System Safety**, vol. 71, n. 1, 2001, p.249-260.

BRACHMAN, R. J.; ANAND, T. The Process of Knowledge Discovery in Databases. **The KDD Process for Extracting Useful Knowledge from Volumes of Data**. v. 94, n. 3, 1996, p. 37-57.

CANALTECH. **Empresa** **instagram**. Disponível em: <<https://canaltech.com.br/empresa/instagram/>>. Acesso em: 06 abr. 2020.

CANALTECH. **Empresa** **Pinterest**. Disponível em: <<https://canaltech.com.br/empresa/pinterest/>>. Acesso em: 07 abr. de 2020.

CANALTECH. **Guia do Snapchat para iniciantes**. Disponível em: <<https://canaltech.com.br/redes-sociais/guia-do-snapchat-para-iniciantes/>>. Acesso em: 07 abr. 2020.

CARDOSO, O. N. P.; MACHADO, R. T. M. Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras. **Rev. Adm. Pública**, v. 42, n. 3, 2008, p. 495-528.

CESTARI, D. M. **Classificação de sinais de epilepsia utilizando redes complexas**. Tese (doutorado) – Universidade de São Paulo, 2017, 192 p.

COSTA, E.; BAKER, R. S.; AMORIN, L.; MAGALHÃES, J.; MARINHO, T. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. **Jornada de Atualização em Informática na Educação**, v. 1, n. 1, 2013, p. 1-29.

DANTAS, Tiago. "Youtube". **Brasil Escola**. Disponível em: <<https://brasilecola.uol.com.br/informatica/youtube.html>>. Acesso em: 06 abr. 2020.

D'EL REY, G. J. F. Fobia social: mais do que uma simples timidez. **Arquivos de Ciências da Saúde da Unipar**, v. 5, n. 3, 2001, p.273-276.

DIGGLE, P.; LIANG, K.; ZEGGER, S. **Analysis of longitudinal data**. Oxford University Press: New York, 1996.

DUARTE, M. & REGO, M. Comorbilidade entre a Depressão e doenças clínicas em um ambulatório de geriatria. **Cadernos de Saúde Pública**, v. 23, n. 3, 2007, p. 691-700.

EFRAIM, I. **Tudo o que a grande mente capta**. Rio de Janeiro: Editora Gente, 1993.

ESTAVAM, I. D.; FORMIGA, N. S.; FRANCO, J. B. M.; BONIFÁCIO, Élide D. N. C.; FERREIRA, S. V.; COSTA, E. L.; FERREIRA, M. C.; PEREIRA, J. F. A predictive study about Organizational Support and Psychological Capital at work regarding emotional disorder during Covid-19 isolation period. **Research, Society and Development**, v. 11, n. 7, 2022, p. 1-12. Disponível em: <<http://doi.org/10.33448/rsd-v11i7.29883>>. Acesso em: 27 dez. 2022.

EY, H.; BERNARD, P.; BRISSET, C. As psicoses esquizofrênicas. **Manual de psiquiatria**. Rio de Janeiro: Masson, 1985, p. 535-615.

FACEBOOK. **Company**. Disponível em: <<https://about.fb.com/br/company-info/>>. Acesso em: 10 abr. 2020.

FACEBOOK MENSSENGER. **Features**. Disponível em: <<https://www.messenger.com/features>>. Acesso em: 10 abr. 2020.

FAYYAD, U. M.; PIATESKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery: An Overview. **Advances in Knowledge Discovery and Data Mining**. California: AAAI Press, 1996, p. 1-36. Disponível em: <<https://dl.acm.org/doi/book/10.5555/257938>>. Acesso em: 15 jan. 2023.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining: um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações**. São Paulo: Elsevier, 2005.

HAN J.; KAMBER, M. **Data mining: concepts and techniques**. 2 ed. Boston: Elsevier, 2006.

HAYKIN, S. **Redes Neurais: Princípios e Prática**. 2.ed. Porto Alegre: Bookman, 2001.

HIRSHFELD, D. R.; ROSENBAUM J. F.; FREDMAN, S. J.; KAGAN J. The neurobiology of childhood anxiety disorders. **Neurobiology of mental illness**. New York: Oxford University Press; 1999. p. 823-38.

HOUAISS, A.; VILLAR, M. S.; FRANCO, F. M. **Dicionário da Língua Portuguesa**. Rio de Janeiro: Moderna, 2001, p 1264.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 4 ed. New Jersey: Prentice-Hall, 1999.

JUST, A. P.; ENUMO, S. R. F. Problemas emocionais e de comportamento na adolescência: o papel do Estresse. **Bol. - Acad. Paul. Psicol.**, São Paulo, v. 35, n. 89, 2015, p. 350-370.

KATON, W. J. Clinical and Health Services Relationships Between Major Depression, Depressive Symptoms, and General Medical Illness. **Biol. Psychiatry**, v. 54, [s.n.], 2003, p. 216-26.

KHATTREE, R.; NAIK, D. N. **Multivariate data reduction and discrimination with SAS software**. North Caroline: SAS Institute Inc., 2000.

LINCOLN, T. M.; RIEF, W.; HAHLOWEG, K.; FRANK, M.; SCHROEBER, B; FIEGENBAU, W. Effectiveness of an empirically supported treatment for social phobia in the field. **Behaviour Research and Therapy**, v. 41, n. 11, 2003, p. 1251- 1269.

LINKEDIN. **About**. Disponível em: <https://about.linkedin.com/pt-br?trk=homepage-basic_directory_aboutUrl>. Acesso em: 06 abr. 2020.

LOPES, P. D. A. **Agrupamento de Dados Semissupervisionado na Geração de Regras Fuzzy**. Dissertação (Mestrado). Universidade Federal de São Carlos, 2010, 74 p.

LOVIBOND, SH; LOVIBOND, PF. **Manual for the Depression Anxiety Stress Scales**. 4. ed. Sydney: Psychology Foundation, 2004.

LOVIBOND, P. F.; LOVIBOND, S. H. The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck depression and anxiety inventories. **Behaviour Research and Therapy**, v. 33, n. 3, 1995, p. 335-343.

MECHIE, D.; SPIEGELHALTER, D.; TAYLOR, C. Machine Learning, Neural and Statistical Classifications. **American Statistical Association**, v. 91, n. 433, Mar. 1996, p. 436-438. Disponível em: <<https://www.jstor.org/stable/2291432>>. Acesso em: 15 ago. 2022.

MEIRA, C. A. A.; RODRIGUES, L. H. A.; MORAES, S. A. Análise da epidemia da ferrugem do cafeeiro com árvore de decisão. **Trop Plant Pathol**, v. 33, n. 2, 2008, p.114-24. Disponível em: <<https://doi.org/10.1590/S1982-56762008000200005>>. Acesso em: 15 ago. 2022.

NEAPOLITAN, R. E. **Learning Bayesian Networks**. Upper Saddle River: Pearson, 2004.

OLIVEIRA, Luan S.; CAMPOS, Guilherme O., DA SILVA, Rodrigo S. **Mineração de Dados em Redes Sociais**. Disponível em: <https://www.researchgate.net/publication/261359327_Mineraao_de_Dados_em_Redes_Sociais>. Acesso em: 5 fev. 2021.

OMS. (Organização Mundial de Saúde). Classificação de transtornos mentais e de comportamento da CID-10. **Descrições clínicas e diretrizes diagnósticas**. Porto Alegre: Artes Médicas, 1993.

PAPALIA, D. E.; OLDS, S. W.; FELDMAN, R. D. **Desenvolvimento Humano**. Trad. D. Bueno. 10 ed. Porto Alegre: AMGH, 2010, p. 395-462.

PARTHASARATHY S. Data mining at the crossroads: successes, failures and learning From Them. **13nd. international conference on Knowledge discovery and data mining**, [s.v.], [s.n], August. 2007, p.1053-1055. Disponível em: <<https://dl.acm.org/doi/abs/10.1145/1281192.1281311>>. Acesso em: 12 nov. 2021.

PERERA, L. C. J.; KIMURA, Herbert; HORTA, R. A. M.; LIMA, F. G. Uma Análise em Data Mining: Árvores de Decisão, Redes Neurais e Support Vector Machines. **Anais do XXXV Encontro da ANPAD**, v.1, [s.n.], set, 2011, p.1-15.

PINTEREST. **About**. Disponível em: <<https://about.pinterest.com/pt-br>>. Acesso em: 07 abr. 2020.

REGAZZI, A. J. **Análise multivariada, notas de aula INF 766**. Departamento de Informática da Universidade Federal de Viçosa. v.2. ED. UFV: 2000.

RODRIGUES, R. J. Information systems: the key to evidence based health practice. **Bull World Health Organ**, v. 78, n.11, 2000, p.1344-51. Disponível em <<https://pubmed.ncbi.nlm.nih.gov/11143195/>>. Acesso em: 11 Dez. 2022.

RODRIGUES, F. A. **Caracterização, classificação e análise de redes complexas**. Instituto de Física de São Carlos. EDUSP: São Carlos, 2007.

ROSEN, J. B.; SCHILKIN, J. From normal fear to pathological anxiety. **Psychol. Rev.**, [s.v.], n. 105, 1998, p. 325-50.

SALOMAN, Laurie. Depression and Anxiety in College Students: A True Epidemic? **Psycom**. Disponível em: <<https://www.psycom.net/depression-anxiety-college-students>>. Acesso em: 5 fev. 2021.

SAS. **Company information old before redesign profile**. Disponível em: <https://www.sas.com/pt_br/company-information-old-before-redesign/profile>. Acesso em: 04 ago. 2020.

SAS. **Enterprise Miner**. Disponível em: <https://www.sas.com/content/dam/SAS/en_us/doc/factsheet/sas-enterprise-miner-101369.pdf>. Acesso em: 06 ago. 2020.

SCHANK, T.; WAGNER, D. **Approximating clustering-coefficient and transitivity** Fakultät Karlsruhe fakultät für informatike. Karlsruhe Press, 1998.

SILVA, A.; FERREIRA, M. Gestão do conhecimento e capital social: as redes e sua importância para as empresas. **Informação & Informação**, v. 12, n. esp., 2007, p. 1-33. Disponível em: <<http://doi.org/10.5433/1981-8920.2007v12n1espp125>>. Acesso em: 18 fev. 2022.

SNAPCHAT. **About**. Disponível em: <<https://www.snap.com/pt-BR/>>. Acesso em: 07 abr.2020.

STEIN, M. B.; TORGRUD, L. J.; WALKER, J. R. Social phobia symptoms, subtypes and severity. **Archives of General Psychiatry**, v. 57, n.9, 2000, p.1046-1052.

SUMATHI S.; SIVANANDAM S. N. **Data mining trends and knowledge discovery, studies in computational intelligence (SCI)**. London: Springer Press, 2006, p. 185–194. Disponível em: <<https://doi.org/10.1007/1-84628-183-0>>. Acesso em: 22 jan. 2022.

SYLVESTER, C. Separation anxiety disorder and other anxiety disorder. **Kaplan and Sadock's comprehensive textbook of psychiatry**. 7 ed. Baltimore: Lippincott Williams & Wilkins; 2000, p. 2770-2781.

TOMAÉL, M. I.; MARTELETO, R. M. Redes sociais: posição dos atores no fluxo da informação. **Revista Eletrônica de Biblioteconomia e Ciência da Informação**, [s.v.], n. esp, 2006, p. 75-91.

TORRES, C. **A bíblia do marketing digital**. São Paulo: Editora Novatec, 2009.

TRANI, D. B. D. A. A Viabilidade da Mineração de Dados como Suplemento nas Redes Sociais Visando ao Monitoramento Pedagógico e a Aprendizagem Colaborativa do Educando. **Revista Prospectus**, v. 2, n. 2, Ago/Fev 2020, p. 182-223.

TWITTER. **About**. Disponível em: <<https://about.twitter.com/pt.html>>. Acesso em: 06 abr. 2020.

TWITTER. **Help**. Disponível em: <<https://help.twitter.com/pt/new-user-faq>>. Acesso em: 06 abr. 2020.

VAIE M. M.; MOURA, D. J.; NÄÄS, I. A.; OLIVEIRA, S. R. M.; RODRIGUES, L. H. A. Data mining to estimate broiler mortality when exposed to heat wave. **Sci Agric.**, v.65, n.3, 2008, p. 223-229.

VALLADA FILHO, H.; BUSATTO FILHO, G. Esquizofrenia. In: ALMEIDA, P.; DRACTU, L.; LARANJEIRA, R. (Orgs.). **Manual de psiquiatria**. Rio de Janeiro: Guanabara Koogan, 2006, p. 127-150.

VERSIANI, M.; REIS, R.; FIGUEIRA, I. Diagnóstico do transtorno depressivo na infância e adolescência. **J. Bras Psiquiatria**, v. 49, n. 10, 2000, p. 367-382.

VIEIRA, L. M. **Antivacinação, redes sociais e desinformação: uma análise em um grupo do Facebook através da mineração de dados**. Dissertação (Mestrado em Comunicação) - Universidade Federal de Goiás, 2020, 84 p.

WHATSAPP. **About**. Disponível em: <<https://www.whatsapp.com/about/>>. Acesso em: 10 abr. 2020.

ZACHARIAS, Carlos Eduardo Kerbeg. **Quais são os principais perigos à saúde mental dos jovens na atualidade?** Disponível em: <<https://hospitalsantamonica.com.br/quais-sao-os-principais-perigos-a-saude-mental-dos-jovens-na-atualidade>>. Acesso em: 10 fev. 2021.