UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

THIAGO DE ANDRADE CORRÊA

Uso de ontologias para filtragem de base de dados para melhora de desempenho de classificadores em análise de sentimentos

Goiania
2021

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

AUTORIZAÇÃO PARA PUBLICAÇÃO DE TRABALHO DE CONCLUSÃO DE CURSO EM FORMATO ELETRÔNICO

Na qualidade de titular dos direitos de autor, **AUTORIZO** o Instituto de Informática da Universidade Federal de Goiás – UFG a reproduzir, inclusive em outro formato ou mídia e através de armazenamento permanente ou temporário, bem como a publicar na rede mundial de computadores (*Internet*) e na biblioteca virtual da UFG, entendendo-se os termos "reproduzir" e "publicar" conforme definições dos incisos VI e I, respectivamente, do artigo 5º da Lei nº 9610/98 de 10/02/1998, a obra abaixo especificada, sem que me seja devido pagamento a título de direitos autorais, desde que a reprodução e/ou publicação tenham a finalidade exclusiva de uso por quem a consulta, e a título de divulgação da produção acadêmica gerada pela Universidade, a partir desta data.

Título: Uso de ontologias para filtragem de base de dados para melhora de desempenho de classificadores em análise de sentimentos

Autor(a): Thiago de Andrade Corrêa

Goiania, 21 de Janeiro de 2021.

Thiago de Andrade Corrêa – Autor

Dra. Deborah S. A. Fernandes – Orientadora

THIAGO DE ANDRADE CORRÊA

Uso de ontologias para filtragem de base de dados para melhora de desempenho de classificadores em análise de sentimentos

Trabalho de Conclusão apresentado à Coordenação do Curso de Computação do Instituto de Informática da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Bacharel em Computação.

Área de concentração: Ciência de dados, Análise de sentimentos, Inteligência artificial.

Orientadora: Profa. Dra. Deborah S. A. Fernandes

Goiania
2021

THIAGO DE ANDRADE CORRÊA

Uso de ontologias para filtragem de base de dados para melhora de desempenho de classificadores em análise de sentimentos

Trabalho de Conclusão apresentado à Coordenação do Curso de Computação do Instituto de Informática da Universidade Federal de Goiás como requisito parcial para obtenção do título de Bacharel em Computação, aprovada em 21 de Janeiro de 2021, pela Banca Examinadora constituída pelos professores:

Profa. Dra. Deborah S. A. Fernandes
Instituto de Informática – UFG
Presidente da Banca

Profa. Dra. Luciana de Oliveira Berretta
Instituto de Informática – UFG

Prof. MSc. Márcio Giovane Cunha Fernandes
Universidade Estadual de Goiás – CCET

Thiago de Andrade Corrêa

Dedicado a todxs que buscam o conhecimento. Somente em tempos sombrios vemos a fagulha da liberdade que pulsa.

Agradecimentos

Quero agradeçer a todos que participaram direta ou indiretamente nesse trabalho, amigos, familiares, professores e orientadores. Destaque especial para Leidiane Beatriz que me apoiou e me deu suporte durante o trabalho; para professora Deborah Fernandes, minha orientadora, pelo carinho e atenção dispensados e seu companheiro Márcio Giovane pelas aulas; para Luciana, Juliana, Jean, Arthur, Maria Lúcia, Alice, Aline, Vítor e Snow.

Resumo

Correa, Thiago A.. **Uso de ontologias para filtragem de base de dados para melhora de desempenho de classificadores em análise de sentimentos**. Goiania, 2021. 61p. Relatório de Graduação. Instituto de Informática, Universidade Federal de Goiás.

O presente trabalho visou determinar se usando ontologias haveria melhora na precisão de um classificador para inclinação à tristeza no contexto de análise de sentimentos. Concluiu-se que houve melhora significativa quando a ontologia foi empregada na filtragem dos termos usados como entrada para o classificador.

Palavras–chave

Análise de sentimentos, Ciências de dados, Inteligência artificial.

Abstract

Correa, Thiago A.. **Use of ontologies for database filtering to improve classifier performance in sentiment analysis**. Goiania, 2021. 61p. Relatório de Graduação. Instituto de Informática, Universidade Federal de Goiás.

The present work aimed to determine whether using ontologies would improve the accuracy of a classifier for depression and anxiety in the context of the sentiment analysis. It was concluded that the precision improved when ontology was applied to filter the terms used as input for the classifier.

Keywords

Sentiment analysis, Data science, Artificial intelligence

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

1.1 Contextualização

Nunca antes na história da humanidade se compartilhou tanta informação quanto agora, segundo dados do [Stats 2018] o mundo tem cerca de quatro bilhões de usuários ativos na *internet*, o que representa cerca de 59% da população mundial. Essa informação toda disponível pode fornecer importantes dados tanto para empresas quanto para instituições [Liu 2010] e é o que a ciência de dados aborda. Ciência de dados é o estudo dos dados usando estatística para predição e aprendizado de padrões [Dhar 2013], ou seja, usando modelos computacionais em conjunto com dados, minerados ou não, para auxiliar na tomada de decisão baseada nos dados de entrada. A coleta e separação dos dados de entrada é importante para que o modelo seja adequado ao problema a ser resolvido, pois facilita o uso e a seleção dos dados julgados como importantes. Uma técnica para coleta e separação de dados é a mineração de dados.

A Mineração de Dados trata da análise de dados não estruturados para extração e a sumarização de características importantes de forma que seja compreensível e utilizável para humanos e máquinas [Hand e Adams 2014]. Com isso, pode-se coletar dados avulsos da *internet*, estruturar, ou seja, classificar e sumarizar, para simplificar a utilização por máquinas e algoritmos. Porém, ter os dados separados não significa que todos são úteis para uso, há de ter uma limpeza, uma seleção dos dados que serão úteis é necessária. Uma forma de filtrar e selecionar dados, por restrição de contexto e remoção de ambiguidades, é usando ontologias.

Ontologias promovem a restrição dos dados a fim de se evitar ambiguidades e ruídos. Trata-se de modelos para representação do conhecimento, uma forma de agrupar coisas afins, suas entidades, seus relacionamentos, suas tarefas e suas funções, uma forma de modelar o mundo [Mizoguchi 2003]. Existem trabalhos que relacionam a ontologia a uma melhoria em aspectos de visualização e segmentação de dados quando em conjunto com a ciência de dados [Kumar e Joshi 2017, Schouten, Frasincar e Jong 2017, Ali et al. 2019, G.Tau et al. 2004], além das vantagens supracitadas. A filtragem pela ontologia gera uma conjunto de dados correspondente com o contexto de termos e ex-

pressões do domínio. Esse conjunto pode ser utilizado como base de treinamento de uma rede inteligente, com o propósito de identificar padrões e associações entre esses termos e expressões.

Inteligência artificial é a área da computação que estuda o comportamento de entidades inteligentes e tenta replicar esse comportamento computacionalmente [Russell e Norvig 2002]. Esta área engloba diversas abordagens e algoritmos, alguns serão descritos com mais ênfase ao longo desse relatório: redes neurais, *machine learning*, redes bayesianas, árvores de decisão [Russell e Norvig 2002] e processamento de linguagem natural. Essa última propõe o estudo da linguagem falada ou escrita, com a finalidade de extrair conhecimento dessas, para predições, sugestões, extração de opinião, entre outros [Liu 2010]. A opinião é um dos alvos deste trabalho.

Toda opinião expressa pode ser definida como sendo de um emissor para uma entidade, sendo localizada temporalmente, tendo uma avaliação positiva ou negativa. Para coletar opiniões e classificá-las temos uma subárea do processamento de linguagem natural, a análise de sentimentos. Essa tem como objetivo a extração da emoção ou sentimento expresso na forma de texto escrito ou fala, que expressam opinião, resenha, sensação, avaliação em relação a uma entidade, seja ela um objeto, um produto, um serviço ou sistema [Liu 2012, Liu 2010]. Essa opinião pode envolver sentimentos negativos em relação a si mesmo, nesse caso tornando possível a detecção de sinais ou traços de doenças mentais como depressão e ansiedade, por exemplo.

Segundo dados do Ministério da Saúde, a depressão só cresce [OMS], sendo uma das principais doenças mentais no mundo. O mal afeta cerca de 15% da população brasileira [Ministério da Saúde 2014] e traz diversas consequências para o indivíduo, como tristeza comum e cotidiana sem motivo específico, apatia, falta de motivação, falta da vontade de viver. A depressão é o principal motivo de suicídio no mundo todo [OMS]. Mesmo com toda a conectividade que há na atual conjuntura globalizada, esses dados indicam que somente crescem os números dessa doença no mundo inteiro. A depressão, por envolver diversos aspectos de causalidade, acaba envolvendo outras doenças consigo, fenômeno chamado de comorbidade.

Uma comorbidade associada à depressão é o transtorno da ansiedade, caracterizada por preocupação, tensão ou medo exagerado, sensação de alerta constante, entre outros sintomas. Segundo dados da OMS, a ansiedade está em plena ascensão em todo o mundo [OMS], e alguns trabalhos associam o ritmo frenético do mundo atualmente aos sintomas da ansiedade [Méa, Biffe e Ferreira 2016, Bhat e Rather 2012], sendo o estresse o principal fator para tal comorbidade [Saúde].

Em síntese, a motivação desse trabalho reside na necessidade de verificar se o emprego de uma ontologia melhora o desempenho de um classificador para ansiedade e depressão no contexto de análise de sentimentos, para dados minerados de redes sociais.

1.2 Definição do problema

Dado o contexto geral na Seção 1.1, temos que a depressão é a doença que mais cresce no mundo todo, uma comorbidade que traz consigo outros fatores como a ansiedade. Para evitar ou mitigar os sintomas, tais como solidão, falta de apetite, insônia, melancolia, ansiedade exagerada, faz-se necessário intervenção com a finalidade de proporcionar melhora na qualidade de vida dos indivíduos acometidos e redução na taxa de mortalidade associada a fatores psíquicos. Como o fluxo de dados é muito volumoso na internet atualmente, extrair dados de redes sociais é uma saída, uma vez que a abundância e a disponibilidade dessas redes facilita a coleta. Porém, essa vastidão precisa ser filtrada para reduzir o ruído nos dados. A ontologia define o contexto no qual está inserida, reduzindo o escopo do universo de aplicação, fazendo uma filtragem de assuntos relacionados. Processamento de linguagem natural é uma subárea da inteligência artificial, que trata de analisar a linguagem escrita e tirar dela informações importantes, como o sentimento ou opiniões que elas expressam. Contudo, essa filtragem poderia não remover de fato o ruído para necessariamente haver melhora, ou seja, o ponto chave é saber o quão preponderante é essa fase de filtragem por ontologia no desempenho de um classificador supervisionado.

1.3 Objetivo Geral

O objetivo geral desse trabalho é verificar se o uso de ontologia para filtragem dos dados melhora o desempenho de um classificador em análise de sentimentos para inclinação à tristeza de dados minerados da rede social Twitter.

1.4 Objetivos Específicos

- Compreender os aspectos da depressão e ansiedade e como esses se manifestam nos usuários de redes sociais.
- Verificar, testar e ajustar a ontologia desenvolvida em [Rodrigues 2019] para ser utilizada neste trabalho.
- Produzir um modelo de detecção para depressão e ansiedade baseando-se em aprendizado de máquina supervisionado.
- Treinar o classificador e comparar se houve melhoria com e sem o uso de ontologia.

O trabalho segue dividido da seguinte forma: no Capítulo 2 são apresentados alguns trabalhos relacionados ao tema e o que foi produzido na área de ontologia, depressão e análise de sentimentos; no Capítulo 3 a fundamentação teórica; no Capítulo 4 tem-se a apresentação da arquitetura do modelo desenvolvido para este trabalho; no Capítulo 5 tem-se os resultados obtidos e finalmente, no capítulo 6 a conclusão e os trabalhos futuros.

CAPÍTULO 2

Trabalhos Correlatos

2.1 Ontologia

Considerando o contexto descrito no Cap. 1, há alguns trabalhos relacionados que desbravam o mesmo universo de conhecimento. A Tabela 2.1 apresenta alguns desses trabalhos com a motivação que originou a pesquisa.

Trabalhos	Área de estudo	Motivação
[Kumar e Joshi 2017]	Mineração de dados	Satisfação com o governo
[Schouten, Frasincar e Jong 2017]	Análise de sentimentos	Melhoria no desempenho
[Mendonça e Soares 2017]	Ontologias	Novo modelo de construção
[Köhler et al. 2017]	Medicina	Glossário de doenças
[Dragoni, Poria e Cambria 2018]	Análise de sentimentos	Dicionário de termos
[Ali et al. 2019]	Engenharia de trânsito	Smart Cities
[G.Tau et al. 2004]	Sistema sensível ao contexto	Nova abordagem

Tabela 2.1: *Trabalhos sobre Ontologia.*

Os autores [Kumar e Joshi 2017] utilizaram análise de sentimentos para estimar a aprovação do governo para com o povo indiano, detectando a satisfação geral em relação às medidas aprovadas e decisões tomadas, além das críticas aos governantes. O objetivo dessa análise é mostrar-se mais transparente e próximo do povo, promovendo o engajamento em assuntos estatais, principalmente entre os jovens indianos. Para tal, coletaram dados do Twitter e adotaram ontologias para filtrarem e redes bayesianas para o aprendizado de máquina e treinamento do classificador. O modelo proposto pelos autores [Kumar e Joshi 2017] atingiu uma acurácia de 77%.

Em [Schouten, Frasincar e Jong 2017] há a descrição de como análise de sentimentos e ontologias foram utilizadas para alcançar tanto eficiência quanto acurácia na execução da rotina de aprendizado. Como resultado os autores obtiveram evidências que o uso de ontologias melhorou a acurácia do sistema de classificação.

O artigo [Mendonça e Soares 2017] apresenta uma nova metodologia para a criação de ontologias, com etapas de construção mais detalhadas e claras. O argumento

principal é que muitas vezes, quem monta a ontologia não detém vasto conhecimento nessa atividade. Assim, analisaram várias metodologias de criação e reuniram o que julgaram como o melhor de cada, dessa forma, apresentaram uma abordagem didática para construção de ontologias. Os autores concluíram que a metodologia proposta por eles era mais eficiente que as outras abordagens já existentes.

Os trabalhos [Köhler et al. 2017] e [Dragoni, Poria e Cambria 2018] trazem a construção de ontologias para determinados fins. O primeiro trata-se da construção de um ontologia médica transversal entre humanos e alguns animais, a fim de associar doenças afins, seja pelo órgão afetado ou pelo patógeno causador para auxílio ao diagnóstico. A motivação do estudo foi a ampliação desse modelo de representação do conhecimento via ontologia. Já o segundo, trata de uma ontologia voltada para o mapeamento de termos usados em análise de sentimentos. O objetivo do trabalho era coletar dados de artigos sobre análise de sentimentos já publicados, extrair deles termos comuns à área e usar a mesma classificação obtida nas pesquisas, para montar um dicionário. Os autores coletaram palavras primitivas, minerados de artigos publicados na área de análise de sentimentos, associada a classiicações como positivo, negativo ou neutro. Essa classificação foi extraída em conjunto com a mineração em artigos, aproveitando o desenvolvimento já feito para dicionarizá-la. O estudo feito ajudaria outros pesquisadores quando usarem essas palavras, bastaria somente a utilização da classificação pronta.

Com a finalidade de prever congestionamentos, engarrafamentos, aumentar a fluidez do trânsito e evitar acidentes, em [Ali et al. 2019] é adotada a análise de sentimentos em dados coletados de redes sociais para previsões sobre trânsito, para fins de engenharia de transportes e controle de tráfego. Os autores sugeriram uma ontologia para filtragem dos termos coletados e aumento da eficácia do modelo de aprendizado. Como resultado, obtiveram que a abordagem baseada em ontologia foi mais eficiente que o modelo sem a mesma.

Em [G.Tau et al. 2004] é apresentado o desenvolvimento de um *middleware* para auxiliar a comunicação e configuração de sistemas sensíveis ao contexto, voltado para ambientes com Internet das Coisas. Com a popularização dessa área, tornou-se imperativo o desenvolvimento de tais sistemas, uma vez que eles controlam várias variáveis do ambiente comunicando entre si, muitas vezes entre sensores heterogêneos dificultando a tarefa. Para contornar essa heterogeneidade, a abordagem dos pesquisadores sugere uma ontologia que seria capaz de identificar o hardware e utilizar o modelo de comunicaçã específico para cada um desses componentes. Eles concluíram com a implementação do *middleware* e demonstraram que houve melhoria em sistemas sensíveis quando o middleware foi adotado.

2.2 Ansiedade e Depressão

A Tabela 2.2 apresenta uma visão geral dos artigos dessa seção, baseada em dois parâmetros, área de estudo e motivação. O objetivo da tabela é mostrar de maneira resumida o que motivou a pesquisa dos artigos e a área na qual a pesquisa está incluída.

Trabalhos	Área de estudo	Motivação
[Petry 2016]	Programação Ubíqua	Auxílio ao tratamento
[Duque, Raymundo e Neto 2018]	Mineração de dados	Apoio ao diagnóstico
[Dias 2018]	Gamificação	Auxílio ao tratamento
[Cremasco e MN]	Depressão e suicídio	Revisão sistemática
[Moraes e Others 2020]	Processamento de áudio	Apoio ao diagnóstico
[Silveira e Others 2019]	Mineração de dados	Apoio ao diagnóstico
[Dainez e Dainez 2015]	Inteligência artificial	Apoio ao diagnóstico
[Dorneles e Others 2019]	Inteligência artificial	Apoio ao diagnóstico
[Choudhury et al. 2013]	Mineração de dados	Glossário comportamental

Tabela 2.2: *Trabalhos sobre Depressão e Ansiedade.*

Uma rede de monitoramento ubíqua é apresentada por [Petry 2016] para identificação de crises em indivíduos já notoriamente depressivos. O sistema, composto por sensores e monitores ligados ao indivíduo, como medidores de pressão e frequência cardíaca, armazena as variáveis das crises anteriores de forma a comparar a atual situação do indivíduo aos momentos armazenados e contactar os familiares e/ou um médico em caso de reincidência. O estudo conclui que esse modelo ubíquo é eficiente, pois evita a rotina invasiva de testes clínicos, permitindo um monitoramento contínuo para melhor intervenção.

Com o propósito de identificar traços de doenças mentais em textos coletados de redes sociais, os autores [Duque, Raymundo e Neto 2018] coletaram dados do *Twitter* e classificaram manualmente em depressivos e neutros. Após essa classificação, treinaram uma rede bayesiana para aprender a classificação usando ferramentas de processamento em linguagem natural. O objetivo era a construção de um classificador capaz de identificar indícios de depressão em textos coletados da rede social *Twitter*. O artigo conclui com uma acurácia de 75%, sendo considerado satisfatório pelos autores.

Em [Dias 2018] é demonstrada uma aplicação ubíqua de gamificação para o engajamento de pessoas com sintomas de doenças depressivas e de transtorno de ansiedade. O sistema desenvolvido baseia-se em intervenções usando *mindfullness*, uma técnica de concentração usada em meditações e terapia cognitiva. Trata-se de uma forma de terapia na qual o terapeuta e o paciente trabalham juntos para identificação de pensamentos que podem ser nocivos e como lidar com estes nos momentos de crises para aliviar o estresse.

Através da coleta de dados biológicos, a aplicação conseguiu aliviar a maioria das crises, segundo os autores.

O trabalho de [Cremasco e MN] faz uma correlação entre a sintomatologia depressiva e o suicídio em alunos do departamento de psicologia. O objetivo era determinar se havia uma correlação entre casos de depressão e suicídio na graduação superior. Os dados utilizados foram obtidos via entrevista e trouxeram uma revisão sistemática de outras universidades que fizeram o mesmo levantamento.

Para detecção de depressão em áudios, os pesquisadores de [Moraes e Others 2020] descreveram a coleta e estudo de áudios de fala para identificação de pacientes com depressão usando redes neurais profundas. Na conclusão, os autores obtiveram resultados satisfatórios, porém apontaram as limitações do estudo como: base de dados restrita, o fato de ser em língua inglesa e pouco poder computacional para treinamento do classificador.

O trabalho de [Silveira e Others 2019] sugere a construção de um classificador para detecção de depressão em usuários de redes sociais como o *Twitter*. Através da coleta de dados de contas assumidamente depressivas, esses foram utilizados como entrada para ferramentas de aprendizado (SVM - *Support Vector Machine*, KNN - *K-Nearest Neighbors* e *Random Forest*). O estudo alcançou uma acurácia de 60% na predição de depressão.

Com a finalidade de auxiliar o rastreio da qualidade de vida de idosos, [Dainez e Dainez 2015] baseou em instrumentos de classificação existentes como o WHOQOL-bref e GDS-15, e empregou redes neurais para facilitar a rotina de aplicações desses instrumentos de detecção . A metodologia consistiu no emprego desses instrumentos para tentar mapear o nível de felicidade e satisfação geral em relação à vida. Perguntas sobre como o indivíduo se sentiu na última semana, se ele tem se sentindo sozinho, entre outras foram adortadas no estudo. A rede facilitou para que a aplicação desse questionário fosse menos invasiva e eficiente, segundo os autores.

Para detectar traços de depressão em *tweets*, o artigo [Dorneles e Others 2019] apresenta a utilização de máquina de vetores de suporte (SVM) em conjunto com o Inventário de Beck, uma ferramenta de diagnóstico para depressão. Os autores retiraram termos que remetiam às questões do teste para a filtragem dos dados coletados e também aplicaram *stemming* e remoção de *stop words*. Esses dados coletados foram manualmente separados em opinativos e não opinativos para o treinamento do modelo de aprendizado de máquina. O modelo obteve acurácia de 78% para dados com conteúdo depressivo.

Em [Choudhury et al. 2013] foram avaliados perfis em redes sociais de pessoas com diagnóstico clínico de depressão para levantar comportamentos e atitudes que ajudassem a classificar um padrão de comportamento. Entre as características levantadas estão o tempo de conexão, engajamento na rede, estilo linguístico e citação de medicação anti-

depressiva. Os autores ainda usaram essas características como entrada de um algoritmo de aprendizagem supervisionado, alcançando uma acurácia de 70%.

CAPÍTULO 3

Fundamentação teórica

3.1 Depressão e Ansiedade

3.1.1 Depressão

Depressão é uma doença crônica grave [Ministério da Saúde 2014], com mais de 264 milhões de pessoas afetadas no mundo [James et al. 2018]. Ela é caracterizada por tristeza profunda e constante, insônia, falta de energia e de apetite, redução da libido, cansaço e dor sem motivo aparente [Ministério da Saúde 2014], pode ser causada por fatores genéticos, bioquímicos e gatilhos emocionais. Fatores bioquímicos caracterizam-se por alterações de certas substâncias no cérebro, como excesso de cortisol e ausência de dopamina [What Is Depression?]. Os fatores genéticos somente se manifestam depois de um gatilho emocional, como traumas e fortes emoções, estresse físico e psicológico, doenças como hipotireoidismo e síndrome do intestino irritável [Baniasadi et al. 2017]. A sintomatologia padrão [Beck e Alford 2016] é definida, além dos sintomas já mencionados, por alteração de peso; culpa excessiva; dificuldade de concentração; ideias suicidas; baixa autoestima e problemas psicomotores, como agitação ou apatia. Essa doença é classificada em três níveis, leve, moderada e grave.

Segundo o manual de DSM-5, *Diagnostic and Statistical Manual of Mental Disorders* [Association et al. 2014], o diagnóstico da doença é feito de acordo com o histórico do paciente, se o paciente apresentou um sentimento depressivo, perda de interesse ou prazer por atividades cotidianas e tem pelo menos quatro ou cinco dos sintomas citados acima no decorrer dos últimos dias [NIHM 2019].

Segundo dados da OMS, cerca de oitocentas mil pessoas morrem por suicídio todo ano, afetando mais mulheres entre 15 e 29 anos [OMS] sendo a depressão, a maior causa relacionada aos altos índices de mortes. Dados da Sociedade Brasileira de Psiquiatria, num estudo de 2012, verificou que 96,8% dos casos de suicídio estavam relacionados com transtornos mentais e a depressão é o mais comum [Psiquiatria]. Tamanho é o impacto dessas mortes, que o Brasil, desde de 2014, criou a campanha

do Setembro Amarelo para conscientizar a população geral para esse problema e tentar educar a população para identificar e ajudar indivíduos que se encontram nessa condição.

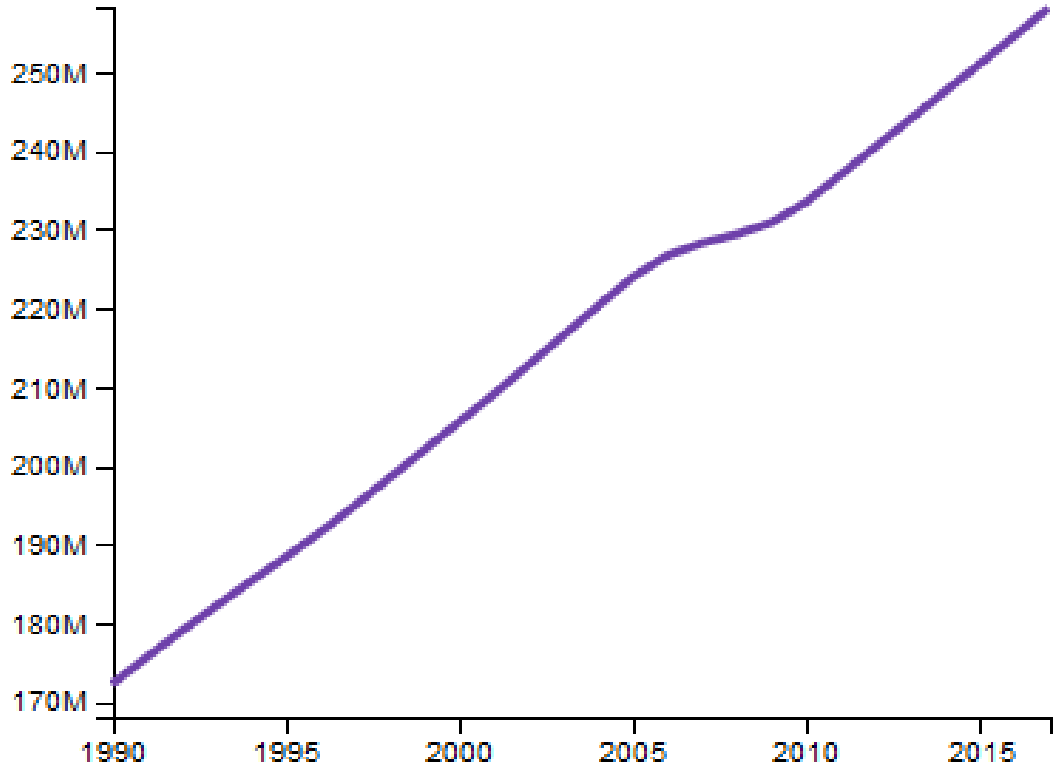

Figura 3.1: *Taxa de incidência de depressão global, ambos os sexos e todas as idades.*

A Figura 3.1 [IHME DATA 2019] mostra o crescimento dos casos globais de depressão ao longo dos anos, para todos os sexos. Apesar da curva ascendente há autores que defendem que muitos casos não entram na contagem oficial, sendo uma das doenças com maior número de subdiagnóstico [Teng, Humes e Demetrio 2005].

Os tratamentos envolvem uso de medicamentos associados a terapias acompanhadas por psicológicos, embora em países emergentes e subdesenvolvidos 76% a 85% dos pacientes não recebam tratamento algum [Wang et al. 2007]. As terapias tentam reestabelecer o paciente em três vias, comportamental, cognitiva-comportamental e interpessoal. Segundo dados do Ministério da Saúde, 90% a 95% dos pacientes mostram remissão total com tratamento antidepressivo [Ministério da Saúde 2014]. Ainda segundo o Ministério, há formas de se evitar depressão através da adoção de um estilo de vida saudável, dieta equilibrada, atividade física regular, evitando o consumo de álcool e drogas ilícitas. Terapias de socialização ou em grupo podem ser usadas em conjunto com outras abordagens e tem apresentado boas taxas de recuperação e menos reincidência em indivíduos que foram submetidos aos tratamentos associando medicamentos e terapias [Cardoso 2017].

Alguns estudos citados na Seção 2 apresentam uma abordagem computacional para o estudo da depressão. Em [Choudhury et al. 2013] os pesquisadores tentaram entender os aspectos do comportamento das pessoas depressivas em redes sociais, tais como engajamento, emoção e estilo linguístico. Os autores definiram engajamento como a taxa de atividade do usuário e o horário no qual ela ocorre. No trabalho ficou evidente que a taxa de atividade era restrita a poucos contatos pessoais e o horário de uso era majoritariamente noturno nos indivíduos com transtorno de depressão. Outro parâmetro observado foi o quão emotivo eram os textos publicados, essa emoção foi classificada em positiva, negativa, ativa ou dominadora, onde ativa é o quão forte são os termos usados e dominadora é grau de controle podendo ser submisso ou dominador. O estilo linguístico também foi observado e foi definido por como o usuário escreve. Verificou-se que pessoas com transtorno depressivo usavam mais palavras possessivas, termos auto depreciativos e negativos além de termos médicos ligado à doença, como o nome de remédios antidepressivos [Choudhury et al. 2013].

3.1.2 Ansiedade

Ansiedade se refere a um grupo de desordens mentais que vem em crises caracterizadas pela sensação de ansiedade, isto é, a sensação incômoda de inquietude e irritabilidade provocada por apreensão constante pelo futuro próximo ou a médio prazo, e medo, incluindo síndrome do pânico, fobias, desordem de ansiedade social, TOC (transtorno obsessivo compulsivo) e desordem por estresse pós-traumático [WHO 2017]. A doença torna-se um empecilho a partir do momento em que prejudica o andamento do dia a dia da pessoa, deixando de fazer tarefas simples do cotidiano por conta da ansiedade. Os sintomas associados incluem, de maneira geral:

- Preocupação, tensão ou medo exagerado;
- Sensação de desastre iminente;
- Medo exagerado de ser humilhado, de algum objeto ou de alguma situação;
- Pavor depois de trauma, leve ou moderado.

O tratamento segue por duas vias, uma medicamentosa, usando ansiolíticos, e uma por psicoterapia, com acompanhamento de psicólogo ou médico psiquiatra, porém, segundo o Ministério da Saúde, o mais adequado é a junção das duas [Saúde]. Geralmente as pessoas com distúrbios de ansiedade possuem certos fatores de risco, exposição ao estresse durante a infância e durante toda a fase adulta, histórico de ansiedade familiar, condições físicas e alimentares, uso de substâncias estimulantes como cafeína ou remédios que tenham o mesmo efeito, podem agravar as ocorrências de crises, o recomendado é evitar esses gatilhos [Health 2016].

Existem vários tipos de males que são englobados sob o mesmo aspecto da ansiedade, assim como referenciado ao início da seção, são exemplos:

- A desordem de ansiedade generalizada é caracterizada pela excessiva ansiedade ou preocupação que perpassa por vários aspectos cotidianos, seja na vida pessoal ou na profissional. É chamada comumente de ansiedade e alguns sintomas são irritabilidade, tensão muscular, dificuldade de concentração e problemas de sono, e sensação de alerta a todo tempo. É tratável com medicação ansiolítica e terapia;
- A síndrome do pânico é definida por acessos intensos de pânico inesperados, disparado por algum gatilho emocional, seja de uma situação ou objeto em específico. As crises tem os seguintes sintomas, tremores nervosos, sensação intensa de desconforto ou perigo iminente, de sufocamento, de palpitação e sudorese excessiva. O paciente desenvolve uma tensão entre as crises, de forma a querer evitar o gatilho dessas, causando uma séries de dificuldades na vida pessoal e profissional do paciente. O tratamento inclui terapia para evitar as crises e medicação para o paciente [NIHM 2019];
- Fobias são medos irracionais por objetos ou situações. Diferenciam-se por pela pontualidade das crises, ou seja, os gatilhos são explícitos e as crises somente ocorrem na presença do disparador. Algumas fobias são medo de altura, medo de aranha, medo de insetos, entre outros [NIHM 2019];
- O transtorno obsessivo compulsivo é caracterizado por recorrentes pensamentos obsessivos e comportamentos repetitivos incontroláveis que trazem uma série de prejuízos ao paciente. Os sintomas incluem germofobia, pensamentos agressivos contra si ou os outros, perfeccionismo exagerado, compulsão por simetria, arranjo e/ou limpeza, alguns pacientes possuem espasmos/rituais involuntários. A terapia associada à medicamentação adequada trazem bons resultados aos pacientes [NIHM 2019];
- Estresse pós traumático é uma desordem que afeta pessoas que vivenciaram algum acontecimento traumático. Geralmente o paciente evita lugares que disparem o gatilho da memória, ou sofre com lembranças recorrentes, tem dificuldade para dormir ou concentrar, sensação de alerta e/ou medo constante e tremores involuntários. A terapia segue por duas vias, uma terapia da exposição, onde o terapeuta confronta o paciente com o trauma vivenciado e com o tempo acredita-se que o evento perca a carga de estresse; ou reconstrução cognitiva, onde o terapeuta remonta os eventos passados, guiando o paciente para entender a situação e controlar as emoções. O tratamento combinado com medicamentos geram bons resultados [NIHM 2019].

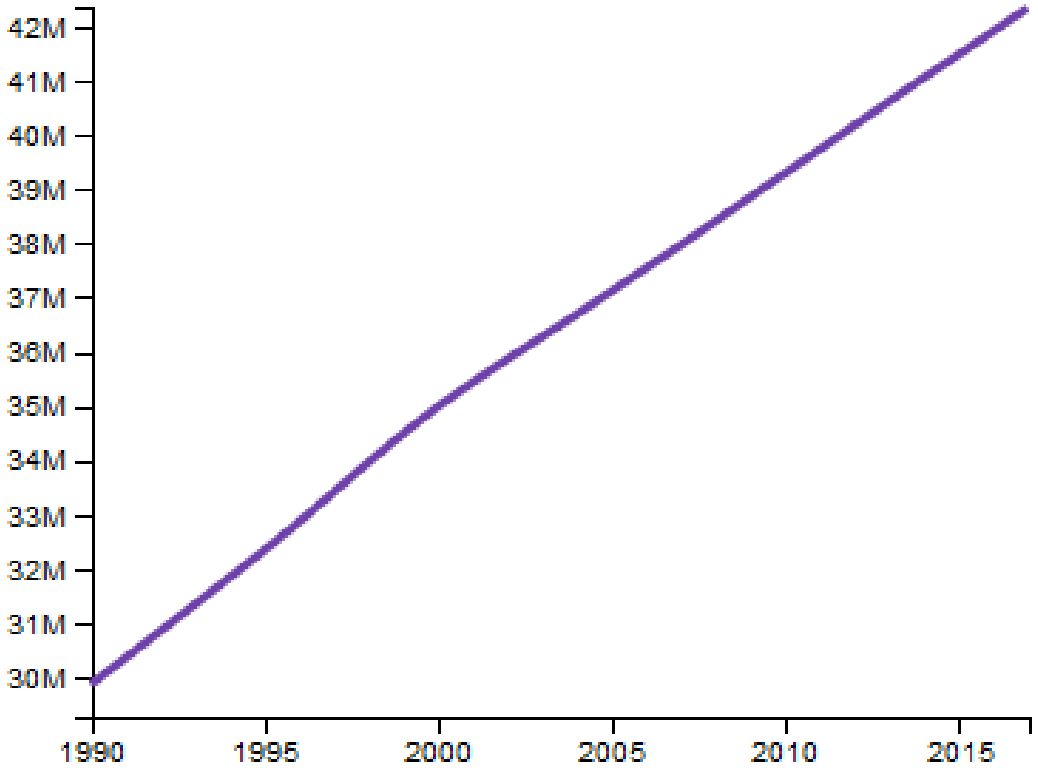

Figura 3.2: *Taxa de incidência de ansiedade global, ambos os sexos e todas as idades.*

A Figura 3.2 [IHME DATA 2019] exibe a taxa de incidência dos casos de ansiedade, considerando ambos os sexos e todas as idades. A curva ascendente no gráfico expõe que o número de casos só tende a aumentar [Souza e Machado-De-Sousa 2017].

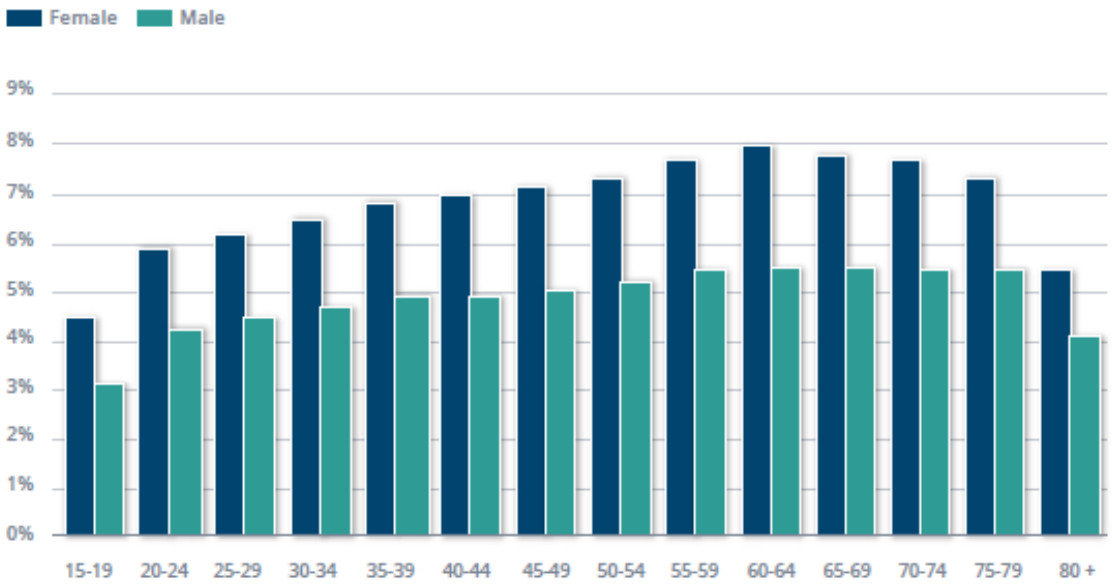

Figura 3.3: *Prevalência global de ansiedade, por sexo, por idade.*

A Figura 3.3 [IHME DATA 2019] mostra os casos de incidência considerando o sexo afetado. A Figura expressa que a taxa de mulheres com transtorno de ansiedade é maior do que para homem em todas as idades consideradas.

3.2 Ontologia

Ontologias são modelos de representação do conhecimento, uma forma de agrupar coisas afins, suas entidades, seus relacionamentos, suas tarefas e suas funções, uma forma de modelar o mundo [Mizoguchi 2003]. Surgiu como um termo nos estudos do grego Aristóteles e designa o estudo (Logos) do ser (Onto), e suas idiossincrasias. O Dicionário Oxford de Filosofia define como derivado que significa ”Ser”, porém passou a representar o ramo da metafísica que estuda aquilo que existe [Blackburn 1997].

Na computação, assume outra definição, um conjunto de conceitos fundamentais e suas relações, que capta como as pessoas entendem o domínio em questão e permite a representação de tal entendimento de maneira formal, compreensível para humanos e computadores [Mizoguchi 2003]. Há utilidade em diversas áreas como inteligência artificial, banco de dados, engenharia de software e a sua função é representar o conhecimento inerente ao domínio de aplicação, mapear entidades e seus relacionamentos e definição de termos e conceitos.

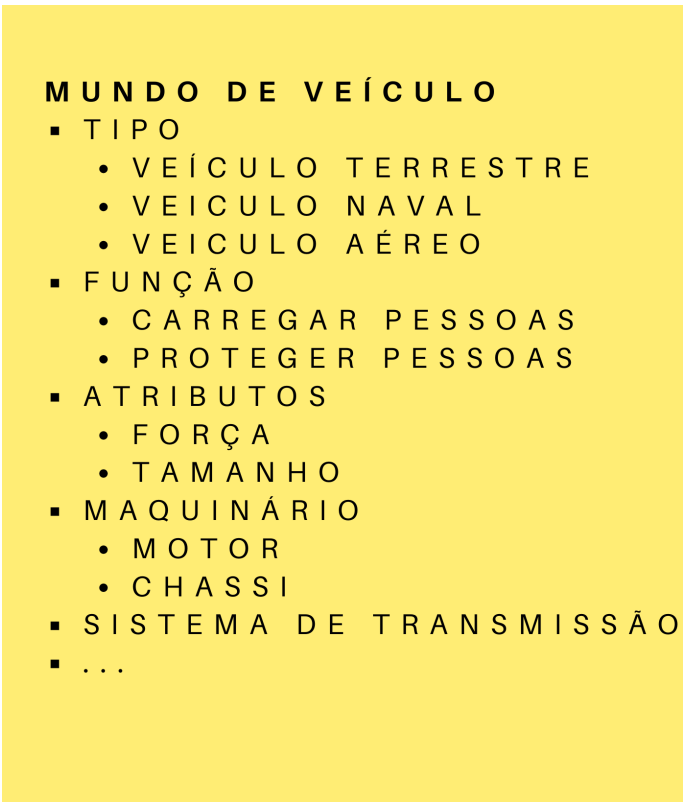

Figura 3.4: *Ontologia de veículo.*

Os componentes básicos de uma ontologia são [Noy e McGUINNESS 2001]:

- Classes, organizadas em taxonomias;
- Relações, interação entre os conceitos de um domínio;
- Axiomas, sentenças sempre verdadeiras;
- Instâncias, para representar os dados em si.

As instâncias devem ser definidas em linguagem formal para evitar ambiguidade na definição. O domínio é o universo no qual a ontologia está inserida. As entidades são os elementos que são mapeados do domínio para a ontologia. Relacionamentos ou relações são as associações possíveis entre essas entidades e axiomas são definições formais em sentenças lógicas que descrevem essa entidade [Morais e Ambrósio 2007].

A Figura 3.4 exemplifica os componentes básicos de uma ontologia. As classes seriam essas taxonomias, tipo, função, atributo; as entidades seriam o objeto real ou observado mapeado para a ontologia; as relações seriam as associações entre entidades, como o pneu que se encaixa na roda, como o eixo transmite energia para as rodas, como o manche regula a força do motor. Os axiomas, nesse contexto, poderiam ser as restriões que se aplicam aos veículos, como ter somente um volante no carro, ou ter duas rodas na motocicleta, ter duas asas em aviões. A instância estaria ligado a como as entidades estão representados, nesse caso, por escrito em extenso e o domínio seria o universo dos veículos, como os aviões, carros, motocicletas, triciclos, entre outros.

Numa crescente demanda por informação, em que ela se apresenta desconexa e difusa, tornou-se imperativo o uso de ontologias para reduzir o contexto da informação, reduzindo as ambiguidades por limitar o contexto no qual está sendo aplicada. Há algumas definições propostas por diferentes autores da área da computação, algumas são:

> *Uma especificação de um vocabulário representacional para um domínio compartilhado do discurso - definições de classes, relações, funções e outros objetos - é chamada de ontologia [Gruber 1993].*

> *Ontologias são descrições formais de conhecimento compartilhado em um domínio. Uma ontologia pode ser usada como uma especificação de um sistema baseado em conhecimento, porque especifica o necessário para as tarefas que o sistema deve executar [Borst e Borst 1997].*

Ambos os autores utilizam um termo oriundo da inteligência artificial, a conceitualização, que é, segundo [Genesereth e Nilsson 1987], um conjunto de objetos que o observador acredita que exista no universo de interesse e suas relações, logo é composta por entidades e relacionamentos que foram mapeados para uma abstração desse universo. Tal conceitualização é definida formalmente por $C = (D, W, R)$ em [Guarino, Oberle e Staab 2009], na qual:

- C é conceitualização, um recorte do mundo real que se deseja representar, o universo de interesse.
- D é o domínio no qual a ontologia está inserida, o universo de interesse.
- W é o conjunto de possibilidades de arranjo entre as entidades da ontologia e seus relacionamentos.
- D é o conjunto das relações dentro do espaço de domínio (D,W), dos relacionamentos entre as entidades.

Usando o exemplo adotado na Figura 3.4, a ontologia de veículos, a conceitualização é a abstração do universo composto por todos os veículos, todas as entidades que os compõem e todos os relacionamentos entre essas entidades. O domínio seria esse universo de veículos, o conjunto de possibilidades seriam as possíveis formas de montar um veículo e o conjunto de relações seria a forma como essas entidades se relacionam para montar o veículo.

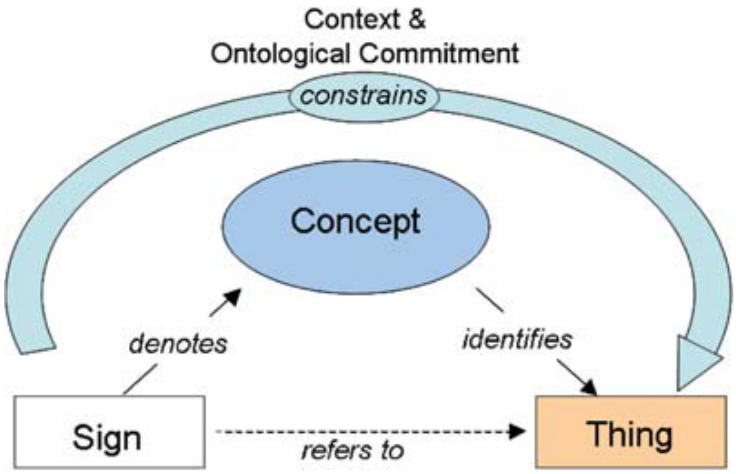

Figura 3.5: *Contexto ontológico.*

A Figura 3.5 [Guarino, Oberle e Staab 2009] exemplifica o que é o contexto ontológico. São as restrições formais impostas para compactar o mundo de possibilidades a fim de limitar as interpretações, reduzindo ambiguidades e facilitando o processo comunicativo. Por exemplo, usando uma analogia, duas tribos A e B decidem previamente combinar um sinal de fumaça de indicaria a presença de algum predador. O objeto, referenciado como coisa na figura, seria o predador, o objeto que dispara um sinal. O sinal combinado é transmitido após a visualização do objeto. O conceito é o perigo iminente que o predador representa. Caso o sinal não fosse previamente especificado ou se o sinal não fosse explícito, a mensagem poderia não ser entendida por uma das partes, dificultando o processo comunicativo e lesando uma das partes. Ao ver a fumaça levantada pela tribo A, a tribo B se protege para evitar o ataque, assim sendo a fumaça é um sinal que denota um conceito (presença do predador) que identifica um objeto (predador), restrito a

um contexto comum às ambas tribos, de forma a que não haja dúvida que o sinal enviado seja unicamente pela presença do predador.

O compromisso, descrito na Figura 3.5 [Guarino, Oberle e Staab 2009], caracteriza-se pela formalização do contexto no qual a ontologia estará incluída, restringindo a amplitude e a transversalidade do universo do discurso e reduzindo ambiguidades. O vocabulário seria o conjunto de constantes e símbolos de predicados, como por exemplo uma fórmula lógica $X(A,B)$, em que A é x em relação a B e pertence a uma linguagem L. O compromisso ontológico é o mapeamento dos predicados e constantes do vocabulário para o domínio da ontologia ou para o conjunto das relações presentes na ontologia, assim tem-se que para todo elemento dentro do conjunto vocabulário há um elemento associado no conjunto $D \cup R$.

A Figura 3.5 [Guarino, Oberle e Staab 2009] descreve o contexto ontológico definido em uma linguagem L lógica de primeira ordem, com um vocabuário V apropriado e uma conceitualização $C = (D,W,R)$, assim tem-se que é definido por $K = (C,I)$, no qual:

- K é o compromisso ontológico.
- C é a conceitualização, o universo de interesse.
- I é uma função de intensionalidade $I : V \rightarrow D \cup R$.

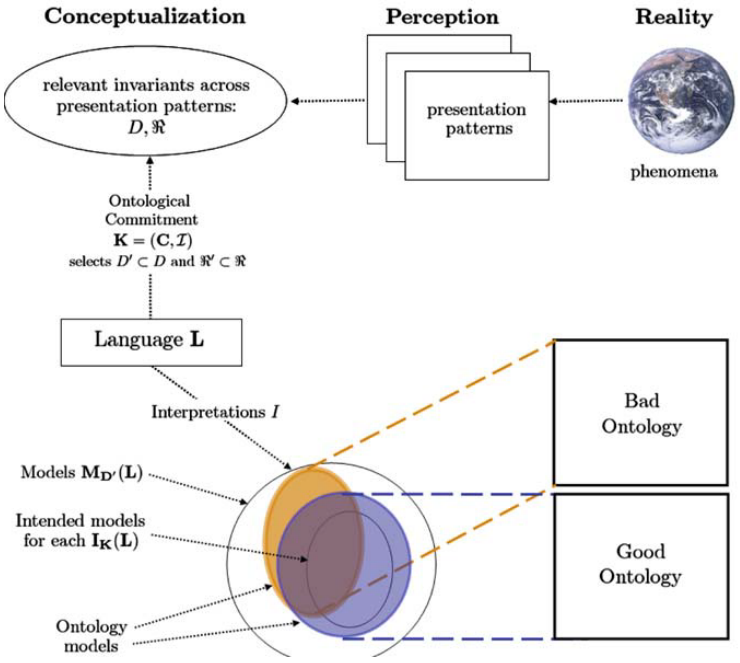

Figura 3.6: *Exemplo de ontologia.*

Na Imagem 3.6 [Guarino, Oberle e Staab 2009] há um exemplo de ontologia, onde observa-se a realidade, obtém-se dela uma observação, monta-se a conceitualização, que é um recorte dessa realidade, um microuniverso de interesse, e usando

a linguagem L monta-se uma ontologia, que é um modelo, uma representação formal dessa conceitualização do universo de interesse. A boa ontologia descreve de forma mais geral o modelo proposto, sem aproximar-se de interpretações ou inferências [Guarino, Oberle e Staab 2009].

Pela definição formal de ontologia, tem-se que uma ontologia O qualquer, para uma conceitualização C, um vocabulário V e um compromisso ontológico K é uma teoria lógica, consistindo em um conjunto de sentenças e axiomas em linguagem L, ou seja, uma ontologia é formada por uma abstração de interesse do mundo real, descrita usando uma linguagem formal lógica de primeira ordem, possuindo um vocabulário consistindo em constantes e predicados, fundamentado por um compromisso que determina que todo elemento do vocabulário tenha um elemento associado no conjunto de domínio ou no das relações da ontologia, sendo assim, uma teoria formal e válida [Guarino, Oberle e Staab 2009].

Algumas áreas possíveis de aplicação de ontologias são recuperação de dados, processamento de linguagem natural, gestão do conhecimento [Morais e Ambrósio 2007], é considerada estrutural em aplicações baseadas em conhecimento, como sistemas de representação do conhecimento, integração de dados e aplicações baseadas em *web semântica* [Kharbat e El-Ghalayini 2008]. Outras aplicações estão descritas na seção 2, são exemplos: engenharia de trânsito [Ali et al. 2019], medicina e apoio ao diagnóstico [Köhler et al. 2017], análise de sentimentos [Dragoni, Poria e Cambria 2018, Schouten, Frasincar e Jong 2017], mais exemplos estão na arte [Beloozerov et al. 2019], segurança da informação [Ding, Wu e Zhang 2019], *smart houses* [Ning et al. 2019], programação [Zhong et al. 2019] e robótica [Azevedo, Belo e Romero 2019, Asprino et al. 2019].

Alguns autores propuseram a tipificação das ontologias, [Uschold e Gruninger 1996] sugeriu a separação das ontologias levando em conta o nível de formalismo da linguagem L adotada, indo desde altamente informais, expressa em linguagem natural; passando por semi-informais, pouco uso de formalidade; semiformais, algum rigor formal e fortemente formais, onde os termos são definidos logicamente com teoremas e provas. Outro tipo de ontologia foi descrito por [Guarino 1998], cuja classificação leva em conta a finalidade para a qual a ontologia é construída, tendo algumas subclasses como:

- Ontologias genéricas: representação geral de uma ontologia, construção de teorias abstratas do mundo;
- Ontologias de domínio: conceitos específicos de um domínio, por exemplo, uma ontologia de veículos;
- Ontologias de tarefas: descrição específica de atividades, processos ou diagnósticos, contém o passo a passo para a realização dessas;

- Ontologias de representação: usada para formalizar o conhecimento de um sistema, explicam as conceitualizações por trás dos formalismos de representação.

3.3 Análise de sentimentos

Opiniões são centrais em diversas áreas, é comum ver avaliações de outros usuários quando se quer baixar um aplicativo novo, ou ver a avaliação de um restaurante antes de reservar uma mesa, empresas usam de opiniões para melhoria ou descontinuidade de um produto ou serviço. Opiniões, sentimentos, avaliações, atitudes e emoções são a base do estudo da análise de sentimento [Liu 2012]. O crescimento da demanda pela análise de sentimentos cresceu juntamente com a ascensão das redes sociais na *Internet* e o crescente volume de opiniões, comentários, avaliações que são postadas a todo momento na *Web* [Liu 2012].

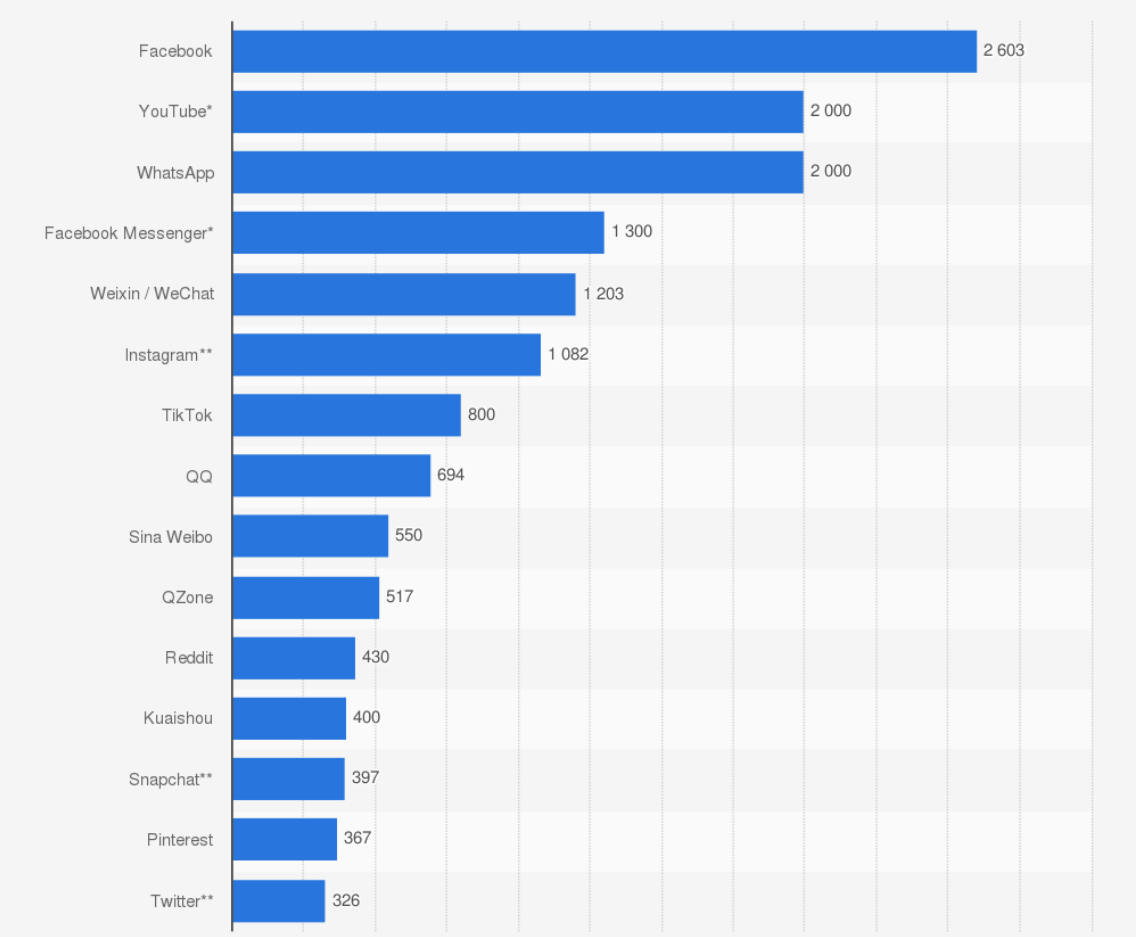

Figura 3.7: *Número de usuários por redes sociais em milhões.*

A Figura 3.7 [Kemp] mostra os usuários totais no mês de julho em 2020. Usuários ativos foram aqueles que acessaram pelo menos uma vez alguma dessas redes no mês de referência. Tornou-se interessante separar e analisar todo esse fluxo para tirar

informações úteis para a sociedade e as organizações. Na seção 2 é apresentado um trabalho que utiliza de dados coletados de redes sociais para estimar a satisfação popular ante ao governo indiano.

A análise de sentimentos baseia-se na classificação de dados de interesse em relação a um objeto, seja ele um produto, uma pessoa, uma organização, serviços, eventos, entre outros [Liu 2012], exigindo o uso de sistemas automatizados, uma vez que seria difícil a um leitor humano procurar, ler, separar, classificar e organizar de forma usável todo o texto dessas redes sociais e da *Web*.

Para facilitar o estudo da opinião, é comum defini-la com um conjunto quíntuplo formado por $O(E,A,S,h,t)$, em que:

- O é a opinião de interesse.
- E é a entidade, objeto da opinião.
- A é o aspecto alvo da opinião, pertence a E.
- S é a sentimento em relação ao aspecto analisado.
- h é a quem emitiu a opinião.
- t é o tempo pontual em que a opinião foi emitida.

Por exemplo, João comeu em um novo restaurante X que abriu no bairro, depois de terminar ele emitiu uma opinião em relação à qualidade do almoço, classificando como bom em 2019. A opinião será formalmente descrita como:

$$O(almocorestaurante_x, qualidadealmoco, BOM, Joao, 2019)$$

A classificação é um dos tópicos mais estudados dentro da área de análise de sentimentos [Liu 2010]. Diversos autores publicaram estudos sobre como classificar um texto em relação a um objeto em positivo, quando a opinião é boa em relação à entidade ou negativo, quando a opinião é ruim em relação à entidade.

Os classificadores operam em três níveis, de acordo com a granularidade [Liu 2010], a saber:

- Nível de documento: os artigos [Turney 2002],[Pang, Lee e Vaithyanathan 2002] discorrem acerca da opinião emitida em relação a uma entidade, considerando a opinião geral de um documento como um todo, podendo ser positiva ou negativa.
- Nível de sentenças: a análise opera sentença por sentença e tenta classificar essas em positivo, negativo e neutro, esse último é quando uma opinião não é encontrada no texto analisado. Como há essa separação de opinião neutra, esse nível se aproxima bastante da classificação de subjetividade [Wiebe, Bruce e OHara 1999].
- Nível de aspectos: aqui a análise é somente na opinião, identificando o alvo e o sentimento apenas e ignorando o resto, sendo considerada a análise mais refinada de sentimentos [Liu 2012].

3.3.1 Classificação em nível de documento

O classificador percorre todo o texto avaliando o conteúdo e emitindo a opinião geral do texto em relação a uma entidade. Essa análise pode ser feita de algumas maneiras: o trabalho [Pang, Lee e Vaithyanathan 2002] sugere o emprego de aprendizado de máquina supervisionado, usando ferramentas como *Naïve Bayes*, máquinas de vetores de suporte e classificação de entropia máxima; em [Turney 2002] é proposto o emprego de aprendizado de máquina não supervisionado. Esse nível de análise está muito restrito ao domínio e à linguagem, ou seja, é complicado classificar quando um documento apresenta opiniões acerca de diferentes áreas, como cinema e automobilismo, por exemplo, pois não se consegue separar os domínios, a classificação ocorrerá no documento com um todo [Liu 2010].

Aprendizado supervisionado

No aprendizado supervisionado, à máquina é apresentado um conjunto selecionado de dados que vai mostrá-la como se comporta o objeto em análise. No caso da análise de sentimentos, o conjunto de aprendizado é formado por opiniões coletadas já pré avaliadas, sendo notório se são positivas ou não. Dessa maneira, o computador estabelece um padrão entre as avaliações positivas e as negativas [Coppin 2004]. Após essa etapa inicial, é apresentado à máquina um conjunto de teste, formado por opiniões mistas, uma parte positiva, outra negativa, e avalia-se a convergência das respostas, quanto maior o número de acertos, maior a acurácia do modelo de aprendizado [Pang, Lee e Vaithyanathan 2002]. Análise de sentimentos é um problema de classificação de texto, logo, empregam-se as mesmas ferramentas como *Naïve Bayes*, máquinas de vetor de suporte e classificação de entropia máxima [Liu 2012]. A análise do texto é feita levando em consideração alguns recursos textuais como:

- Termos e frequências: as palavras são processadas e as frequências dessas podem definir o contexto geral do texto. Palavras repetidas com valor positivo tendem a tornar o aspecto geral do documento mais positivo, e vice versa.
- *Part of Speech*: as palavras são computadas de acordo com o valor morfológico do termo, assim adjetivos possuem maior peso na análise, pois são considerados importantes quando se emite uma opinião, por exemplo. [Liu 2012].
- Palavras de sentimentos e frases: são palavras que indicam uma tendência positiva ou negativa, por exemplo, advérbios como muito, bom, ou adjetivos como fantástico, maravilhoso, ou substantivos como porcaria, lixo, ou até verbos como odiar, amar, por exemplo.
- Trocadores de sentimento: são palavras que alteram toda a classificação do texto, invertendo positivo em negativo e vice versa. O não é um termo trocador.

Essas partes possuem pesos que somados, demonstram a inclinação do texto como um todo, sendo positivo ou negativo, dependendo do objeto de análise escolhido.

Aprendizado não supervisionado

No aprendizado não supervisionado, o computador consegue agrupar elementos relacionados sem que haja treinamento prévio. O uso dessa ferramenta ocorre quando não se sabe um critério de seleção ou quando se quer separar em *clusteres* um conjunto de dados [Coppin 2004]. As redes neurais são um exemplo de aprendizado não supervisionado, nas quais são carregados dados (conjunto) e ela os classifica, usando uma estrutura desenvolvida baseada no cérebro humano e neurônios. Esses neurônios preveem elementos afins de acordo com teorias probabilísticas, como regressão logística [Coppin 2004]. Na análise de sentimentos, os trabalhos [Turney 2001] e [Turney 2002] exploram essa abordagem usando aprendizado não supervisionado para classificar trechos formados por substantivos e adjetivos em positivos ou negativos. O algoritmo funciona coletando tais trechos e comparando, numa rede neural, a semelhança desses com as palavras "excelente"e "pobre"(*poor* no texto original) [Turney 2002] e está dividido em três partes:

- Primeira fase: O algoritmo extrai os termos da oração e anota a classificação sintática, sendo o termo um adjetivo, ou advérbio. A análise leva em consideração uma tabela que organiza o peso dos termos em conjunto, de acordo com o tag de *Part of Speech*. Numa frase composta por um adjetivo seguido de um advérbio, com um substantivo depois, é isolado os termos que denotam opinião, adjetivo e advérbio.
- Segunda fase: ocorre o cálculo da orientação do sentimento, uma medida de dependência entre dois termos, para tanto ele obtém a razão entre a co-ocorrência dos termos associados e a co-ocorrência dos termos caso eles fossem independentes. O resultado é a subtração da medida de dependência com o termo "excelente"e o termo "pobre".
- Terceira fase: é feito o somatório dos resultados de cada sentença para montar o contexto sentimental geral do texto, caso os termos estejam inclinados para a positividade do "excelente", o texto vai ser considerado como positivo durante a classificação, ou negativo, caso tenha um valor que aponte para tal.

3.3.2 Análise em nível de sentença

O foco da análise nesse nível não é um documento como um todo, mas em sentenças menores. Segundo [Liu 2010], esse nível pode ser considerado como uma extensão do anterior, mas em escala reduzida, uma vez que sentenças podem ser entendidas como documentos menores, a diferença é a quantidade de opiniões, maior no documento. É

comum dividir a análise em duas partes, o primeira parte é classificar se uma sentença contém ou não uma opinião, já a segunda é classificar se essa opinião é positiva ou negativa. Uma das premissas nessa etapa é que uma sentença é a opinião emitida por um único emissor em relação a um produto, ou objeto.

Classificação de subjetividade

A etapa inicial, citada acima, é também chamada de classificação de subjetividade e decide se uma sentença é objetiva, contém fatos ou expressões não emocionais, ou subjetiva, contém opinião explícita. A maioria dos algoritmos de classificação de subjetividade usa aprendizado de máquina supervisionado [Liu 2010]. No artigo [Wiebe, Bruce e OHara 1999], os autores propuseram o uso de *naïve bayes* para classificação binária dos termos da oração e o quanto elas podem expressar opiniões. O mesmo autor [Wiebe et al. 2000] sugere usar aprendizado não supervisionado para essa etapa, o trabalho consiste em identificar subjetividade em sentenças para classificá-las em termos de opinião, para tanto usaram similaridade distribucional, para achar palavras similares e o sentimento que elas representam [Liu 2010], baseando-se num corpo léxico para filtragem da "gradabilidade"(*gradability* no texto original), que são palavras modificadores de intensidade, adjetivos que aumentam ou diminuem o teor significativo da oração. Esse corpo léxico foi apresentado em [Hatzivassiloglou e Wiebe 2000] e é composto por advérbios e frases nominais montado manualmente. O artigo [Moraes et al. 2016] propôs a criação de um corpo léxico em português para a classificação de subjetividade, o *Computer-BR Corpus*, onde trechos de textos retirados de redes sociais foram manualmente mapeados em ironia, negativo, neutro e positivo.

Classificação de sentimentos em sentenças

Na segunda etapa o objetivo é classificar qual polaridade o trecho subjetivo tem, sendo negativa, neutra ou positiva. Uma premissa que há nessa etapa é a de que uma sentença subjetiva é composta por uma opinião pontual de uma entidade para uma outra, sendo um emissor e o outro o objeto em questão. Os modelos que tentaram esclarecer esse etapa são de aprendizado supervisionado ou baseado em corpus léxicos. Um exemplo de algoritmo que usa aprendizado supervisionado é o [Kim e Hovy 2004], que usa o somatório dos pesos dos termos para a classificação geral da sentença, o aprendizado foi para identificar outras formas específicas de opinião. Em [Hu e Liu 2004], os autores se basearam em *WordNet*, um corpo léxico de apoio composto por antônimos e sinônimos e suas respectivas classificações. O objetivo do artigo é análise em nível de aspecto, mas também pode ser usado em nível de sentença. Novamente a polaridade da sentença vai ser a soma dos pesos intermediários colhidos em comparação com o corpus *WordNet*.

Nome	Descrição	NS	S
Emoticons	Possui uma lista de emoticons dividida em positivos(":)") e negativos(":("). O texto é classificado de acordo com a classe que tiver mais emoticons. Apesar de possuir uma alta taxa de acertos este método depende muito da presença do emoticon no texto.	✓	
Opinion Lexicon [Hu and Liu, 2004]	Também conhecido como Sentiment Lexicon, consiste de uma lista com cerca de 6.800 palavras rotuladas como positivas e 6.800 palavras rotuladas como negativas, incluindo gírias e abreviações no idioma Inglês. Este é um método léxico criado a partir de textos coletados em reviews de produtos em sites de compra.	✓	
Opinion Finder (MPQA) [Wilson et al., 2005a] [Wilson et al., 2005b]	É uma ferramenta considerada híbrida pois utiliza um léxico de sentimentos mas utiliza Naive Bayes para distinguir se uma sentença é subjetiva ou objetiva.	✓	✓
Happiness Index [Dodds and Danforth, 2009]	É uma escala de sentimentos que utiliza o popular ANEW (um conjunto de palavras ligadas a emoções do Inglês). Este método foi contruido para avaliar textos entre 1 a 0, indicando a quantidade de felicidade existente. Em particular os autores utilizaram este método para mostrar que a "quantidade de felicidade" nas letras das músicas diminui entre 1961 e 2007.	✓	
SentiWordNet [Esuli and Sebastiani, 2006] [Baccianella et al., 2010]	É um léxico contruído a partir de outro léxico já conhecido chamado WordNet [Miller, 1995]. No WordNet os autores agruparam adjetivos, substantivos, verbos em conjuntos de palavras que fossem similares formando uma rede de palavras. Já os autores do SentiWordNet associaram uma polaridade entre algumas palavras-semestes do WordNet e propagaram essa polaridade nas palavras similares da WordNet criando um amplo léxico de sentimentos.	✓	✓
LIWC [Tausczik and Pennebaker, 2010]	O LIWC é uma ferramenta bem estabelecida e utilizada em diversas áreas, e contou com o aval de psicólogos, sociólogos e linguistas durante seu desenvolvimento. Ela possui um dicionário léxico de aproximadamente 4500 palavras e raízes de palavras, fazendo parte de oitenta categorias das mais variadas(ansiedade, saúde, lazer, etc).	✓	
SenticNet [Cambria et al., 2010]	SenticNet é um dicionário semântico e afetivo para opinião em nível de conceito e análise de sentimento. Ele foi construído através do que é denominado pelos autores de sentic computing, um pardigma que explora Inteligência Artificial e técnicas de Web semântica para processar opiniões via mineração de grafos e redução de dimensionalidade. Ele é público e provê um bom material para mineração de opniões em nível semântico e não apenas sintático.	✓	
AFINN [Nielsen, 2011b]	É um léxico contruído a partir do ANEW mas com o foco em redes sociais, contendo gírias e acrônimos e palavrões da lingua Inglesa. Ele possui uma lista de 2.477 termos classificados entre -5(mais negativo) e +5(mais positivo).	✓	
SO-CAL [Taboada et al., 2011]	É um método léxico que leva em conta a orientação semântica das palavras(SO). Criado contendo unigramas (verbos, advérbios, substanticos e adjetivos) e multi-gramas (intensificadores e frases) numa escala entre + 5 e -5. Os autores também incluíram analisador de partes do discurso e negação.	✓	
Emoticons DS (Distant Supervision)[Hannak et al., 2012]	É um léxico que possui termos gerados a partir de uma extensa base de dados do Twitter. Estes termos foram classificados automáticamente baseando-se na frequência de emoticons positivos ou negativos nas sentenças.	✓	

Figura 3.8: *Trecho de tabela comparativa entre os algoritmos de classificação.*

O Trabalho [Benevenuto, Ribeiro e Araújo 2015] apresenta a Tabela 3.8 comparativa entre os métodos de classificação, seja supervisionados ou não, em nível de sentença.

3.3.3 Análise em nível de aspecto

Uma vez que a análise em nível de documento pode ser muito geral, omitindo opiniões específicas em prol do contexto geral do documento e a análise em nível de sentença pode também não ser suficiente para identificar o objeto da opinião, a análise em nível de aspecto tenta reconhecer todas as opiniões definidas em 3.3 em um dado documento [Liu 2012]. Essa análise é basicamente dividida em duas etapas, extração de aspecto e classificação de sentimento em nível de aspecto.

Extração do aspecto

A extração do aspecto envolve uma premissa que já foi abordada anteriormente, a opinião sempre tem um alvo, sendo assim tem-se quatro formas de fazê-lo:

- Baseado em frequência de nomes: minerando opiniões acerca de um objeto comum, percebeu-se que havia uma convergência na linguagem utilizada. Assim

determina-se o principal aspecto pela maior frequência. Nomes menos frequentes são menos relevantes e não precisam ser levados em consideração, é descrito em [Hu e Liu 2004].

- Baseado em proximidade de aspecto: estudos sugerem que o aspecto principal está sempre próximo ao adjetivo ou palavra que expressa alguma opinião e sentimento, é descrita em [Hu e Liu 2004].
- Aprendizado supervisionado: extração de aspecto pode ser considerado como um caso de extração de informações. Como precisa de dados de treinamento manualmente rotulados, é dependente de um corpo léxico previamente existente. Os métodos mais usados são HMM (*Hidden Markov Models*) [Rabiner 1989] e CRF (*Conditional Random Fields*) [Lafferty, McCallum e Pereira 2001].
- Modelos de tópicos: método não supervisionado que usa dados estatísticos para construir clusteres de palavras, chamados de tópicos. As palavras são escolhidas pela distribuição estatística dessas no documento com um todo. Os mais usados são pLSA (*Probabilistic Latent Semantic Analysis*) [Hofmann 1999] e LDA (*Latent Dirichlet allocation*) [Griffiths e Steyvers 2003].

Classificação do sentimento no aspecto

Nessa etapa ocorre a determinação do sentimento em relação ao aspecto selecionado, ocorre a identificação do objeto alvo da análise e do emissor, assim como o sentimento expressado. Para tal extração há duas abordagens, a baseada em aprendizado supervisionado e baseado em corpo léxico. A abordagem baseada em aprendizado supervisionado que classificam em nível de sentença também podem ser utilizadas para nível de aspecto. A utilização de analisadores de dependência (*dependency parsers* no original) é comum [Liu 2012] para o reconhecimento do aspecto principal na sentença. Como em [Jiang et al. 2011], onde um analisador de dependência foi usado para gerar um conjunto de recursos dependentes de aspecto para classificação. Uma das limitações dessa abordagem é a dependência em relação aos dados de entrada, sendo muito restrito ao domínio no qual está inserido, além de não escalar bem quando apresentado a muitos domínios [Liu 2012]. A abordagem léxica tenta resolver esse problema, escalando bem e não sendo tão restrita ao domínio de aplicação [Liu 2012]. O trabalho apresentado em [Ding, Liu e Yu 2008] fornece um método em quatro fases para essa abordagem.

- Marcação das palavras de sentimento e frases: Marca-se todas as palavras que contém ou expressam emoção o frases que contenham mais de um aspecto emocional, somando um a cada palavra dita como positiva e subtraindo um a cada palavra negativa.
- Aplicação de modificadores de sentimentos: Marcação das negações ou palavras que modificam o sentido da emoção.

- Orações adversativas: conjunções adversativas alteram o sentido da frase, palavras como "mas"introduzem o efeito oposto ao falado posteriormente na frase. Por exemplo, na frase "Fui mal atendido, mas gostei do prato principal"pode parecer uma opinião ruim, porém a conjunção muda esse aspecto.
- Agregação dos aspectos: somatório das atribuições de valor das etapas anteriores. Caso a soma seja positiva, a oração é dada como positiva, analogamente é negativa caso a soma seja negativa.

Em [Taboada et al. 2011], os autores comparam as duas abordagens e citam vantagens e desavantagens de cada uma, tendo ligeira preferência para a léxica.

CAPÍTULO 4

Descrição do Experimento

4.1 Modelo da ontologia

A base de dados utilizada nesse trabalho foi coletada por [Rodrigues 2019], onde os autores montaram uma ontologia de domínio para a depressão e ansiedade. Essa base de dados continha 117604 linhas, cuja coleta foi realizada no dia 12 de abril de 2019, com a ajuda do Twittology, um *crawler* de busca e armazenamento de dados públicos da plataforma *Twitter*. Os dados incluíam *tweets*, nome de usuário, data da coleta, linguagem utilizada, entre outras informações.
O Trabalho [Rodrigues 2019] consistiu em mapear palavras relacionadas ao universo da ansiedade e depressão utilizando questionários e materiais da área da psicologia com a ajuda de um especialista. Com a obtenção dessas palavras, os autores montaram um conjunto de regras ontológicas que determinariam o nível de classificação definido pela Tabela 4.1.

As regras propostas por [Rodrigues 2019] foram:

- Leve: Mais de um termo Leve e nem termo Grave ou Moderado;
- Moderado: Mais de um termo Grave e mais de três termos Moderado e menos de dois termos Leve ou Um termo Grave e dois termos Moderado e mais de quatro termos Leve;
- Grave: Mais de um termo Grave e mais de cinco termos Leve.

Esse presente papel visou dar continuidade ao [Rodrigues 2019], testando e adequando as regras proposta pelos autores.

Grave	Moderado	Leve
Sofrer	Não gosta de mim	Estresse com pessoas
Suicídio	Vida	Tempo desperdiçado
Auto-estima baixa	Sem amor próprio	Se dar valor
Sem vontade de viver	Estou triste	
Pensar em suicídio	Insegura	
Vacilar comigo	Depressiva	
Tudo errado	Vivendo triste	
Cansada de existir	Solidão	
Frustrado		
Quero morrer		
Vida sem sentido		
Dificuldade de suportar		
Sentindo mal		
Depressão frescura		
Ansiedade		
Me cortar		

Tabela 4.1: *Termos e a classificação sugerida em [Rodrigues 2019].*

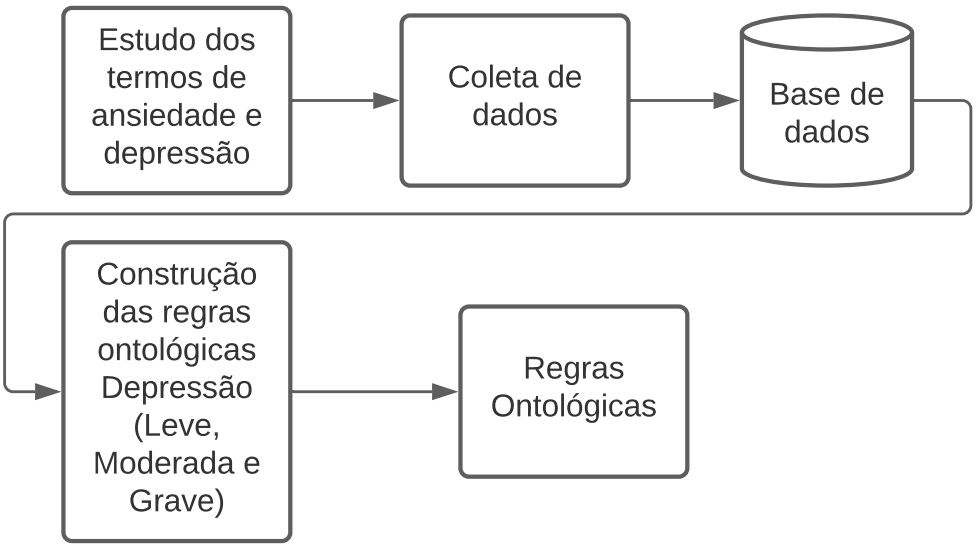

Figura 4.1: *Definição das etapas realizadas em [Rodrigues 2019].*

A Figura 4.1 define as etapas realizadas naquele trabalho, desde a coleta até a consolidação das regras ontológicas, usadas nesse trabalho.

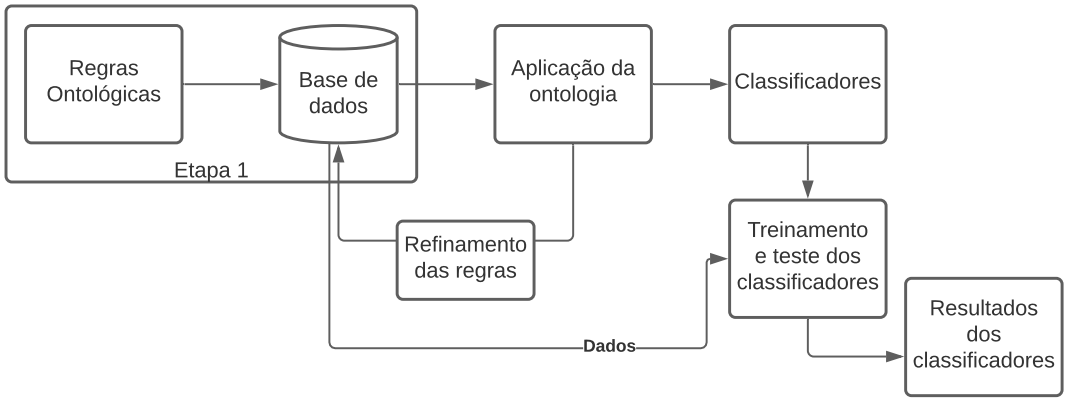

Figura 4.2: *Proposta desse trabalho a partir do trabalho feito em [Rodrigues 2019].*

A Figura 4.2 esclarece a continuidade, demostrando o escopo da etapa 1, feita pelo trabalho [Rodrigues 2019] e esse trabalho.

O teste das regras foi feito via *Notebook Jupyter* em linguagem *Python* usando a base de dados já mencionada. O resultado do teste foi que os *tweets* apresentavam somente um termo da categoria à qual pertencia.

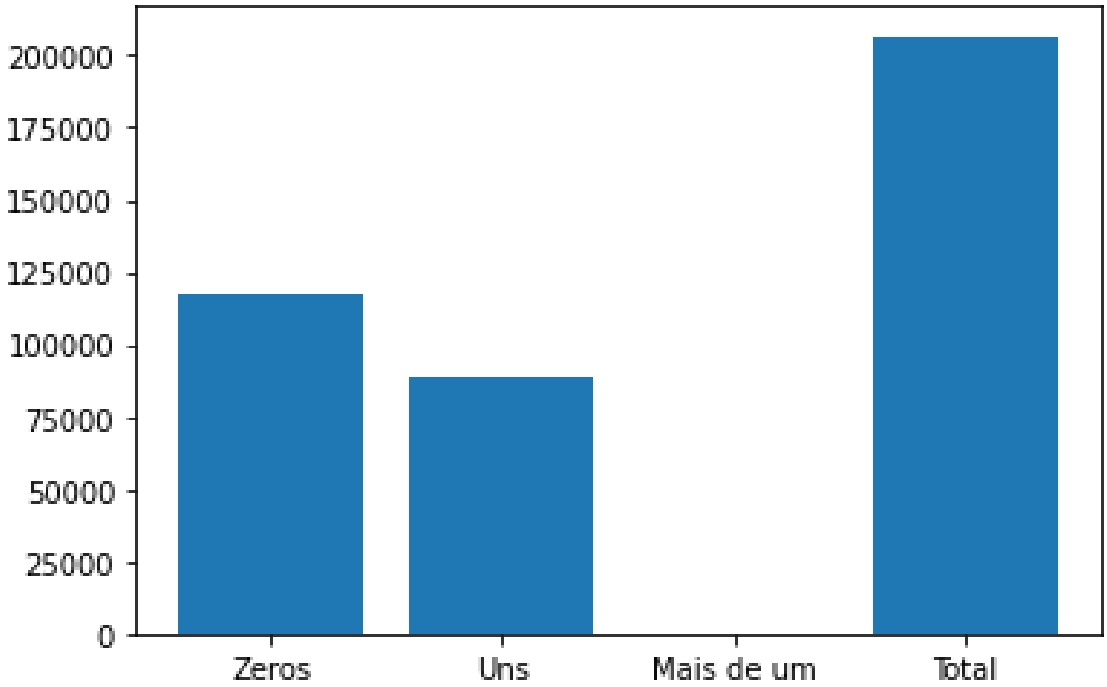

Figura 4.3: *Gráfico comparativo entre a coocorrência dos termos pertencentes às categorias.*

A Figura 4.3 mostra a ocorrência dos termos em um *tweet*, evidenciando que a classificação usando a cardinalidade não é viável, uma vez que poucos termos ocorrem simultaneamente no contexto analisado. A primeira coluna à esquerda são a quantidade de *tweets* em que não apareceram termos, a segunda é a quantidade de *tweets* em que aparece somente um termo, a terceira é quando aparecem mais de um e ao lado o total da base de dados. Concluindo, como não se podia assegurar a execução das regras propostas em [Rodrigues 2019], levando em conta a cardinalidade dos termos que nelas ocorriam, foi necessário modificar as regras para que a ontologia retornasse mais *tweets* com traços de ansiedade e depressão.

4.1.1 Modificações realizadas

Remoção de *retweets*

Percebeu-se que muitas das palavras mais frequentes na base estavam relacionadas com *retweets*, prática comum em tal rede social, na qual os usuário reenviam uma mensagem de outro usuário. Como o intuito é identificar inclinação à tristeza, chegou-se à conclusão que esses *retweets* não eram o alvo da análise. Sendo assim, fez-se um novo teste retirando da base de dados os *retweets*, a limpeza retornou 43244 *tweets*, conforme a Figura 4.4.

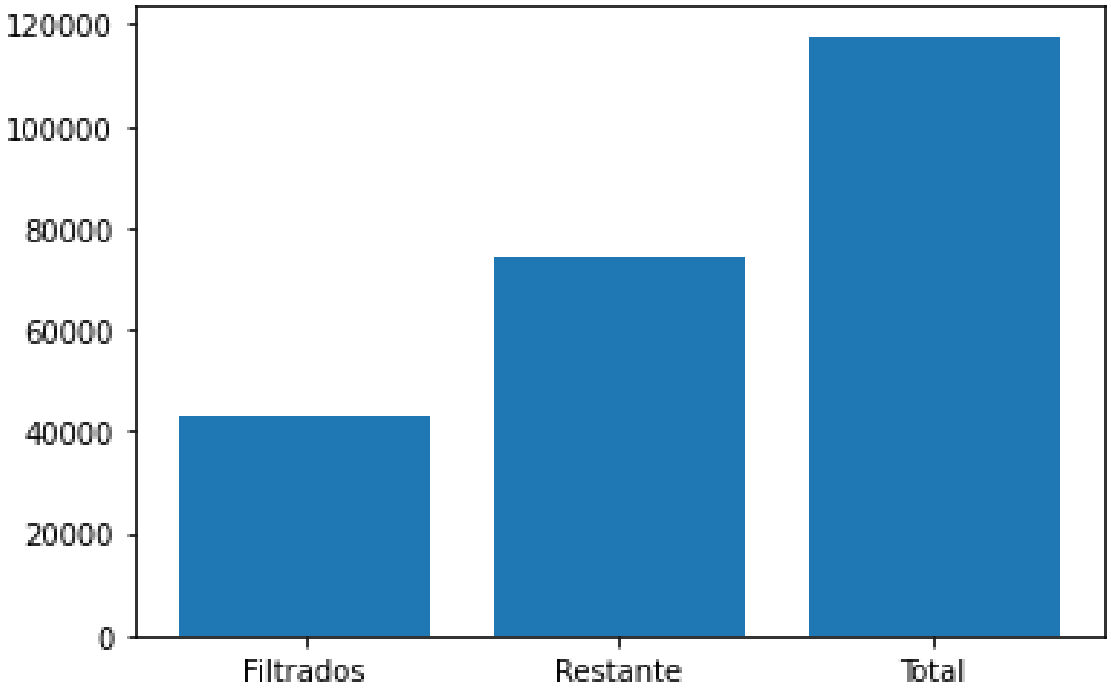

Figura 4.4: *Gráfico comparativo entre a quantidade filtrada e o total da base.*

A filtragem por *retweets* gerou uma nuvem de palavras mais fiel ao contexto analisado, conforme explicitado pela Figura 4.5. Apareceram palavras já esperadas como Vida, Ansiedade e Pessoa.

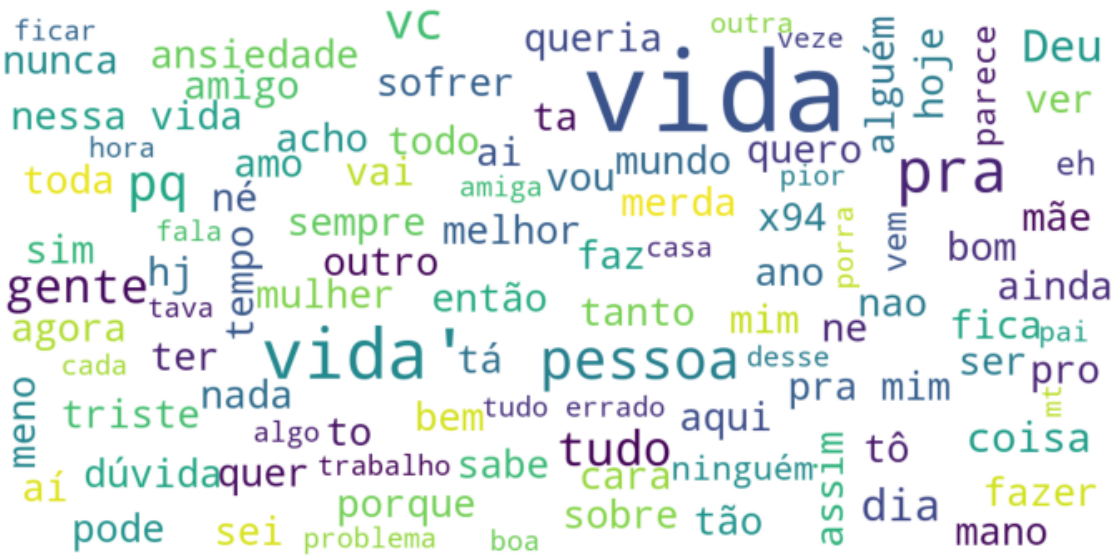

Figura 4.5: *Nuvem de palavras da filtragem por retweets.*

Remoção da hierarquização

Com a base mais fiel ao contexto, foi necessária a aplicação de um novo filtro pelos termos mencionados na Tabela 4.1, sem a hieraquização defendida por [Rodrigues 2019]. Com isso, reduziu-se a base de dados para 37566 *tweets*. A Figura 4.6 mostra a diferença no tamanho da base de acordo com as novas filtragens utilizadas, mostrando que houve pouca mudança, sugerindo que as palavras utilizadas nesse processo eram muito genéricas, contribuindo para a generalidade da base.

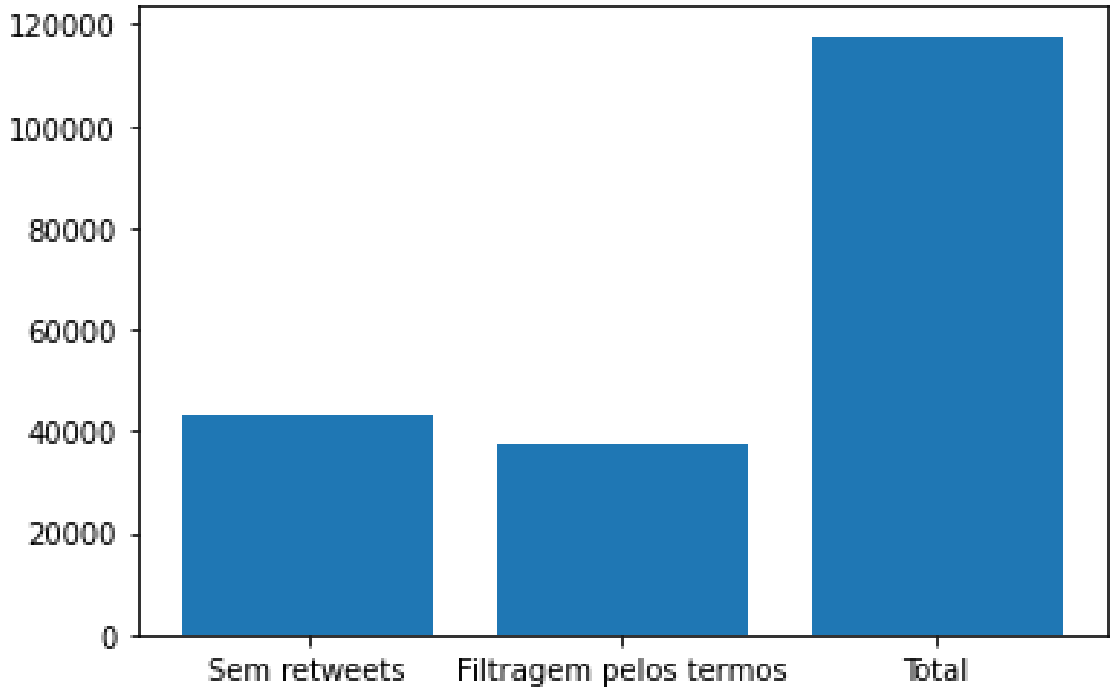

Figura 4.6: *Gráfico comparativo entre a filtragem por retweets e pelos termos sem hierarquia.*

A nuvem de palavras resultante desse processo ficou muito parecida com a anterior, ratificada pela Figura 4.7, com a exceção de poucos termos:

Figura 4.7: *Nuvem de palavras da filtragem usando os termos já citados.*

Nova classificação

Fez-se necessário, mediante as análises supracitadas, uma nova filtragem, que tornasse a base mais fideldigna ao contexto em estudo. Uma revisão sistemática dos trabalhos correlatos em Cap. 2, mostrou que outros trabalhos utilizaram ferramentas de diagnóstico para a construção do conjunto de termos. Usando como base o Inventário de Beck, chegou-se aos seguintes termos:

Novos termos para filtragem	
triste	infeliz
decepcionad	desanimad
fracass	cansad
preocupad	irrita
alguem	nada
odeio	culp
err	pior
chorar	fraqueza
matar	nunca
durm	dor
morrer	ansiedade

Tabela 4.2: *Novos termos alcançados via revisão sistemática dos trabalhos correlatos.*

Consoante Tabela 4.2, o radical foi retirado para não restringir a busca pela base de dados. Colocando em prática o novo método, obteve-se a uma base de cerca de nove mil *tweets* e uma nuvem de palavras mais fiel ao contexto do que as anteriores, o que pode ser visto na Figura 4.8:

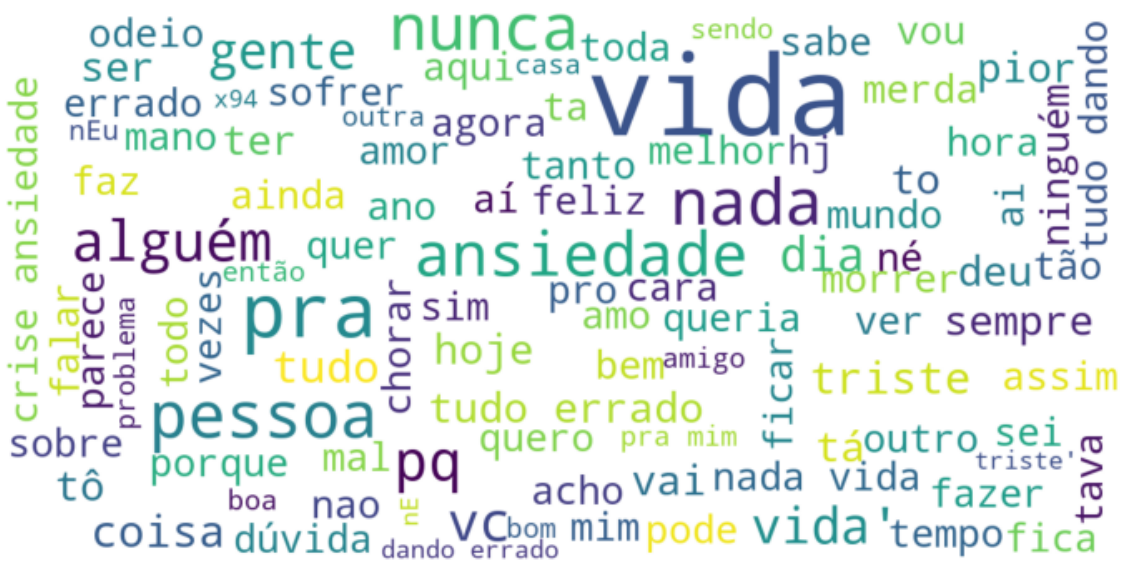

Figura 4.8: *Nuvem de palavras da filtragem usando os novos termos*

Palavras e termos como odeio, crise, morrer, pior, tudo errado, triste, aparecem somente na nuvem acima. Palavras como ansiedade aumentaram bastante a frequência de ocorrência.

Dessa forma, concluiu-se que a aplicação das regras desenvolvidas por [Rodrigues 2019] não funcionaram e necessitaram de ajustes, uma vez que poucos termos estão associados com as categorias propostas, na maioria somente um aparece. Da mesma forma ocorreu na filtragem pelos termos sem hierarquia, mesmo com a retirada dos *retweets*, não encerrou a generalidade da base, que retornou uma nuvem de palavras semelhante à inicial. Conseguiu-se somente atingir o objetivo aplicando-se na base uma nova filtragem utilizando os termos citados na Tabela 4.2, o que gerou uma nuvem mais fiel ao contexto desejado. Sendo assim, essas modificações realizadas serão adotadas como abordagem ontológica para a continuação deste trabalho.

4.2 Modelo do classificador

Os classificadores utilizados para o ensaio e testes foram os modelos regressão logística, redes neurais, KNN e SVM para uma análise supervisionada dos dados. A realização do treinamento, validação e classificação foram conduzidas através do uso da ferramenta Orange [Demšar et al. 2013] para análise e estudo em mineração de dados e processamento de linguagem natural.

Utilizando dos conhecimentos adquiridos e relatados no Capítulo 3 sobre análise de sentimentos, implementou-se a classificação binária dos *tweets* em: possui traços de tristeza (termos associados à ansiedade e depressão) ou não possui.

Cada *tweet* é encarado como sendo um documento, dessa forma, a classificação é realizada à nível de documento conforme descrito na Seção 3.3.1.

A base de dados original contém 117603 *tweets* coletados por [Rodrigues 2019] utilizando como filtro as palavras contidas na Tabela 4.1.

4.2.1 Treinamento dos classificadores

Esses *tweets* foram selecionados manualmente de acordo com a tendência à tristeza que cada um continha, sendo classificados em presente e ausente para essa característica. Essa seleção gerou uma base de quatrocentos *tweets* que possuíam inclinação à tristeza e quatrocentos que não possuíam. Essa base de dados selecionados passou por um pré-processamento manual, onde foram removidas acentuações, símbolos, pontuações e *links*, depois os termos foram separados usando uma *bag of words* e foram colocados como entrada em classificadores.

Tweets com negatividade	Tweets sem negatividade
eu faco tudo errado depois fico chorando com as consequencias	eu estou muito apaixonada nesse video
Agora lembrei me da pior coisa que ja passei na vida	Eu fico muito emocionada de ver alguem com um coracao tao lindo e gigante igual o seu

Figura 4.9: *Exemplos de termos obtidos.*

Após o treinamento foram obtidas métricas frequentemente adotadas em Inteligência Artificial [Hossin e Sulaiman 2015] para auxiliar na escolha do melhor classificador para a continuação do trabalho:

- AUC: Área abaixo da curva do gráfico de ROC, sendo esse último composto pela taxa de positivos verdadeiros sobre os negativos verdadeiros. Quanto mais próximo de 1, melhor;
- Acurácia (CA): Proporção de exemplos classificados corretamente;
- F-1: Média harmônica ponderada entre precisão e *recall*;
- Precisão: Razão entre os positivos verdadeiros sobre o total classificado como positivo;
- Recall: Razão entre os positivos verdadeiros sobre o total de positivos.

O treinamento e validação realizados no Orange gerou a matriz de precisão disposta na Figura 4.10. Conforme os resultados obtidos, verificou-se que os classificadores de redes neurais e regressão logística tinham as melhores métricas.

Model	AUC	CA	F1	Precision	Recall
kNN	0.888	0.802	0.802	0.806	0.802
SVM	0.814	0.686	0.667	0.743	0.686
Neural Network	0.892	0.818	0.817	0.818	0.818
Logistic Regression	0.906	0.820	0.820	0.821	0.820

Figura 4.10: *Resultado do treinamento dos classificadores.*

Dessa forma, os classificadores regressão logística e redes neurais foram selecionados para continuar o experimento.

4.2.2 Teste dos classificadores

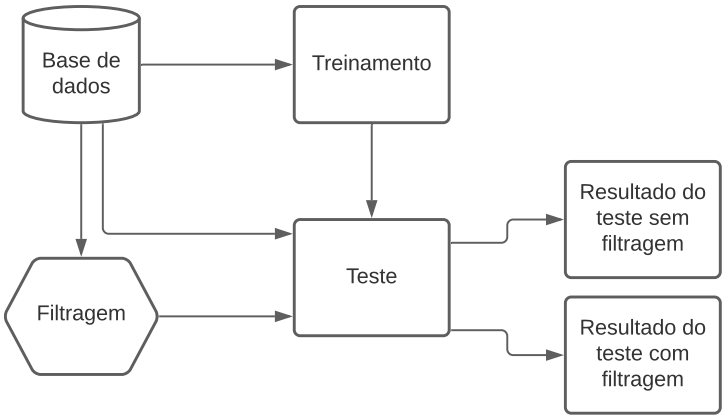

Figura 4.11: *Diagrama do teste dos classificadores.*

A Figura 4.11 exemplifica processo de teste dos classificadores, a partir dos dados coletados anteriormente, em [Rodrigues 2019], separamos uma amostra para testar o quão a filtragem por ontologia melhoraria o desempenho desses sistemas de classificação.

Para realizar o teste dos classificadores obtidos na Seção 4.2.1, foi selecionado outro conjunto de dados. Partindo da mesma base utilizada para a coleção de treinamento, separou-se novamente um subconjunto de cem *tweets* com inclinação à tristeza e cem que não o continham. Essa base de teste passou pelo mesmo pré-processamento que a base de treinamento, sendo, então, testada nos classificadores.

Para testar a ontologia proposta em Seção 4.1.1, realizou-se uma filtragem na base supracitada usando os métodos descritos na Seção 4.1, usando a classificação baseada nos termos do Inventário de Beck [Beck e Alford 2016]. Essa base filtrada foi composta por cinquenta *tweets*, sendo esses quarenta e seis para inclinação e quatro para ausência dessa característica, sendo novamente testada no classificador treinado.

CAPÍTULO 5

Resultados

Na Seção 4.2.2 foi descrito o método de teste que foi realizado nesse trabalho, consistindo na separação de uma amostra da base inicial, classificação manual e pré-processamento, para então testar o desempenho do classificador treinado com essa base filtrada e sem filtragem, conforme a Figura 4.10. Os referidos testes geraram os resultados que seguem.

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.908	0.835	0.835	0.836	0.835
Logistic Regression	0.910	0.800	0.798	0.812	0.800

Figura 5.1: *Matriz de precisão do teste sem filtragem.*

A Figura 5.1 tem-se que base pré-processada foi testada nos classificadores sem ser filtrada pela ontologia. Esse teste gerou a matriz de precisão acima. Pelos parâmetros descritos em 4.2.1, vê-se que a precisão da rede neural foi melhor do que a da regressão logística. Isso significa que a rede neural teve mais positivos verdadeiros no total classificado como positivo do que a da regressão logística.

Já o segundo teste, foi feito usando as técnicas de filtragem pela ontologia, descritas na Seção 4.1.1 e exemplificada na Figura 4.11:

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.886	0.875	0.893	0.925	0.875
Logistic Regression	0.881	0.771	0.814	0.882	0.771

Figura 5.2: *Matriz de precisão para o teste com filtragem ontológica.*

Pela Figura 5.2 pode-se perceber que houve melhora significativa quando os termos foram filtrados usando uma ontologia adequada ao contexto.

Para comprovar a validade da saída dos experimentos, utilizou-se o teste de aderência, que consiste em calcular a distância entre o resultado esperado e os resultados

obtidos, conforme a equação $\Sigma\frac{(O-E)}{E}^2$, onde O é a frequência observada e E é a frequência esperada, [McHugh 2013]. Os valores supracitados são dispostos em uma tabela. O cálculo da frequência esperada é expressada pela equação $\frac{M_R * M_C}{n}$, onde M_R é a soma da linha para aquela célula, M_C é a soma da coluna para aquela célula e n é o tamanho total da amostra [McHugh 2013].

A frequência comparativa entre os testes, dos *tweets* com inclinação e sem, com o que os classificadores categorizaram como positivo ou negativo estão dispostos na Figura 5.3.

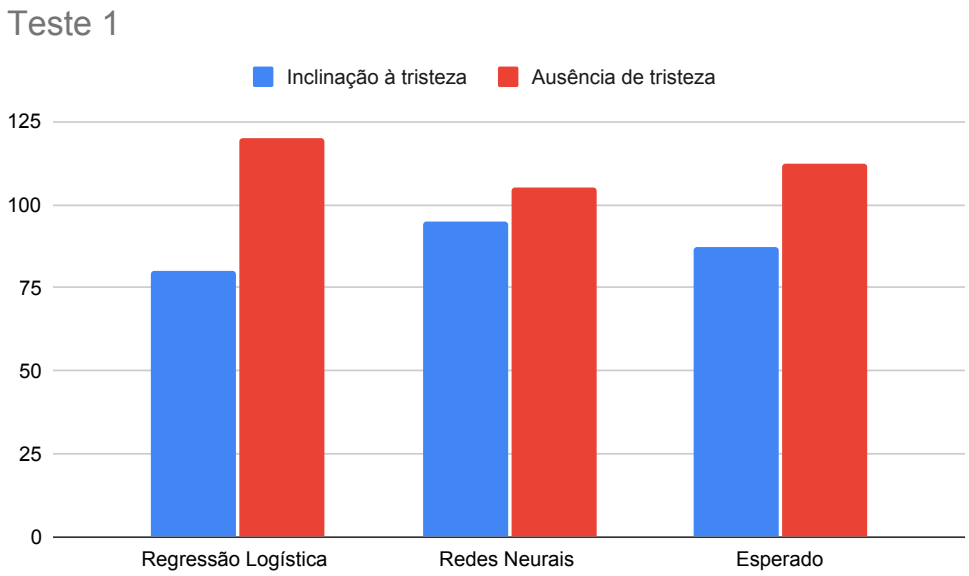

Figura 5.3: *Frequência dos termos sem filtragem.*

Na Figura 5.3 tem-se três colunas, onde as duas primeira representam a taxa de classificação para cada categoria de acordo com o classificador. A terceira coluna, mais à direita, representa os verdadeiros positivos para cada categoria, sendo assim, o classificador apresenta melhor resultado quando a classificação fica próxima ao esperado, representado pela coluna à direita. Dito isso, vê-se que a rede neural apresentou resultados mais próximos à classificação que se esperava.

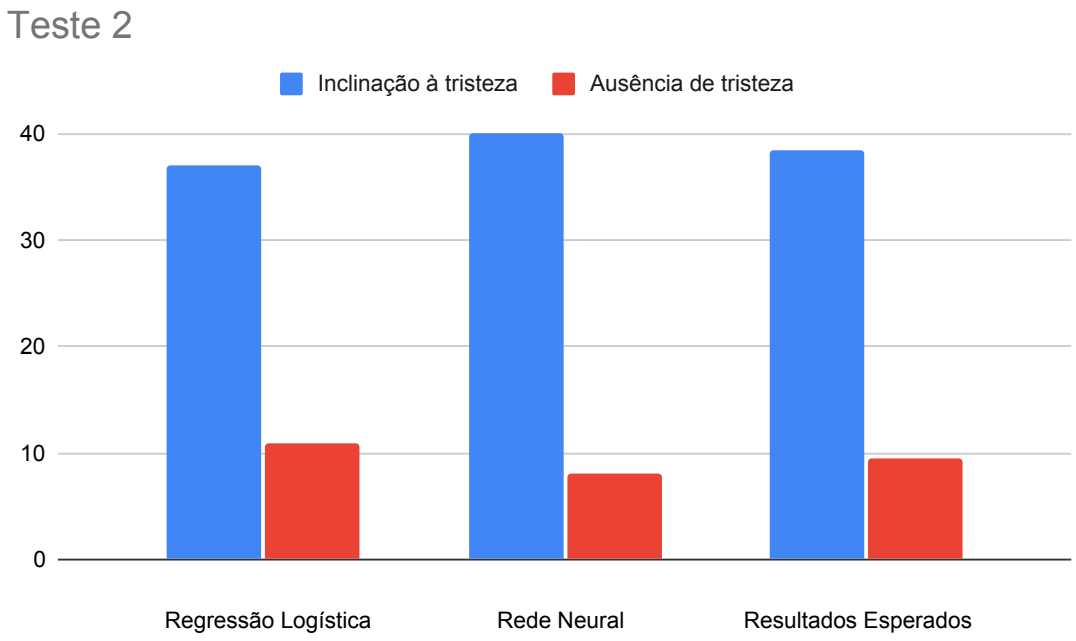

Figura 5.4: *Frequência dos termos com filtragem ontológica.*

Na Figura 5.4 tem-se que os resultados esperados, os verdadeiros positivos para cada categoria citada, ficou mais próximo dos obtidos pela rede neural. Para a rede neural, quarenta *tweets* continham inclinação à tristeza, desses, trinta e seis eram verdadeiros positivos, representados pela coluna Resultados Esperados.

O somatório resultante da aplicação do teste qui-quadrado é então comparado com uma tabela de distribuição padrão, de acordo com os graus de liberdade e o nível de significância. Para o teste deste trabalho, foi considerado um grau de liberdade, pelo fato de ter-se duas categorias, inclinação à tristeza e ausência dessa, e um nível de significância de vinte por cento, justificado pela precisão obtida pelo classificador de oitenta por cento, sendo o complemento desse. Essa precisão foi definida no treinamento do classificador, para verificar se haveria melhora nos valores do teste quando comparados com a etapa de treinamento.

Teste	Valor calculado	Valor de referência
Teste 1	2,2857	1,642
Teste 2	0,591	1,642

Tabela 5.1: *Resultados do teste qui-quadrado.*

De acordo com a Tabela 5.1, temos os valores calculados, sendo esse o somatório dos valores obtidos em cada teste. Para um grau de liberdade e um nível de significância de 0,20, obtemos 1,642 usando a tabela de qui-quadrado. Quando os valores são comparados com o de referência, tem-se que o do Teste 2 ficou abaixo do referencial, indicando que

está dentro do que se esperava. Sendo assim, houve melhora na aderência do teste 2 em comparação ao Teste 1.

CAPÍTULO 6

Conclusão

Pelos resultados descritos no Cap. 5, verificou-se que o treinamento gerou dois classificadores consistentes para as etapas seguintes, de acordo com a Figura 4.10, capazes de detectar a tendência à tristeza em *tweets*. O teste dos classificadores também foi satisfatório, vide Figuras 5.1 e 5.2, tanto para a base filtrada quanto à outra e os resultados comprovam que houve melhora observável em utilizar uma ontologia para a seleção dos dados de entrada de um classificador, verificável pelo aumento da aderência como visto na Tabela 5.1, mostrando uma melhora no segundo teste, quando foi utilizado ontologia com os termos descritos em 4.2. Essa melhora se deve ao fato da redução do campo de palavras quando a ontologia foi empregada, redução essa descrita na Seção 3.2. As nuvens de palavras mostradas no Cap. 4 indicam essa tendência, evidenciando o contexto e tornando mais visível expressões e palavras que pertenciam ao domínio da inclinação à tristeza, exemplificado pela Figura 4.9. Os dados selecionados para o teste se mostraram mais sensíveis ao contexto após a filtragem, como ficou exposto pela Figura 5.2, onde a precisão para o classificador de redes neurais alcançou noventa e dois por cento, em contraste com os oitenta e três por cento registrados quando não houve filtragem pela ontologia.

Assim, fica claro que a utilização de ontologia melhorou a eficiência dos classificadores escolhidos, como visto no Cap. 5, pois ao reduzir o contexto ontológico, o escopo léxico ficou mais adequado à classificação, uma vez que ficaram mais nítidas as associações e o reconhecimento de padrões a que ela visa a procurar.

Como trabalhos futuros, temos que essa mesma metodologia pode ser usada em outros campos da área de análise de sentimentos, para detecção de tendência à compra, tendência ante à política e em casos onde se faz necessário saber a inclinação geral em relação a uma entidade ou sentimento. Ou visando a expansão desse trabalho, poderia-se abordar melhor a montagem da ontologia, de forma a melhorar a filtragem realizada por essa.

Referências Bibliográficas

[Ali et al. 2019]ALI, F. et al. Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowledge-Based Systems*, Elsevier B.V., v. 174, n. xxxx, p. 27–42, 2019. ISSN 09507051. Disponível em: <https://doi.org/10.1016/j.knosys.2019.02.033>.

[Asprino et al. 2019]ASPRINO, L. et al. Ontology-based knowledge management for comprehensive geriatric assessment and reminiscence therapy on social robots. In: *Data Science for Healthcare*. [S.l.]: Springer, 2019. p. 173–193.

[Association et al. 2014]ASSOCIATION, A. P. et al. *DSM-5: Manual diagnóstico e estatístico de transtornos mentais*. [S.l.]: Artmed Editora, 2014.

[Azevedo, Belo e Romero 2019]AZEVEDO, H.; BELO, J. P. R.; ROMERO, R. A. Using ontology as a strategy for modeling the interface between the cognitive and robotic systems. *Journal of Intelligent & Robotic Systems*, Springer, p. 1–19, 2019.

[Baniasadi et al. 2017]BANIASADI, N. et al. ASSESSING THE SLEEP QUALITY AND DEPRESSION-ANXIETY-STRESS IN IRRITABLE BOWEL SYNDROME PATIENTS. *Arquivos de Gastroenterologia*, scielo, v. 54, p. 163 – 166, 06 2017. ISSN 0004-2803.

[Beck e Alford 2016]BECK, A. T.; ALFORD, B. A. *Depressão: causas e tratamento*. [S.l.]: Artmed Editora, 2016.

[Beloozerov et al. 2019]BELOOZEROV, V. et al. Conception of ontology for computer methods of paintings analysis. In: IEEE. *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*. [S.l.], 2019. p. 0903–0908.

[Benevenuto, Ribeiro e Araújo 2015]BENEVENUTO, F.; RIBEIRO, F.; ARAÚJO, M. Métodos para análise de sentimentos em mídias sociais. In: *Brazilian Symposium on Multimedia and the Web (Webmedia), Manaus, Brasil*. [S.l.: s.n.], 2015.

[Bhat e Rather 2012]BHAT, M. A.; RATHER, T. A. Socio-economic factors and mental health of young people in india and china: An elusive link with globalization. *Asian Social Work and Policy Review*, Wiley Online Library, v. 6, n. 1, p. 1–22, 2012.

[Blackburn 1997]BLACKBURN, S. *Dicionário Oxford de filosofia*. [S.l.]: Jorge Zahar Editor, 1997.

[Borst e Borst 1997]BORST, W.; BORST, W. *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. Tese (Doutorado) — University of Twente, Netherlands, 9 1997.

[Cardoso 2017]CARDOSO, L. R. D. Psicoterapias comportamentais no tratamento da depressão. *Psicologia argumento*, v. 29, n. 67, 2017.

[Choudhury et al. 2013]CHOUDHURY, M. D. et al. Predicting depression via social media. In: *Seventh international AAAI conference on weblogs and social media*. [S.l.: s.n.], 2013.

[Coppin 2004]COPPIN, B. *Artificial Intelligence Illuminated*. [s.n.], 2004. 768 p. ISSN 02650096. ISBN 0763732303. Disponível em: <https://dochub.com/thiagoanc/DL7JlEGV1NrZE6OwrWe0oa/artificial-intelligence-illuminated-pdfdrive-com-pdf>.

[Cremasco e MN]CREMASCO, G. S.; MN, B. *Depressão, motivos para viver eo significado do suic{\'\i}dio em graduandos do curso de psicologia. Estud Interdiscip Psicol. 2017; 8 (1): 22-37*.

[Dainez e Dainez 2015]DAINEZ, P. S.; DAINEZ, E. C. L. Rede Neural artificial Aplicada em um Sistema de Auxilio no Rastreio de Depressão e de Qualidade de Vida de Idosos. *Learning and Nonlinear Models*, v. 13, n. 2, p. 67–72, 2015.

[Demšar et al. 2013]DEMŠAR, J. et al. Orange: data mining toolbox in python. *the Journal of machine Learning research*, JMLR. org, v. 14, n. 1, p. 2349–2353, 2013.

[Dhar 2013]DHAR, V. Data science and prediction. *Communications of the ACM*, ACM New York, NY, USA, v. 56, n. 12, p. 64–73, 2013.

[Dias 2018]DIAS, L. P. S. iAwre: um modelo para cuidado ub{\'\i}quo de pacientes com transtornos de ansiedade, depressão e estresse utilizando gamificação e biodata. Universidade do Vale do Rio dos Sinos, 2018.

[Ding, Liu e Yu 2008]DING, X.; LIU, B.; YU, P. S. A holistic lexicon-based approach to opinion mining. In: *Proceedings of the 2008 international conference on web search and data mining*. [S.l.: s.n.], 2008. p. 231–240.

[Ding, Wu e Zhang 2019]DING, Y.; WU, R.; ZHANG, X. Ontology-based knowledge representation for malware individuals and families. *Computers & Security*, Elsevier, v. 87, p. 101574, 2019.

[Dorneles e Others 2019]DORNELES, B. S.; OTHERS. Support vector machines na identificação de opiniões depressivas em redes sociais. Universidade Federal da Grande Dourados, 2019.

[Dragoni, Poria e Cambria 2018]DRAGONI, M.; PORIA, S.; CAMBRIA, E. OntoSenticNet: A Commonsense Ontology for Sentiment Analysis. *IEEE Intelligent Systems*, IEEE, v. 33, n. 3, p. 77–85, 2018. ISSN 19411294.

[Duque, Raymundo e Neto 2018]DUQUE, J. W. G.; RAYMUNDO, A. L.; NETO, P. F. Uma aplicação de big data para classificação de sentenças depressivas do Twitter. *Revista H-TEC Humanidades e Tecnologia*, v. 2, n. 1, p. 82–95, 2018.

[Genesereth e Nilsson 1987]GENESERETH, M. R.; NILSSON, N. J. Logical foundations of. *Artificial Intelligence. New York: Morgan Kaufmann Publishers*, 1987.

[Griffiths e Steyvers 2003]GRIFFITHS, T. L.; STEYVERS, M. Prediction and semantic association. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2003. p. 11–18.

[Gruber 1993]GRUBER, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition*, Academic Press, v. 5, n. 2, p. 199–220, jun 1993. ISSN 10428143.

[G.Tau et al. 2004]G.TAU et al. An ontology based context model in Intelligent environments. *Proc. of CNDS*, 2004.

[Guarino 1998]GUARINO, N. *Formal ontology in information systems: Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy*. [S.l.]: IOS press, 1998.

[Guarino, Oberle e Staab 2009]GUARINO, N.; OBERLE, D.; STAAB, S. What is an ontology? In: *Handbook on ontologies*. [S.l.]: Springer, 2009. p. 1–17.

[Hand e Adams 2014]HAND, D. J.; ADAMS, N. M. Data mining. *Wiley StatsRef: Statistics Reference Online*, Wiley Online Library, p. 1–7, 2014.

[Hatzivassiloglou e Wiebe 2000]HATZIVASSILOGLOU, V.; WIEBE, J. Effects of adjective orientation and gradability on sentence subjectivity. In: *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*. [S.l.: s.n.], 2000.

[Health 2016]HEALTH, T. N. I. of M. *NIMH ż Anxiety Disorders*. 2016. Disponível em: <https://www.nimh.nih.gov/health/topics/anxiety-disorders/index.shtml>.

[Hofmann 1999]HOFMANN, T. Probabilistic latent semantic indexing. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. [S.l.: s.n.], 1999. p. 50–57.

[Hossin e Sulaiman 2015]HOSSIN, M.; SULAIMAN, M. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, Academy & Industry Research Collaboration Center (AIRCC), v. 5, n. 2, p. 1, 2015.

[Hu e Liu 2004]HU, M.; LIU, B. Mining and summarizing customer reviews. In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2004. p. 168–177.

[IHME DATA 2019]IHME DATA. *GBD Results Tool | GHDx*. 2019. Disponível em: <http://ghdx.healthdata.org/gbd-results-tool>.

[James et al. 2018]JAMES, S. L. et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 19902017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, Lancet Publishing Group, v. 392, n. 10159, p. 1789–1858, nov 2018. ISSN 01406736. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0140673618322797>.

[Jiang et al. 2011]JIANG, L. et al. Target-dependent twitter sentiment classification. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. [S.l.: s.n.], 2011. p. 151–160.

[Kemp]KEMP, S. *Digital 2020: Global Digital Overview. DataReportal. 2020*.

[Kharbat e El-Ghalayini 2008]KHARBAT, F.; EL-GHALAYINI, H. *Building ontology from knowledge base systems*. [S.l.]: INTECH Open Access Publisher, 2008.

[Kim e Hovy 2004]KIM, S.-M.; HOVY, E. Determining the sentiment of opinions. In: *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. [S.l.: s.n.], 2004. p. 1367–1373.

[Köhler et al. 2017]KÖHLER, S. et al. The human phenotype ontology in 2017. *Nucleic Acids Research*, v. 45, n. D1, p. D865–D876, 2017. ISSN 13624962.

[Kumar e Joshi 2017]KUMAR, A.; JOSHI, A. Ontology driven sentiment analysis on social web for government intelligence. *ACM International Conference Proceeding Series*, Part F1276, p. 134–139, 2017.

[Lafferty, McCallum e Pereira 2001]LAFFERTY, J.; MCCALLUM, A.; PEREIRA, F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[Liu 2010]LIU, B. Sentiment analysis and subjectivity. In: *Handbook of Natural Language Processing, Second Edition*. [s.n.], 2010. p. 627–666. ISBN 9781420085938. Disponível em: <https://dochub.com/thiagoanc/dbnaAMqK9Y53JGvKGNXJm0/nlp-handbook-sentiment-analysis-pdf>.

[Liu 2012]LIU, B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, v. 5, n. 1, p. 1–184, 2012. ISSN 19474040. Disponível em: <https://dochub.com/thiagoanc/r4D6EkZVZZ2LZQJVpQXW7O/estudo-dtm-sentimentanalysisandopinionmining-bingliu-pdf?pg=40>.

[McHugh 2013]MCHUGH, M. L. The chi-square test of independence. *Biochemia medica: Biochemia medica*, Medicinska naklada, v. 23, n. 2, p. 143–149, 2013.

[Méa, Biffe e Ferreira 2016]MÉA, C. P. D.; BIFFE, E. M.; FERREIRA, V. R. T. Padrão de uso de internet por adolescentes e sua relação com sintomas depressivos e de ansiedade. *Psicologia Revista*, v. 25, n. 2, p. 243–264, 2016.

[Mendonça e Soares 2017]MENDONÇA, F. M.; SOARES, A. L. Construindo ontologias com a metodologia ontoforinfoscience: Uma abordagem detalhada das atividades do desenvolvimento ontológico. *Ciencia da Informacao*, v. 46, n. 1, p. 43–59, 2017. ISSN 01001965.

[Ministério da Saúde 2014]Ministério da Saúde. *Depressão: como diagnosticar, quais os sintomas e qual o tratamento*. 2014. 1–2 p. Disponível em: <https://saude.gov.br/saude-de-a-z/depressao http://saude.gov.br/saude-de-a-z/depressao>.

[Mizoguchi 2003]MIZOGUCHI, R. *Tutorial on Ontological Engineering*. 2003. 363–364 p. Disponível em: <https://pt.scribd.com/document/56477570/Mizoguchi-2004-Tutorial-on-Ontological-Engineering-I>.

[Moraes e Others 2020]MORAES, L. V. de; OTHERS. Detecção de depressão pela fala empregando rede neurais profundas. Universidade Federal de Goiás, 2020.

[Moraes et al. 2016]MORAES, S. M. et al. Comparing approaches to subjectivity classification: A study on portuguese tweets. In: SPRINGER. *International Conference on Computational Processing of the Portuguese Language*. [S.l.], 2016. p. 86–94.

[Morais e Ambrósio 2007]MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Ontologias: conceitos, usos, tipos, metodologias, ferramentas e linguagens. p. 21, 2007. Disponível em: <http://www.portal.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-07.pdf>.

[NIHM 2019]NIHM. *Major Depression*. 2019. Disponível em: <https://www.nimh.nih.gov/health/statistics/major-depression.shtml>.

[Ning et al. 2019]NING, H. et al. A novel ontology consistent with acknowledged standards in smart homes. *Computer Networks*, Elsevier, v. 148, p. 101–107, 2019.

[Noy e McGUINNESS 2001]NOY, N. F.; MCGUINNESS, D. L. Ontology development 101: a guide to create your first ontology. *Knowledge Systems Laboratory Technical Report KSL-01-05, Stanford University. 25p*, 2001.

[OMS]OMS. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/depression>.

[Pang, Lee e Vaithyanathan 2002]PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. [S.l.], 2002. p. 79–86.

[Petry 2016]PETRY, M. M. Hígia: um modelo para cuidado ubíquo de pessoas com depressão. Universidade do Vale do Rio dos Sinos, 2016.

[Psiquiatria]PSIQUIATRIA, S. B. de. *Prevenção do Suicídio Setembro Amarelo*. Disponível em: <https://www.setembroamarelo.com/ https://www.polbr.med.br/ano17/wal0817.php>.

[Rabiner 1989]RABINER, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, Ieee, v. 77, n. 2, p. 257–286, 1989.

[Rodrigues 2019]RODRIGUES, L. B. P. *Comparative study of ontology construction processes for the sentiment analysis and the formation of dictionaries.* Tese (Doutorado) — Federal University of Goias, Brazil, 6 2019.

[Russell e Norvig 2002]RUSSELL, S.; NORVIG, P. Artificial intelligence: a modern approach. 2002.

[Saúde]SAÚDE, M. da. *Ansiedade*. Disponível em: <https://bvsms.saude.gov.br/dicas-em-saude/470-ansiedade>.

[Schouten, Frasincar e Jong 2017]SCHOUTEN, K.; FRASINCAR, F.; JONG, F. de. Ontology-enhanced aspect-based sentiment analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 10360 LNCS, p. 302–320, 2017. ISSN 16113349.

[Silveira e Others 2019]SILVEIRA, J. N.; OTHERS. Modelo de detecção de depressão através das m{\'\i}dias sociais. Florianópolis, SC, 2019.

[Souza e Machado-De-Sousa 2017]SOUZA, I. M. de; MACHADO-DE-SOUSA, J. P. Brazil: World leader in anxiety and depression rates. *Revista Brasileira de Psiquiatria*, Associacao Brasileira de Psiquiatria, v. 39, n. 4, p. 384, oct 2017. ISSN 15164446.

[Stats 2018]STATS, I. W. *World Internet Users Statistics*. 2018. Disponível em: <https://internetworldstats.com/stats.htm https://www.internetworldstats.com/stats.htm https://www.internetworldstats.com/stats.htm%0Ahttp://www.internetworldstats.com/stats.htm>.

[Taboada et al. 2011]TABOADA, M. et al. Lexicon-based methods for sentiment analysis. *Computational linguistics*, MIT Press, v. 37, n. 2, p. 267–307, 2011.

[Teng, Humes e Demetrio 2005]TENG, C. T.; HUMES, E. d. C.; DEMETRIO, F. N. Depressão e comorbidades clínicas. *Archives of Clinical Psychiatry (São Paulo)*, SciELO Brasil, v. 32, n. 3, p. 149–159, 2005.

[Turney 2001]TURNEY, P. D. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In: SPRINGER. *European conference on machine learning*. [S.l.], 2001. p. 491–502.

[Turney 2002]TURNEY, P. D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 40th annual meeting on association for computational linguistics*. [S.l.], 2002. p. 417–424.

[Uschold e Gruninger 1996]USCHOLD, M.; GRUNINGER, M. Ontologies: Principles, methods and applications. *The knowledge engineering review*, Cambridge University Press, v. 11, n. 2, p. 93–136, 1996.

[Wang et al. 2007]WANG, P. S. et al. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. *Lancet*, NIH Public Access, v. 370, n. 9590, p. 841–850, sep 2007. ISSN 01406736.

[What Is Depression?]WHAT Is Depression? Disponível em: <https://www.psychiatry.org/patients-families/depression/what-is-depression>.

[WHO 2017]WHO. Depression and Other Common Mental Disorders: Global Health Estimates. *World Health Organization*, World Health Organization, v. 1, p. 24, 2017.

[Wiebe, Bruce e OHara 1999]WIEBE, J.; BRUCE, R.; OHARA, T. P. Development and use of a gold-standard data set for subjectivity classifications. In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 1999. p. 246–253.

[Wiebe et al. 2000]WIEBE, J. et al. Learning subjective adjectives from corpora. *Aaai/iaai*, Austin, TX, v. 20, n. 0, p. 0, 2000.

[Zhong et al. 2019]ZHONG, Y. et al. Automatically generating assembly sequences with an ontology-based approach. *Assembly Automation*, Emerald Publishing Limited, 2019.