UNIVERSIDADE FEDERAL DE PELOTAS
Centro de Desenvolvimento Tecnológico
Programa de Pós-Graduação em Computação

Dissertação

ABSAuDA - Análise de Sentimentos baseado em Aspectos utilizando Análise de Dependência

Francisco Dias Franco

Pelotas, 2024

Francisco Dias Franco

ABSAuDA - Análise de Sentimentos baseado em Aspectos utilizando Análise de Dependência

Dissertação apresentada ao Programa de Pós-Graduação em Computação do Centro de Desenvolvimento Tecnológico da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Ulisses Brisolara Corrêa
Coorientadora: Prof[a]. Dr[a]. Larissa Astrogildo de Freitas

Pelotas, 2024

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação da Publicação

F826a Franco, Francisco Dias

ABSAuDA - Análise de Sentimentos baseado em Aspectos utilizando Análise de Dependência [recurso eletrônico] / Francisco Dias Franco ; Ulisses Brisolara Corrêa, orientador ; Larissa Astrogildo de Freitas, coorientadora. — Pelotas, 2024.
166 f. : il.

Dissertação (Mestrado) — Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2024.

1. Análise de Sentimento. 2. Árvore de Dependência Sintática. 3. Aspecto. 4. Ontologia. 5. Hontology. I. Corrêa, Ulisses Brisolara, orient. II. Freitas, Larissa Astrogildo de, coorient. III. Título.

CDD 005

Elaborada por Maria Inez Figueiredo Figas Machado CRB: 10/1612

Francisco Dias Franco

ABSAuDA - Análise de Sentimentos baseado em Aspectos utilizando Análise de Dependência

Dissertação aprovada, como requisito parcial, para obtenção do grau de Mestre em Ciência da Computação, Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas.

Data da Defesa: 21 de fevereiro de 2024

Banca Examinadora:

Prof. Dr. Ulisses Brisolara Corrêa (orientador)
Doutor em Computação pela Universidade Federal de Pelotas.

Prof. Dr. Thiago Berticelli Ló
Doutor em Engenharia Agrícola pela Universidade Estadual do Oeste do Paraná.

Prof. Dr. Ricardo Matsumura de Araujo
Doutor em Computação pela Universidade Federal do Rio Grande do Sul.

Dedico este trabalho a meus pais, Clóbis e Isabel, a meus irmãos, Carina, Dini e Fernanda, e a minha namorada, Alice, pelo apoio incondicional, bem como a meus orientadores, Ulisses e Larissa, pela valiosa orientação.

AGRADECIMENTOS

Agradeço primeiramente a Deus e a todo o povo espiritual por permitirem que eu ingressasse em meu sonho e concluísse a primeira etapa da minha jornada, o mestrado em Computação. Agradeço também a Mariângela Gill e Adriano Gill por me auxiliarem em minha caminhada, bem como ao Templo Espiritualista Ogum Sete Espadas pelo apoio e proteção.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, pelo apoio financeiro durante o meu mestrado.

Agradeço ao meu amigo Fabiano D'Ávila por ter-me informado sobre a disponibilidade do edital de ingresso para o mestrado em Computação, pois, sem seu aviso, eu teria perdido o prazo de inscrição e teria cursado o mestrado em Matemática.

Agradeço à Banca Examinadora pela disponibilidade de tempo, sugestões e atenção ao meu estudo.

Agradeço ao Prof. Dr. Ricardo Matsumura de Araújo por ter me indicado ao meu orientador Ulisses, responsável por iniciar a minha jornada na Análise de Sentimentos.

Agradeço aos meus orientadores, Ulisses e Larissa, aos professores do mestrado em Computação e aos membros do Laboratório 333 pela paciência em ensinar um matemático sem noção de programação avançada, e à secretária, Regina, pelo constante auxílio durante o meu mestrado.

Nossa maior fraqueza está em desistir. O caminho mais certo de vencer é tentar mais uma vez.
— THOMAS EDISON

RESUMO

FRANCO, Francisco Dias. **Análise de Sentimentos Baseada em Aspectos utilizando Análise de Dependência**. Orientador: Ulisses Brisolara Corrêa. 2024. 166 f. Dissertação (Mestrado em Ciência da Computação) – Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2024.

A Análise de Sentimentos Baseada em Aspectos consiste em examinar opiniões, sentimentos, avaliações, apreciações, atitudes e emoções das pessoas em relação a entidades e seus aspectos, expressas em diferentes formatos, como texto, imagem, áudio ou vídeo. As entidades podem ser, por exemplo, produtos e serviços, já os aspectos podem ser, por exemplo, preço e durabilidade de um produto ou serviço. Contudo, para realizar a Análise de Sentimentos Baseada em Aspectos é necessário utilizar recursos chamados de Recursos Linguísticos, que são usados para apoiar pesquisas e aplicações, como Conjuntos de Dados e Bases de Conhecimento Linguístico. No entanto, em Linguagens de Poucos Recursos, como é o caso do português brasileiro, ocorre uma carência de recursos. Devido à insuficiência recursos existe uma escassez de métodos confiáveis para Linguagens de Poucos Recursos. Este trabalho propõe um sistema de Análise de Sentimentos Baseada em Aspectos, empregando ontologias de domínio e análise de dependência sintática. Utilizou-se a ontologia Hontology, responsável por armazenar conceitos, relações e instâncias no domínio das acomodações, para identificar os aspectos relacionados com o sistema hoteleiro. Já a análise de dependência sintática é utilizada para identificar termos opinativos relacionados aos aspectos. Além disso, o estudo incorporou o uso de diversos léxicos de sentimento (por exemplo, AffectPT-br, EmoLex, LeIA, LIWC2007pt, Onto.PT, OpLexicon, Reli-Lex, SentiLex-PT, SentiWordNet-PT-BR, UNILEX, WordNetAffectBR), que são responsáveis por armazenar a polaridade das palavras, e o uso do EnsembleLex, ensemble dos resultados dos quatro melhores léxicos. Os testes foram realizados visando verificar a viabilidade do uso de correção ortográfica (manual e através das bibliotecas autocorrect, pyspellchecker e LanguageTool) e do uso da moda da união dos adjetivos que estavam relacionados com um aspecto. Os resultados demonstraram que o EnsembleLex superou os resultados obtidos pelos outros léxicos, alcançando um máximo de 0,461 para medida-f (macro) e 0,536 para medida-f (micro), usando *reviews* com anotações neutras quanto ao sentimento dirigido aos aspectos, a moda da união dos adjetivos e a biblioteca de correção ortográfica pyspellchecker.

Palavras-chave: Análise de Sentimento. Árvore de Dependência Sintática. Aspecto. Ontologia. Hontology.

ABSTRACT

FRANCO, Francisco Dias. **ABSAuDA - Aspect Based Sentiment Analysis Using Dependency Analysis**. Advisor: Ulisses Brisolara Corrêa. 2024. 166 f. Dissertation (Masters in Computer Science) – Technology Development Center, Federal University of Pelotas, Pelotas, 2024.

Aspect-Based Sentiment Analysis consists of examining people's opinions, feelings, evaluations, assessments, attitudes and emotions towards entities and their aspects, expressed in different formats, such as text, image, audio or video. Entities can be, for example, products and services, while aspects can be, for example, price and durability of a product or service. However, to perform Aspect-Based Sentiment Analysis it is necessary to use resources called Linguistic Resources, which are used to support research and applications, such as Datasets and Linguistic Knowledge Bases. However, in Low Resources Languages, such as Brazilian Portuguese, there is a lack of resources. Due to insufficient resources there is a shortage of reliable methods for Low Resource Languages. This work proposes an Aspect-Based Sentiment Analysis system, employing domain ontologies and syntactic dependency analysis. The HOntology ontology was used, responsible for storing concepts, relationships and instances in the accommodation domain, to identify aspects related to the hotel system. Syntactic dependency analysis is used to identify opinionated terms related to aspects. Furthermore, the study incorporated the use of several sentiment lexicons (e.g., AffectPT-br, EmoLex, LeIA, LIWC2007pt, Onto.PT, OpLexicon, Reli-Lex, SentiLex-PT, SentiWordNet-PT-BR, UNILEX, WordNetAffectBR), which are responsible for storing the polarity of words, and the use of EnsembleLex, an ensemble of the results of the four best lexicons. The tests were carried out to verify the feasibility of using spelling correction (manual and through the autocorrect, pyspellchecker and LanguageTool libraries) and the use of the fashion of joining adjectives that were related to an aspect. The results demonstrated that EnsembleLex surpassed the results obtained by other lexicons, reaching a maximum of 0.461 for f-measure (macro) and 0.536 for f-measure (micro), using *reviews* with neutral annotations regarding the sentiment directed at aspects, the fashion of adjective union and the pyspellchecker spelling correction library.

Keywords: Sentiment Analysis. Syntactic Dependency Tree. Aspect. Ontology. Hontology.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABSA	*Aspect-Based Sentiment Analysis*
ABOX	*Assertional Box*
ACL	*adjectival clause (clausal modifier of noun)*
ADV	*adverb*
ADVCL	*adverbial clause modifier*
ADVMOD	*adverbial modifier*
ADJ	adjetive
AI	*Artificial Intelligence*
AMOD	*adjectival modifier*
ANEW	*Affective Norms for English Words*
AR	*Affective Reasoner*
AWHHE	*adjetives with high-human evidence*
AWLHE	*adjetives with low-human evidence*
BMIR	*Stanford Center for Biomedical Informatics Research*
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CC	*coordinating conjunction*
CCONJ	*coordinating conjunction*
CL	*Computational Linguistics*
CS	*Computer Science*
CONJ	*conjunct*
COP	*copula*
DEP	*Dependency*
DEP tags	*Dependency tags*
DET	*determiner*
DG	*Dependency Grammar*
DL	*Description Logic*

DM	*Data Mining*
DMF	*Digital Media Files*
DP	*Dependency Parser*
DSS	*Decision Support Systems*
DT	*Dependency Tree*
EmoLex	*NRC Word-Emotion Association Lexicon*
GI	*General Inquirer*
GT	*Google Translate*
HITs	*Human Intelligence Tasks*
HDP	*Hindi Dependency Parser*
HRL	*Hight Resources Languages*
HSWN	*Hindi SentiWordNet*
IDF	*Inverse Document Frequency*
IR	*Information Retrieval*
KB	*Knowledge Base*
LeIA	Léxico para Inferência Adaptada
LI	*Linguistic Items*
LIWC	*Linguistic Inquiry and Word Count*
LKBs	*Lexical Knowledge Bases*
LM	*Language Models*
LR	*Linguistics Resources*
LRL	*Low Resources Languages*
LSKB	*Lexical-Semantic Knowledge Base*
ME	*Maximum Entropy*
ML	*Machine Learning*
MT	*Microsoft Translator*
MWE	*Multi-Word Expressions*
NB	*Naïve Bayes*
NLP	*Natural Language Processing*
NOUN	*noun*
NSUBJ	*nominal subject*
NSUBJ:PASS	*passive nominal subject*
NRC	*National Research Council Canada*
OCC	Ortony, Clore e Collins

OpLexicon	*Opinion Lexicon*
OM	*Opinion Media*
OT	*Opinion Text*
OWL	*Web Ontology Language*
PAPEL	Palavras Associadas Porto Editora - Linguateca
PMI	*Pointwise Mutual Information*
PoS	*Part-of-Speech*
PoS tags	*Part-of-Speech tags*
qualiflemma	*qualified lemma*
ReLi	Resenhas de Livros
ReLi-Lex	ReLi *Lexicon*
RDF	*Resource Description Framework*
RDFS	RDF *Schema*
SA	*Sentiment Analysis*
SentiLex-PT	*Sentiment Lexicon for Portuguese*
SDA	*Syntactic Dependency Analysis*
SL	*Sentiment Lexicon*
SVM	*Support Vector Machine*
SW	*Semantic Web*
syngraph	*synonym graph*
synset	*synonym set*
TBox	*Terminological Box*
TEP	*Electronic Thesaurus for Brazilian Portuguese*
TF	*Term Frequency*
TF-IDF	*Term Frequency – Inverse Document Frequency*
UD	*Universal Dependencies*
UGC	*User-Generated Content*
UNILEX	*Unified Lexicon*
VADER	*Valence Aware Dictionary and sEntiment Reasoner*
VERB	*verb*
WE	*Wildcard Expansion*
WNP	WordNet de Princeton
WWW	*World Wide Web*
W3C	*World Wide Web Consortium*

XML	*eXtensible Markup Language*
XSD	XML *Schema Definition*

SUMÁRIO

1 INTRODUÇÃO

Com a popularização do uso da Internet, houve um aumento significativo na base de usuários de plataformas, como sites e aplicativos. Isso ocasionou uma ampliação expressiva no fluxo de informações que circulam nessas plataformas, devido à expansão da quantidade de dados coletados de seus usuários e maior interação entre os usuários de uma mesma plataforma.

A quantidade de dados disponível em uma plataforma cresce à medida que os seus usuários disseminam conteúdos na Internet. Esse tipo de dado é chamado de Conteúdo Gerado pelo Usuário (do inglês, *User-Generated Content* — UGC). O UGC refere-se a qualquer conteúdo criado (ou gerado) e distribuído pelo usuário, seja ele texto ou Arquivos de Mídia Digital (do inglês, *Digital Media Files* — DMF), também chamados de Arquivos de Mídia, como áudio, imagem e vídeo (GEORGE; SCERRI, 2007; FRIDRICH, 2009).

O UGC tornou-se uma valiosa fonte de dados, pois as plataformas podem armazená-los em grande volume e, através da análise desses dados, podemos descobrir vários tipos de informações relevantes, como idade e densidade de clientes em uma determinada região ou a taxa de aprovação de um produto ou serviço.

Reviews (resenhas), *ratings* (avaliações) ou *comments* (comentários) são tipos de conteúdos criados (ou gerados) e distribuídos por usuários através de plataformas, tais como sites e aplicativos de redes sociais (exemplo: Facebook e X), de mensagens instantâneas (exemplo: WhatsApp, Discord e Telegram), de fóruns (exemplo: Disqus e Reddit), de vídeos (exemplo: Youtube, Vimeo e TED Talks), de avaliações de hotéis (exemplo: Trivago, Booking.com e TripAdvisor) e de comércio eletrônico (exemplo: Amazon, MercadoLivre, Alibaba, AliExpress, e eBay), etc.

Texto de Opinião (do inglês *Opinion Text* — OT), também chamado de Texto Opinativo, é uma categoria de texto na qual os indivíduos expressam suas opiniões e pontos de vista sobre determinados assuntos (LIU, 2012; KUMAR; GUPTA, 2021). *Reviews*, *ratings* ou *comments* são tipos de textos que podem ser caracterizados como OT, pois, são caracterizados pela expressão da disposição pessoal dos usuários em relação a produtos ou serviços específicos, ou ainda, em relação a pessoas ou lugares. Já

os DMF que expressam opinião são classificados como Mídia Opinativa (do inglês, *Opinion Media* — OM).

Segundo Liu (2020), o termo 'opinião' consiste em um conceito amplo que abrange sentimentos, avaliações, apreciações ou atitudes e informações associadas, tais como o alvo da opinião e o detentor da opinião (pessoa ou organização que detém a opinião). Além disso, Liu (2020) caracteriza o termo 'sentimento' como sensação positiva ou negativa subjacente implícita na opinião, indicando que o sentimento não é sempre explícito ou diretamente expresso, mas que pode ser inferido a partir do tom utilizado ou do conteúdo da opinião.

A análise da opinião expressa pelos usuários é de extrema importância, pois, segundo Liu (2012), "as opiniões são o centro para quase todas as atividades humanas e são as principais influenciadoras de nossos comportamentos". Dessa forma, a partir da análise da opinião podemos descobrir informações que podem auxiliar no desenvolvimento de novos produtos ou serviços, melhorar os produtos ou serviços existentes ou influenciar a perspectiva de futuros clientes - uma opinião positiva pode fazer com que um usuário compre aquele produto ou contrate aquele serviço, já uma opinião negativa fará com que o usuário busque alternativas melhores.

A área que desenvolve as abordagens responsáveis por analisar e extrair o sentimento expresso em um texto opinativo é chamada de Análise de Sentimento (do inglês, *Sentiment Analysis* — SA). A SA é realizada sobre texto ou transcrição de áudios ou vídeos; ou sobre imagens e vídeos.

A SA é responsável por classificar o sentimento que está atrelado a uma opinião (como, uma avaliação de um produto), tendo em vista emoção (por exemplo: felicidade, tristeza, medo e raiva), polaridade (por exemplo: positivo e negativo) ou intensidade (por exemplo: muito positivo e muito negativo) de um aspecto (por exemplo: preço e durabilidade de um produto) (LIU, 2020).

Nos últimos anos, a SA tem sido utilizada em diversos domínios, desde comércio, saúde, turismo, hotelaria e serviços financeiros até eventos sociais e eleições políticas. A SA tem ganhado tanto destaque, que empresas iniciantes e grandes corporações estabelecidas, como Google, Microsoft, Hewlett-Packard, Amazon, eBay, SAS, Oracle, Adobe, Bloomberg e SAP, construíram ou estão em processo de construção de espaços relacionados com SA (LIU, 2020).

Liu (2020) desenvolveu o *Opinion Parser*, um sistema dedicado à SA, e atuou em projetos destinados a clientes abrangendo mais de quarenta domínios, dentre eles destacam-se os setores: alimentício; automotivo; construção e decoração; eletrodomésticos; eletrônicos; entretenimento; esportes; farmacêutico; financeiro; gastronomia; moda; móveis; político; saúde e bem-estar; sustentabilidade; tecnologia; turismo e hoteleiro. Contudo, a elaboração do *Opinion Parser* ocorreu, em sua maior parte, devido à abundância de recursos disponíveis para o seu desenvolvimento.

O intuito desse trabalho é avaliar o emprego de analisadores de dependência sintática em abordagens baseadas em léxico para Análise de Sentimento Baseado em Aspecto. Foi realizado uma adaptação da pesquisa desenvolvida por Freitas (2015). Nele, a autora desenvolveu um método para SA aplicado a *reviews* de hotéis escritos em português brasileiro, sob o nível de aspecto usando diferentes léxicos de sentimentos e a ontologia de domínio Hontology. O diferencial do presente trabalho é o uso da análise de dependência sintática na identificação de termos opinativos relativos aos aspectos encontrados em *reviews* de hotéis. Para isso, foi utilizado o analisador de dependência sintática da biblioteca spaCy (HONNIBAL et al., 2020).

Além dos léxicos de sentimentos utilizados por Freitas (2015) (LIWC2007pt (BALAGE FILHO; PARDO; ALUÍSIO, 2013), Onto.PT (OLIVEIRA; GOMES, 2014; OLIVEIRA; SANTOS; GOMES, 2014), OpLexicon (SOUZA et al., 2011; SOUZA; VIEIRA, 2012), SentiLex-PT (SILVA; CARVALHO; SARMENTO, 2012)), neste trabalho foram utilizados os seguintes léxicos de sentimentos: AffectPT-br (CARVALHO; SANTOS; GUEDES, 2018), EmoLex (MOHAMMAD; TURNEY, 2010, 2013), LeIA (ALMEIDA, 2018), ReLi-Lex (FREITAS et al., 2012), SentiWordNet-PT-BR (BASTOS, 2023), WordNetAffectBR (PASQUALOTTI; VIEIRA, 2008; PASQUALOTTI, 2015) e UNILEX (SOUZA; PEREIRA; DALIP, 2017).

Foram realizados testes empregando a moda da polaridade dos adjetivos e a correção ortográfica dos *reviews* de hotéis. Além disso, implementou-se o EnsembleLex, ensemble dos resultados dos quatro melhores léxicos. Os léxicos alcançaram um resultado máximo de 0,429 de medida-f (macro) e 0,508 de medida-f (micro) incluindo *reviews* com anotações neutras quanto ao sentimento dirigido aos aspectos, utilizando o léxico OpLexicon. O EnsembleLex superou os resultados obtidos com OpLexicon, alcançando um máximo de 0,461 para medida-f (macro) e 0,536 para medida-f (micro).

O restante deste texto está organizado da seguinte forma: No Capítulo 2, apresentamos o referencial teórico deste trabalho. No Capítulo 3, apresentamos os recursos linguísticos utilizados neste trabalho. No Capítulo 4, apresentamos trabalhos relacionados com a área de análise de sentimento. No Capítulo 5, apresentamos as etapas de desenvolvimento do sistema proposto neste trabalho. No Capítulo 6 apresentamos os resultados e sua discussão. E por fim, no Capítulo 7 apresentamos as considerações finais e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Análise de Sentimento

Como discutido anteriormente, como grande crescimento da Internet uma quantidade massiva de dados não estruturados é produzida dia a dia, muitos deles apresentados em forma textual. Além disso, muitos desses textos possuem opiniões sobre produtos ou serviços.

O Processamento da Língua Natural (do inglês, *Natural Language Processing* — NLP) é uma área da Linguística Computacional (do inglês, *Computational Linguistics* — CL) focada na construção de *softwares*, aplicativos e sistemas computacionais capazes de interpretar e/ou gerar informações em linguagem natural, como tradutores automáticos, *chatterbots*, *parsers*, reconhecedores automáticos de voz, geradores automáticos de resumos, etc. (OTHERO, 2006; RUSSELL; NORVIG, 2004).

A Análise de Sentimento (do inglês, *Sentiment Analysis* — SA) é uma sub-área do NLP que visa auxiliar no processo de tomada de decisões. Segundo Liu (2012), a SA é o campo de estudo que analisa as opiniões, sentimentos, avaliações, apreciações, atitudes e emoções das pessoas em relação a entidades e seus atributos, expressos em texto escrito, imagem, áudio ou vídeo. Essas entidades podem incluir produtos, serviços, organizações, indivíduos, eventos, questões ou tópicos.

A SA representa um campo de estudo com amplo espaço para problemas e desafios. Existem várias terminologias e tarefas ligeiramente distintas associadas a este campo, como análise de sentimentos, mineração de opiniões, análise de opiniões, extração de opiniões, mineração de sentimentos, análise de subjetividade, análise de afeto, análise de emoção e mineração de avaliações. Todas essas variantes, embora tenham suas peculiaridades, agora estão agrupadas sob o grande guarda-chuva da SA (LIU, 2020).

Segundo Cambria et al. (2013) e Taboada et al. (2011), a SA pode ser classificada levando em consideração o método utilizado para realizar a avaliação:

- **Palavras-chave e afinidade léxica:** Métodos baseados em palavras-chave utilizam a presença de palavras sem sentido ambíguo, tais como “feliz”, “triste”,

"medo" e "raiva" para classificar uma *review* ou utilizam um recurso léxico, chamado de léxico de sentimento, como o SentiWordNet (BACCIANELLA; ESULI; SEBASTIANI, 2010), que fornecem informações referente à afinidade entre palavras com sentimento;

- **Aprendizado de máquina:** Métodos baseados em aprendizado de máquina utilizam modelos de aprendizado de máquina, como *Naïve Bayes* (NB) e *Support Vector Machine* (SVM), para realizar a classificação da opinião, aprendendo, por exemplo, com a importância e a frequência das palavras;
- **Orientação semântica:** Métodos baseados em orientação semântica utilizam o cálculo da orientação semântica de uma palavra para realizar a classificação da opinião, baseando-se na coocorrência da palavra analisada com palavras que possuem a mesma orientação;
- **Conceitos:** Métodos baseados em conceitos utilizam ontologias ou redes de palavras-chave (expressões) para realizar a classificação da opinião, analisando expressões que não possuem uma emoção explícita, mas relacionam-se com um sentimento implicitamente.

Além disso, a SA pode ser classificada como híbrida, caso sejam utilizados modelos de abordagens diferentes ao mesmo tempo, por exemplo, misturando léxico de sentimento e aprendizado de máquina. Modelos híbridos tendem a ter uma melhor eficiência se comparados com modelos mais simples, em razão de que os resultados dos modelos que se complementam, fazendo com que um modelo substitua as falhas do outro, tornando-o mais fiável.

Por outro lado, também podemos classificar as abordagens de SA de acordo com o nível de granularidade das porções de texto analisadas pela abordagem. Liu (2012) sugere a seguinte categorização:

- **Nível de Documento:** Métodos desenvolvidos em nível de documento realizam a análise de um documento completo, atribuindo um único sentimento a todo o documento. Classificação em nível de documento leva em consideração o sentimento geral expresso na opinião avaliada. Este nível de análise assume que cada documento expressa opiniões sobre uma única entidade, como, por exemplo, um único produto. Dessa forma, esse tipo de análise não é aplicável a documentos que avaliam ou comparam múltiplas entidades.
- **Nível de Sentença:** Métodos desenvolvidos em nível de sentença realizam a análise de sentenças, classificando cada sentença individualmente, como tendo um sentimento positivo, negativo ou neutro - geralmente significando que não expressa uma opinião. Classificação em nível de sentença está relacionado a

classificação subjetividade, distinguindo sentenças que expressam opiniões de sentenças que não expressam opiniões.

- **Nível de Entidade e Aspecto:** Análises baseadas em nível de documento ou baseados em nível de sentença não são capazes de descobrir o que exatamente o que as pessoas gostam ou não gostam, mas através de métodos baseados em nível de entidade e aspecto é possível realizar uma análise mais refinada. Ao invés de olhar para a construção da linguagem (documentos, parágrafos, sentenças, cláusulas ou frases), métodos baseados em nível de entidade e aspecto olham diretamente para a opinião em si, baseando-se no fato de que uma opinião é composta por um sentimento (positivo ou negativo) e um alvo ao qual a opinião se refere (entidade e aspecto).

Já a apresentação dos dados pode levar em consideração:

- **Polaridade:** Métodos baseados em polaridade possuem a vantagem de ter agilidade na análise e maior consistência na análise manual e automática, mas, possuem a desvantagem de ter pouca profundidade na análise. Pode ser classificado como variações de positivo e negativo, e podem possuir o neutro, por exemplo: positivo, negativo e neutro;

- **Humor:** Métodos baseados em humor possuem a vantagem de ter maior percepção nos sentimentos predominantes, mas, possuem a desvantagem de ter inconsistência na análise e ser uma classificação subjetiva. Pode ser classificado como variações de humor, por exemplo: felicidade, tristeza, medo e raiva;

- **Escala:** Métodos baseados em escala possuem a vantagem de ter um melhor entendimento de nível de satisfação/insatisfação, mas, possuem a desvantagem de ter inconsistência na análise e ser uma classificação subjetiva. Pode ser classificado como, por exemplo: péssimo, ruim, regular, bom e ótimo.

2.1.1 Análise de Sentimento Baseada em Aspecto

Na Análise de Sentimento Baseada em Aspecto (do inglês, *Aspect-Based Sentiment Analysis* — ABSA) buscamos identificar os sentimento expressos em relação aos aspecto encontrados em uma *review*. Aspectos consistem em conceitos ou características de produtos ou serviços, como durabilidade e preço.

A metodologia de uma ABSA consiste em quatro etapas:

1. **Pré-processamento:** Essa etapa consiste em aplicar processos de pré-processamento e tratamento dos dados, que contribuem para a organização e refino dos dados. Tal procedimento é realizado com o intuito de agilizar a análise dos dados. Dessa forma, são realizados processos que trazem clareza para

os dados e auxiliam na redução do volume dos dados, como correção de erros ortográficos, remoção de *stopwords*[1], tokenização[2], lematização[3] e stemização[4].

2. **Identificação das características:** Essa etapa consiste em identificar os aspectos das *reviews*. Isso pode ser feito utilizando uma lista pré-definida ou algoritmos que buscam aspectos correlacionados com os produtos ou serviços analisados;

3. **Identificação da Polaridade:** Essa etapa consiste em identificar a polaridade das opiniões pré-processadas e que contêm aspectos. Isso pode ser realizado utilizando léxicos de sentimento ou aprendizado de máquina.

4. **Pós-processamento:** Essa etapa consiste em aplicar processos de pós-processamentos, que contribuem para minimizar os erros das respostas e melhorar a precisão dos resultados (RIBEIRO, 2015). Dessa forma, caso algum advérbio de negação seja encontrado na *review*, podem ser realizados processos que alterem a polaridade da *review*.

5. **Sumarização dos Resultados:** Essa etapa consiste em apresentar as polaridades referente aos aspectos encontrados na *review* analisada.

2.1.1.1 Extração de Termo de Aspecto

A Extração de Termo de Aspecto é uma técnica importante na ABSA, sendo utilizada para identificar e extrair aspectos mencionados em *reviews*. Essa técnica possibilita uma análise mais detalhada dependendo da profundidade com que se observa os aspectos, sendo eles do tipo explícito ou implícito.

Aspectos explícitos são aspectos que são referenciados explicitamente no texto opinativo. Em contrapartida, aspectos implícitos não são referenciados explicitamente no texto opinativo, mas através de uma referência indireta a ele (CRUZ; GELBUKH; SIDOROV, 2014; AYE MAR; SHIRAI, 2022; XU et al., 2023; MAR; SHIRAI; KERT-KEIDKACHORN, 2023; CAI et al., 2023).

[1] A remoção de *stopwords* consiste em remover palavras que são frequentemente utilizadas em textos e que não agregam informações relevantes ao modelo (como: a, o, e, que, para, quando, etc.).

[2] A tokenização consiste em dividir um texto em unidades menores, chamadas de *tokens*. Os *token* podem ser desde palavras, números e espaços em branco (que ocorrem quando existe um espaçamento duplo entre *tokens*), até símbolos ($\$@\#\&*$) ou sinais de pontuação ($,.;:?!$). A tokenização é um processo fundamental em muitas tarefas de NLP, pois a transformação de um texto extenso em fragmentos menores (*tokens*) torna mais fácil a análise.

[3] A lematização consiste na redução de uma palavra para sua forma base ou lema (do inglês, *lemma*). Dessa forma, verbos são representados através do seu infinitivo. Já substantivos e adjetivos são representados através do seu masculino singular (COSTE; GALLISON, 1983).

[4] A stemização consiste na redução de uma palavra para a raiz (radical) ou, como é chamado na técnica, ao seu *stem*. Dessa forma, ocorre a remoção de sufixos e prefixos que indicam variações morfológicas ou flexionais das palavras (PORTER, 1980; BARION; LAGO, 2008).

Nas frases “O **preço** do restaurante é muito elevado.” e “A **medicação** agiu rápido, tirando minha dor.”, ‘**preço**’ e ‘**medicação**’ são considerados aspectos explícitos, pois os aspectos são referenciados na frase.

Em contrapartida, nas frases “A refeição foi muito *cara*.” e “Após a *injeção* minha dor sumiu.”, ‘*cara*’ e ‘*injeção*’ são considerados aspectos implícitos, uma vez que, o aspecto ao qual estão se referindo não está explícito na frase. ‘*Cara*’ é uma referência implícita ao aspecto ‘**preço**’, enquanto que ‘*injeção*’ é uma referência implícita ao aspecto ‘**medicação**’.

2.1.1.2 Determinação da Orientação de Sentimento

A Determinação da Orientação de Sentimento é um processo crucial na ABSA, que consiste em classificar as opiniões expressas em textos como positivas, negativas ou neutras. Esta técnica é amplamente utilizada para compreender as reações e atitudes do público em relação a produtos, serviços ou tópicos diversos. Por exemplo, ao analisar comentários em redes sociais, a determinação da orientação de sentimento ajuda a identificar a percepção geral do usuário sobre uma marca ou evento. Além disso, é uma ferramenta valiosa para empresas e organizações na tomada de decisões estratégicas, permitindo ajustes em produtos ou campanhas com base nas emoções e opiniões dos consumidores em relação aos seus produtos e serviços.

A Figura 1 mostra um exemplo de conteúdo não-estruturado coletado de uma *review* de hotel.

... tem uma excelente **vista para a praia**... o **cheiro do banheiro** estava terrível ... o **quarto** era arrumado e cheiroso ... a **sala** era bem espaçosa ... o **preço** era acessível

Figura 1 – Conteúdo não-estruturado. Fonte: Autoria Própria.

A Determinação da Orientação de Sentimento consiste na Sumarização dos Resultados através da análise de diferentes estruturas linguísticas. A sumarização dos resultados ocorre, através da soma das intensidades das propriedades relacionadas aos aspectos, como pode ser visto na Figura 2.

Figura 2 – Primeira forma de Sumarização dos Resultados. Fonte: Autoria Própria.

A Sumarização dos Resultados também pode ser apresentado em função da média aritmética arredondada dos resultados obtidos em cada aspecto, como pode ser visto na Figura 3.

Figura 3 – Segunda forma de Sumarização dos Resultados. Fonte: Autoria Própria.

Os *scores* (pontuações) dos resultados da Sumarização dos Resultados também podem ser apresentados através de uma escala na forma textual utilizando a seguinte conversão:

- $score =$ ★★★★★ $=>$ *review* neutra
- ★★★★★ $> score \leq$ ★★★★★ $=>$ *review* extremamente negativa
- ★★★★★ $> score \leq$ ★★½★★ $=>$ *review* ligeiramente negativa
- ★★½★★ $> score \leq$ ★★★★★ $=>$ *review* ligeiramente positiva
- ★★★★★ $> score \leq$ ★★★★★ $=>$ *review* extremamente positiva

A escala textual também pode ser resumida através da ocultação dos advérbios de intensidade (ligeiramente e extremamente) como pode ser observado na escala a seguir:

- $score =$ ★★★★★ $=>$ *review* neutra
- ★★★★★ $> score \leq$ ★★½★★ $=>$ *review* negativa
- ★★½★★ $> score \leq$ ★★★★★ $=>$ *review* positiva

2.2 Recursos Linguísticos

Durante o nosso dia a dia, os Recursos Linguísticos (do inglês, *Linguistics Resources* — LR), são utilizados para entender, compreender e expressar ideias em uma determinada língua (CHAPELLE, 2020). Como exemplo de LR utilizados na nossa língua podemos citar palavras, frases e estruturas (CHAPELLE, 2020). As estruturas são separadas em Estrutura Linguística e Estrutura Extralinguística (KOCH, 2000).

A Estrutura Linguística refere-se à organização interna do texto e abrange a coesão, coerência e tipologia textual. A tipologia textual está relacionada à natureza linguística do texto desenvolvido, sendo determinada pelos tipos textuais, como narração, argumentação, exposição e descrição. A coesão e a coerência são responsáveis, respectivamente, por unir os elementos do texto por meio de conectivos (também conhecidos como conjunções ou palavras de ligação, como “e”, “ou” e “mas”) e estabelecer significado e interpretabilidade ao texto, criando uma conexão entre as palavras e as frases.

Enquanto que, a Estrutura Extralinguística corresponde ao contexto envolvido no processo de escrita. Esse processo refere-se ao uso do conhecimento de mundo dentro do texto, estabelecendo as ideias necessárias para imputar significado às abstrações do escritor, trazendo, assim, significação ao texto.

Dessa forma, os LR compreendem desde o conhecimento a respeito de palavras presentes em um dicionário (como grafia, classe gramatical e significado), até o nosso conhecimento de mundo (como o entendimento sobre técnicas ou conhecimentos específicos utilizados em uma determinada área) (KOCH, 2000). Por outro lado, na área de NLP, os LR são descritos como Conjuntos de Dados e Bases de Conhecimento Linguístico, sendo usados para apoiar pesquisas e aplicações (BRANCO et al., 2012; IDE; ROMARY, 2004). Da mesma forma que precisamos de uma variedade de recursos linguísticos para entender, compreender e expressar sentimentos e opiniões, os modelos de NLP também precisam desses recursos para sintetizar informações (CHAPELLE, 2020).

LR podem ser encontrados na literatura pelos sinônimos de Recursos de Linguagem (do inglês, *Language Resources*) e Recursos Lexicais (do inglês, *Lexical Resources*) (KOCH, 2000; MANNING; SCHUTZE, 1999; NUNES et al., 2005; OLTRAMARI et al., 2013; KRAUWER, 2003; IDE; ROMARY, 2004; BRANCO et al., 2012).

Segundo Oltramari et al. (2013), LR podem ser divididos em dois grupos:

- O primeiros grupo refere-se aos recursos léxico-semânticos, recursos que fornecem informações sobre lexemas[5] e sua relação com outros lexemas, como dicionários, ontologias, bases de conhecimento semântico, etc.
- O segundo grupo refere-se aos recursos de Corpus anotados, que correspondem as coleções de dados anotados com características linguísticas. Tais corpora podem assumir a forma escrita, falada ou gestual.

Nunes et al. (2005) também divide os LR em dois grupos, mas, leva em consideração a realização do processamento do texto:

- O primeiros grupo refere-se aos recursos que oferecem conhecimento linguístico, mas não realizam o processamento do texto, como os dicionários, Corpus e Thesaurus;
- O segundo grupo refere-se aos recursos que realizam o processamento do texto, como analisadores morfológicos, sintáticos (*Parsers*) e semânticos ou etiquetadores morfossintáticos (*Taggers*).

2.2.1 Itens Linguísticos

Itens Linguísticos (do inglês, *Linguistic Items* — LI) tem como função contrastar com outras unidades da língua, desempenhando o papel de descrever não a soma de todos os elementos existentes no mundo, mas descrever um tipo de elemento que se diferencia dos demais, trazendo assim um tipo de representação (significado ou função), ou seja, LI são termos genéricos que se referem a qualquer unidade da língua que carrega significado ou função. Dessa forma, o significado da palavra 'cão' não representaria a soma de todos os cães possíveis no mundo, mas, seria utilizado como diferenciação de outros tipos de animais, como gato, cavalo, etc. (JEFFRIES, 2006). Exemplo de LI, incluem palavras, frases, expressões idiomáticas, gírias ou elementos de linguagem, como *hashtags* ou *emoticon*.

2.2.2 Ontologia de Domínio

Para que ocorra o desenvolvimento em uma área é essencial que pessoas, organizações e sistemas de software possam se comunicar entre si. No entanto, a diversidade nas necessidades e nos contextos, pode resultar em divergência nos pontos de vista, ocasionando em suposições diferentes sobre o que é essencialmente o

[5] O lexema é considerado como lema ou unidade lexical, ou seja, palavra ou termo que tem significado por si só, representam um conjunto de palavras que compartilham um significado básico comum, mas podem ter formas diferentes devido a inflexões morfológicas (flexão em gênero e número). Por exemplo, "correr" é um lexema que representa todas as formas conjugadas do verbo, como "corro", "corres", "correu", "correndo", "corria", etc.

mesmo assunto. Cada cultura pode ter os seus próprios dialetos, assim como conceitos, estruturas e métodos diferentes, sobrepostos e/ou incompatíveis. Dessa forma, a ausência resultante de um entendimento compartilhado provoca uma comunicação deficiente (USCHOLD; GRUNINGER, 1996).

Segundo Uschold; Gruninger (1996), para resolver esse problema é necessário reduzir ou eliminar a confusão conceitual e terminológica através de um consenso comum, pois, desse modo, atuará como um elo integrador e estruturante para as diversas perspectivas, trazendo os seguintes benefícios: reutilização, confiabilidade e especificação.

Ainda, de acordo com os autores, o termo ontologia se refere “ao entendimento compartilhado de algum domínio de interesse que pode ser usado como uma estrutura unificadora para resolver os problemas com confusão conceitual e terminológica”. Já Arp; Smith; Spear (2015) define ontologia como “um artefato representacional, compreendendo como parte própria uma taxonomia, cujas representações pretendem designar alguma combinação de universais, classes definidas e certas relações entre eles”.

Na visão de Uschold; Gruninger (1996) a ontologia seria como o entendimento mútuo de conhecimentos dentro de um certo campo de estudo. Por outro lado, na visão de Arp; Smith; Spear (2015), a ontologia seria considerada uma estrutura com uma hierarquia, baseada em termos que denotam classes ligados por relações de subtipos (da mesma forma que acontece na biologia, com o o uso de filos para separar o reino animal), deliberadamente concebida para atender ao propósito de gerar uma representação visual do conhecimento (conceitos, suas propriedades e relações) de um determinado domínio.

Assim, temos que as ontologias são utilizadas porque fornecem um modelo de representação formal e estruturado do conhecimento, além de possuírem a vantagem de serem reutilizáveis. Para um melhor entendimento sobre ontologia de domínio, consulte o Apêndice A.

2.2.2.1 Base de Conhecimento

Na Lógica Descritiva (DL — do inglês *Description Logic*), conforme descrito por Baader (2003) e Baader et al. (2017), o conhecimento normalmente é dividido em dois componentes, Caixa Terminológica (TBox — do inglês *Terminological Box*), também chamada de Caixa Taxonômica (do inglês *Taxonomic Box*), e Caixa Assercional (ABox — do inglês *Assertional Box*), sendo a combinação desses dois componentes chamada de Base de Conhecimento (KB — do inglês *Knowledge Base*).

A TBox representa o conhecimento relativo à estrutura de um domínio (semelhante a um esquema de banco de dados) e é construída através de declarações que descrevem propriedades gerais dos conceitos. Enquanto que ABox representa o conheci-

mento sobre uma situação concreta (semelhante a uma instância de banco de dados) e é construída através do conhecimento extensional (também chamado de conhecimento assertivo), que representa o conhecimento que é específico dos indivíduos do domínio do discurso.

O conhecimento pode ser considerado intensivo ou extensivo. O conhecimento intensivo é geralmente considerado atemporal, ou seja, é o conhecimento que é pensado para não mudar. Por outro lado, o conhecimento extensivo é geralmente considerado contingente ou dependente de um único conjunto de circunstâncias e, portanto, sujeito a mudanças ocasionais ou mesmo constantes.

Como exemplo de declarações TBox, sobre o domínio universitário, podemos citar, por exemplo, que um professor é uma pessoa que ministra um curso, um aluno é uma pessoa que frequenta um curso e os alunos não ensinam. Em contrapartida, declarações ABox, sobre o mesmo domínio, podemos citar, por exemplo, que Mary é uma pessoa, CS600 é um curso e Mary ensina CS600.

2.2.3 Sistema de Análise de Dependência Sintática

Para saber a qual conceito um sentimento está se referindo são utilizadas as informações extraídas de uma Análise de Dependência Sintática (do inglês, *Syntactic Dependency Analysis* — SDA) (MEL'CUK et al., 1988; DE MARNEFFE; NIVRE, 2019). A SDA tem como objetivo obter as relações hierárquicas entre as palavras de uma frase, ou seja, ela realiza um processo de identificação das relações de dependência (SOUZA, 2023). Essas relações de dependência formam a estrutura sintática da frase, sendo descrita apenas em termos de relações gramaticais binárias direcionadas entre as palavras. Essa estrutura forma uma espécie de árvore, chamada de Árvore de Dependência (do inglês, *Dependency Tree* — DT) (JURAFSKY; MARTIN, 2023).

A Figura 4 apresenta um exemplo da representação de uma DT, na qual as relações de dependência são representadas pelos arcos orientados e os rótulos dessas relações são exibidos em azul.

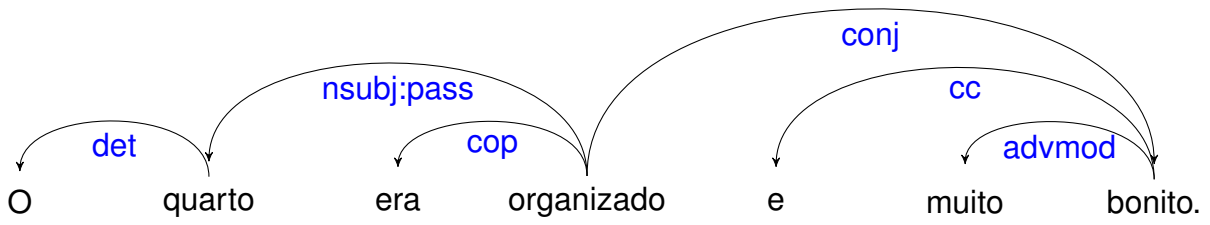

Figura 4 – Representação da Árvore de Dependência para a sentença "O quarto era organizado e muito bonito.".

Cada elemento de uma frase é denominado *token*, os *tokens* podem ser desde palavras e espaços em branco (que ocorrem quando existe um espaçamento duplo entre

as palavras), até símbolos ($@#&∗) ou sinais de pontuação (, .; :?!). Além disso, os *tokens* também podem ser classificados como *spans* ou entidades nomeadas. *Spans* são considerados um agrupamento de *tokens*,e podem ser úteis para representar expressões idiomáticas[6]. Já entidades nomeadas podem ser um local, organização, pessoas, etc.

As relações entre as palavras são ilustradas acima da frase por arcos direcionados e rotulados que partem dos *head tokens* (*tokens* de cabeça) até os seus dependentes, também chamados de *child tokens* (*tokens* filho). Essa estrutura de relações é chamada de Estrutura de Dependência Tipificada, pois os rótulos são extraídos de um inventário fixo de relações gramaticais (JURAFSKY; MARTIN, 2023).

O topo dessa estrutura é constituído pelo *root* (nó raiz), marcando explicitamente a raiz da árvore, ou seja, a partir do nó raiz é que todos os outros nós se originam (JURAFSKY; MARTIN, 2023). Dessa forma, cada frase (ou sentença) terá somente uma raiz, como mostrado Figura 5.

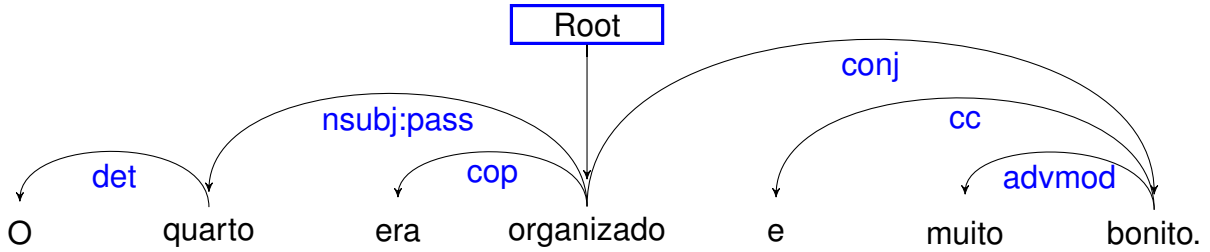

Figura 5 – Representação do *root* da Árvore de Dependência da sentença “O quarto era organizado e muito bonito.”.

O *root* de uma DT fica mais evidente ao transformá-la em um Gramática de Dependência (do inglês, *Dependency Grammar* — DG). Em uma DG, apenas as relações de dependência entre as ocorrências de unidades linguísticas em expressões compostas são descritas (PAGANI, 2015). Portanto, as abordagens baseadas em DG podem abstrair um pouco mais das informações relativas à ordem das palavras (JURAFSKY; MARTIN, 2023). A Figura 6 apresenta um exemplo da representação de uma DG, onde as relações de dependência entre palavras são indicadas pelas arestas que conectam os nós da árvore.

Na Figura 6, onde temos a representação da DG da sentença “O quarto era organizado e muito bonito.”, a palavra “organizado” representa o núcleo central da sentença (***Root***). Deste núcleo, dependem diretamente os seus filhos, as palavras “era”, “quarto”, e “bonito”. “Era” funciona como cópula (**cop**), ligando o sujeito ao predicado. “Quarto” é o sujeito passivo (**nsubj:pass**) da sentença, com “o” atuando como determinante (**det**), especificando o sujeito. Por outro lado, “bonito” é um adjetivo em

[6]Expressões multipalavras que tem um significado específico diferente de quando analisadas separadamente. Por exemplo: grande homem, à céu aberto, abrir o jogo, etc.

conjunção (**conj**) com “organizado”, ampliando a descrição do sujeito, e é modificado pelo advérbio “muito” (**advmod**), que intensifica seu significado. Além disso, a conjunção “e” (**cc**) conecta os adjetivos “organizado” e “bonito”. Assim, nesta árvore, cada palavra é ligada ao núcleo “organizado”, refletindo a distribuição dos papéis temáticos e a organização gramatical da sentença, onde os elementos modificadores e relacionais dependem dos núcleos aos quais estão associados.

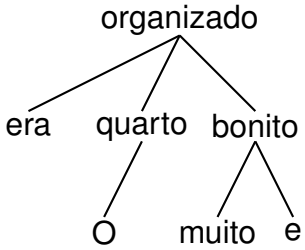

Figura 6 – Representação da Gramática de Dependência para a sentença “O quarto era organizado e muito bonito.”.

A base para as relações binárias que compõem estas estruturas de dependência são fornecidos pela noção linguística tradicional da relação gramatical, sendo os seus argumentos um *token* de cabeça e um *token* filho. O *token* de cabeça desempenha o papel de elemento central, a partir dele que irá surgir os outros ramos (também chamados de descendentes). Já o *token* filho desempenha o papel de modificador. A relação entre *token* de cabeça e *token* filho estabelece explicitamente a ligação das palavras que são imediatamente dependentes deles (JURAFSKY; MARTIN, 2023).

Além de estabelecer os pares cabeça-filho, as gramáticas de dependência desempenham o papel de classificar os tipos de relações gramaticais que os descendentes exercem sobre o seu *token* de cabeça. As relações gramaticais também podem ser chamadas de funções gramaticais (JURAFSKY; MARTIN, 2023).

Os linguistas desenvolveram taxonomias de relações que vão muito além das noções familiares de sujeito e objeto (direto e indireto). Embora haja uma variação considerável entre as teorias, a quantidade de elementos compartilhados propiciou para que padrões interlinguísticos tenham sido desenvolvidos. Podemos citar, como exemplo, o projeto intitulado Dependências Universais (do inglês, *Universal Dependencies* — UD), de Marneffe et al. (2021). Esse projeto foi uma iniciativa colaborativa pública para anotar dependências e outros aspectos da gramática em mais de 100 idiomas, fornecendo um inventário de 37 relações de dependência (MARNEFFE et al., 2021).

Em tese, as relações mais utilizadas podem ser divididas em dois conjuntos: relações clausais e relações modificadoras. As relações clausais descrevem papéis sintáticos em relação a um predicado (geralmente um verbo). Já as relações modificadoras categorizam as maneiras pelas quais as palavras podem modificar seus *token* de cabeça (JURAFSKY; MARTIN, 2023).

2.2.4 Léxicos de Sentimento

Segundo Liu (2020), os termos opinativos são os indicadores mais importantes de sentimento. Eles são constituídos por LI que que carregam sentimento (positivo ou negativo), como amor e “trem bom”, que representam LI com sentimento positivo.

Os Léxicos de Sentimentos (do inglês, *Sentiment Lexicon* — SL) (ou léxicos de opinião) são recursos responsáveis por armazenar uma lista de LI, como palavras e expressões idiomáticas, junto com a sua polaridade (ou valência) (LIU, 2020). Os SL são muito importantes, pois são utilizados durante a etapa de identificação da polaridade. Dessa forma, caso algum dos SL utilizados não seja adequado, seja do ponto de vista de cobertura lexical, em razão do recurso não contemplar um grande volume de LI, seja do ponto de vista de formalização, em razão do processo de desenvolvimento do recurso não foi realizado com a rigorosidade necessária para inferir a polaridade mais correta para os LI, isso pode afetar a qualidade do processamento em fases subsequentes (NUNES et al., 2005). Ao longo dos anos, os pesquisadores desenvolveram diversos métodos para construir SL. Segundo Liu (2020), existem três abordagens para compilar termos opinativos:

- **abordagem manual**: A abordagem manual é trabalhosa e demorada, pois, o sentimento de cada termo é atribuído manualmente. Por essa razão essa abordagem geralmente é utilizada para verificar os resultados obtidos por abordagens automatizadas, pois essas abordagens cometem erros.
- **abordagem baseada em dicionários**: A abordagem baseada em dicionários é separada em um três etapas:
  - na primeira etapa, é coletado manualmente um pequeno conjunto de palavras com polaridades conhecidas (também chamada de lista de sementes);
  - na segunda etapa, é realizado uma consulta por sinônimos e antônimos em um repositório ou dicionário online utilizando a lista de sementes. As palavras encontradas são adicionadas a lista e o processo é repetido até terminar todos os termos da lista.
  - na última etapa, a abordagem manual é realizada.
- **abordagem baseada em corpus**: A abordagem baseada em corpus é separada em um três etapas:
  - na primeira etapa, é coletado o conjunto de palavras do corpus;
  - na segunda etapa, a polaridade dos termos é imputada baseando-se no sentido que é utilizada no corpus.
  - na última etapa, a abordagem manual é realizada.

Essa abordagem também pode ser utilizada para adaptar um léxico de uso geral para um domínio específico. Neste contexto, a polaridade de algumas palavras pode mudar. Embora essa abordagem também possa ser usada para construir um léxico de sentimentos de uso geral, se o corpus utilizado for muito grande e diversificado, a abordagem baseada em dicionário torna-se muito mais eficaz, pois um dicionário contém todas as palavras.

Além da classificação baseada nas abordagens manual, baseada em dicionários e baseada em corpus, os léxicos podem ser classificados de acordo com o método pelo qual a atribuição da polaridade foi desenvolvida. Esta classificação leva em consideração se o léxico foi elaborado manualmente, gerado automaticamente, criado por meio da tradução de outro recurso, ou desenvolvido através de métodos híbridos.

- **Léxicos de sentimentos criados manualmente:** Léxicos criados manualmente são recursos nos quais é utilizada uma abordagem manual para atribuir a polaridade de sentimento.
- **Léxicos de sentimentos criados automaticamente:** Léxicos criados automaticamente são recursos no qual a atribuição da polaridade de sentimento foi realizada de forma automática, seja através da utilização das informações obtidas em outros recursos ou através de abordagens automatizadas, onde a atribuição da polaridade de sentimento é feita através do uso de funções ou através de aprendizado de máquina.
- **Léxicos de sentimentos traduzidos:** Léxicos traduzidos são recursos que consistem na adaptação do seu conjunto de Itens Linguísticos para outra língua, por meio de métodos de tradução.
- **Léxicos de sentimentos híbridos:** Léxicos híbridos são recursos no qual é utilizado mais de um método de atribuição da polaridade de sentimento.

2.3 Linguagens de Muitos e de Poucos Recursos

Segundo Russell; Norvig (2004), o problema de compreender a linguagem é consideravelmente complexo, em razão de que a compreensão da linguagem exige a compreensão do assunto e do contexto ao qual está inserida, não apenas a compreensão da estrutura das frases.

Dessa forma, precisamos de recursos que possam auxiliar pesquisas e aplicações na área de NLP. Esses recursos são denominados LR e são utilizados durante todas as etapas de treinamento e teste de modelos de SA, comumente conhecido como Modelos de Linguagem (do inglês, *Language Models* — LM) (IDE; ROMARY, 2004; HAPKE; HOWARD; LANE, 2019).

As linguagens dos LM podem ser classificadas em: Linguagens de Muitos Recursos (do inglês, *Hight Resources Languages* — HRL) e Linguagens de Poucos Recursos (do inglês, *Low Resources Languages* — LRL). Essa classificação leva em consideração a disponibilidade de LR.

HRL são linguagens em que os LR disponibilizados são demasiados. HRL possuem um grande volume de LR disponibilizados de maneira gratuita, isso ocorre não somente em razão da complexidade menor da língua, mas, também pelo compartilhamento maior de informações, através de grupos de pesquisa.

Em contrapartida, LRL são linguagens em que os LR disponibilizados são escassos, em razão de que a disponibilidade de LR é proporcional a complexidade da língua. Quanto maior a complexidade, maior é o tempo gasto no processamento dos dados e, consequentemente, do tempo de desenvolvimento do LR (GUPTA et al., 2021; CIERI et al., 2016).

Para aumentar a disponibilidade de LR nas LRL pode-se recorrer ao artifício da tradução de LR disponibilizados em outras línguas. No entanto, a tradução pode conter algum tipo de viés, como questões envolvendo gênero ou raça, ou envolvendo questões conceituais e culturais, resultando na mudança do sentido das palavras, influenciado pelo contexto de treinamento e desenvolvimento do tradutor (GOLDFARB-TARRANT; ROSS; LOPEZ, 2023).

LRL também podem ser encontradas na literatura através dos seguintes sinônimos: *Resources-Poor Languages* (Linguagens com Poucos Recursos), *Less-Resourced Languages* (Linguagens com Menos Recursos), *Scarce-Resource Languages* (Linguagens com Recursos Escassos) (GUPTA et al., 2021; CIERI et al., 2016; RANI; KUMAR, 2021; MEHMOOD et al., 2019; AKHTAR et al., 2016).

3 RECURSOS LINGUÍSTICOS UTILIZADOS

3.1 Ontologia de Domínio

Na literatura existem algumas ontologias desenvolvidas para o setor de alojamento, que pertence ao domínio de viagens ou turismo, como as ontologia Mondeca e Harmonet. No entanto, o vocabulário delas são limitadas, pois cobrem um pequeno conjunto de conceitos que muitas vezes descrevem o domínio a partir de diferentes perspectivas devido ao âmbito de aplicação restrito a partir do qual as ontologias foram extraídas (BARTA et al., 2009).

Segundo Chaves; Freitas; Vieira (2012), Hontology[1] é uma ontologia multilíngue para o setor de alojamento na indústria do turismo, desenvolvida inicialmente considerando um contexto de aplicações que processam avaliações sobre acomodações em sites da Web 2.0 e Sistemas de Suporte à Decisão (do inglês, *Decision Support Systems* — DSS). Tendo isso em vista, a Hontology busca atender os seguintes usuários: hóspede, cliente em potencial, gerente de alojamento, especialista em domínios, administrador de banco de dados, desenvolvedor de aplicativos e desenvolvedor de aplicativos. Ela é a ontologia de domínio utilizada para a criação do corpus anotado utilizado neste trabalho, e, portanto, também utilizada na etapa de identificação de aspectos nos textos opinativos.

Além disso, Hontology foi desenvolvida seguindo as diretrizes descritas por Chaves; Trojahn (2010):

1. Identificar ontologias existentes em domínios relacionados;
2. Selecionar os principais conceitos e propriedades;
3. Organizar conceitos e propriedades hierarquicamente em categorias;
4. Realizar a tradução manual da ontologia introduzindo os rótulos em português, espanhol e francês;

[1] O LR Hontology está disponível para *download* através do *link* `https://portulanclarin.net/repository/browse/hontology/a83c9d04cb7a11e1a404080027e73ea2359e10ea62b940109aabe03684aa5ea4/`.

5. Expandir conceitos e propriedades com base em avaliações online avaliadas manualmente;

6. Traduzir os novos conceitos e propriedades;

7. Exportar a ontologia em diversos formatos.

Especialistas do domínio atualizaram Hontology, incorporando o mapeamento de conceitos já existentes em outras ontologias. É possível encontrar categorias diferentes para uma mesma acomodação e, em alguns continentes, tipos específicos de acomodações, como Hotel de gelo e Hotel casa da árvore. Além disso, essa extensão, baseou-se em conceitos extraídos de (NET, 2009) e (CHAVES; GOMES; PEDRON, 2012), onde mostram, respectivamente, conceitos de tipos de alojamento e a descrição de tipos de Pequenos e Médios Hotéis em Portugal.

Hontology possui 282 conceitos, os quais são organizados sob 16 conceitos de alto nível, através de uma hierarquia com profundidade máxima de cinco níveis. Além disso, é importante destacar que todos os conceitos e propriedades são multilíngues, estando disponíveis em quatro idiomas: inglês, português, espanhol e francês.

A definição dos conceitos de nível superior baseou-se nas necessidades básicas dos usuários, proporcionando, através destes conceitos, uma perspectiva abrangente e aprofundada do domínio no qual procuram informações. Os conceitos foram organizados nas seguintes categorias:

- **Acomodação:** Este conceito substituiu o conceito Hospitalidade, utilizado na primeira versão da Hontology. O conceito foi estendido para contemplar os diferentes tipos de acomodações, baseando-se nas categorias fornecidas pela (NET, 2009). O conceito acomodação possui ao todo 19 subcategorias, incluindo seis subcategorias sob o conceito Hotel: Hotel adega (do inglês, *Cave Hotel*), Hotel aquático (do inglês, *Under Water Hotel*), Hotel casa da árvore (do inglês, *Tree House Hotel*), Hotel Casamata (do inglês, *Bunker Hotel*), Hotel cápsula (do inglês, *Capsule Hotel*) e Hotel de gelo (do inglês, *Ice Hotel*).

- **Aparência:** Este conceito está relacionado com o uso do design na estrutura da acomodação para criar um ambiente acolhedor e agradável aos hóspedes.

- **Avaliação da Acomodação:** Este conceito está relacionado à classificação (do inglês, *rating*) recebido pela acomodação, uma medida de qualidade e satisfação.

- **Categorias de Hotéis:** Este conceito está relacionado à classificação dos hotéis, que varia de acordo com a qualidade dos serviços oferecidos. Essa classificação pode levar em consideração o número de estrelas: **1 estrela** (Turista,

Turista Superior); **2 estrelas** (Padrão, Padrão Superior); **3 estrelas** (Conforto, Conforto Classe Superior, Conforto Superior); **4 estrelas** (Primeira Classe) e **5 estrelas** (Luxo, Luxo Superior).

- **Endereço:** Este conceito está relacionado à localização da acomodação, como código postal, rua, cidade e país.
- **Facilidades:** Este conceito engloba os recursos projetados para oferecer comodidade aos hóspedes. Este conceito foi dividido em seis subcategorias: Acessibilidade para Deficientes, Acessibilidade para Motoristas, Facilidades do banheiro, Facilidades do quarto, Instalação Externa, Instalação Interna.
- **Horário:** Este conceito está relacionado com o detalhamento dos diferentes períodos de funcionamento de várias instalações ou serviços, como Horário Piscina Interior, Horário Piscina Exterior, Horário Restaurante e Horário Centro Spa. Cada uma destas subcategorias fornece informações precisas sobre os horários de abertura e fechamento dessas instalações ou serviços, ajudando os usuários a planejar suas atividades e a utilizar os serviços de forma mais eficiente.
- **Hospitalidade:** Este conceito está relacionado ao tratamento oferecido aos hóspedes, incluindo a gentileza e simpatia dos funcionários, visando proporcionar uma experiência acolhedora.
- **Pontos de interesse:** Este conceito está relacionado com locais de interesse próximos ao hotel, como aeroporto, universidade e centro comercial.
- **Preço:** Este conceito está relacionado com o detalhamento do preço de serviços específicos, como preço do bar, do café e da Internet. Cada uma destas subcategorias fornece informações precisas sobre o preço de serviços específicos, ajudando os usuários a planejar suas atividades e a utilizar os serviços de forma mais eficiente.
- **Quadro de Funcionários:** Este conceito está relacionado aos diversos funcionários disponíveis na acomodação, como funcionários da limpeza, porteiro e funcionários responsáveis pelo *check-in* e pelo *check-out*.
- **Quarto:** Este conceito está relacionado com as características dos diferentes tipos de quartos oferecidos pela acomodação, como quarto com duas camas e quarto com vista para o mar.
- **Rede de hotéis:** Este conceito está relacionado à rede a qual o hotel pertence, como AccorHotels e Vila Galé Hotéis.

- **Refeição:** Este conceito está relacionado com o tipo de refeições oferecidas pelo hotel, como almoço, café da manhã e jantar.
- **Serviço:** Este conceito está relacionado aos serviços oferecidos pela acomodação, como aluguel de bicicletas e veículos, serviço de porteiro e babá.
- **Tipos de Hóspede:** Este conceito é comumente utilizado para classificar os hóspedes de acordo com um perfil pré-definido, como casal, família, grupo de amigos e viajante de negócios.

Para informações detalhadas sobre a Hontology consulte a estrutura completa da ontologia no Anexo A.

Ontologias podem ser representadas através de diferentes níveis de representação, assumindo desde estruturas simples até estruturas mais complexas. A sua complexidade pode variar significativamente, dependendo do escopo e da profundidade do conhecimento que se deseja representar.

Estruturas simples consistem em listas de termos ou conceitos com definições básicas. Estas estruturas são fáceis de entender e criar, mas oferecem uma representação limitada da KB, pois não capturam as relações complexas entre conceitos.

Em contrapartida, estruturas complexas são estruturas que possuem vários níveis e podem incorporar uma gama ampla de relações semânticas entre conceitos, além de regras, axiomas e restrições lógicas. Esses tipos de estruturas são capazes de representar conhecimentos complexos e interconectados, como as relações causais em uma rede de conceitos científicos ou técnicos. Este nível de complexidade permite a realização de inferências avançadas e a geração de novos conhecimentos a partir da base existente.

Neste trabalho, ontologias são utilizadas durante a etapa de identificação dos aspectos, cada Item Linguístico (do inglês, *Linguistic Item* — LI) do vocabulário da ontologia representará um aspecto na lista de aspectos, mas somente aspectos do tipo explícito podem ser encontrados na lista de aspectos.

Vale salientar que, uma mesma *review* pode ser analisada em diferentes níveis de conhecimento. Por exemplo, o texto opinativo pode referir-se à ‘café da manhã’, no nível mais alto, e referir-se à ‘refeição’, no nível mais baixo. O nível baixo são representado através de estruturas simples e irão se referir as categorias da ontologia, como acomodação e aparência. Por outro lado, o nível alto são representados através de estruturas complexas, no qual irão se referir as subcategorias de níveis mais altos, como ‘aluguel de bicicletas’ e ‘Adagio’.

3.2 Sistema de Análise de Dependência Sintática

Neste trabalho, o processo de Análise de Dependência Sintática é realizada por uma biblioteca de NLP, denominada spaCy (HONNIBAL; MONTANI, 2017). A biblioteca spaCy pode ser utilizada para o processamento em diversos idiomas, incluindo o português, quanto maior o acervo utilizado para o treinamento do idioma selecionado mais precisa será a classificação do Analisador de Dependência (do inglês, *Dependency Parser* — DP) (HONNIBAL et al., 2020).

O DP do spaCy possui arcos direcionados, rotulados com os Rótulos de Dependência (do inglês, *Dependency tags* — DEP tags[2]) e Rótulos de Parte do Discurso (do inglês, *Part-of-Speech tags* — PoS tags[3]) (ALTINOK, 2021). Os DEP tags são colocados juntos ao arco direcionado e representam as relações entre palavras (como: sujeito-verbo, objeto-verbo e adjunto-nome) (SRINIVASA-DESIKAN, 2018). Já os POS tags representam a classe gramatical da palavra (como: adjetivo, verbo e advérbio) e são colocadas abaixo de cada palavra da frase analisada (PATEL; ARASANIPALAI, 2021).

A Figura 7 apresenta um exemplo de *review* de hotel [“O quarto era organizado e muito bonito. O banheiro era limpo e cheiroso”] após uso do DP da biblioteca spaCy (HONNIBAL et al., 2020).

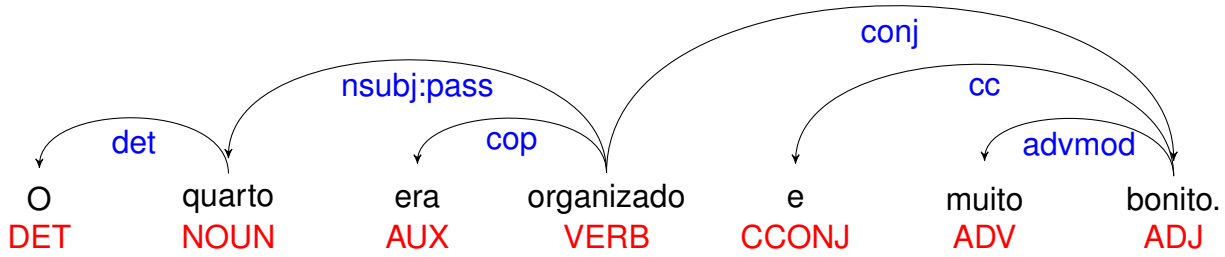

Figura 7 – Exemplo de Análise de Dependência Sintática de uma *review* de Hotel. Fonte: Autoria Própria.

[2]Para mais informações sobre o significado dos rótulos de DEP tags, acesse o link `https://universaldependencies.org/u/dep/index.html`.

[3]Para mais informações sobre o significado dos rótulos de PoS tags, acesse o link `https://universaldependencies.org/u/pos/index.html`.

Onde:

- **advmod** corresponde a *adverbial modifier* (modificador adverbial);
- **cc** corresponde a *coordinating conjunction* (conjunção coordenativa);
- **conj** corresponde a *conjunct* (conjunção);
- **cop** corresponde a *copula* (cópula);
- **det** corresponde a *determiner* (determinante);
- **nsubj:pass** corresponde a *passive nominal subject* (sujeito nominal passivo);

} : DEP tags

- **ADJ** corresponde a *adjective* (adjetivo);
- **ADV** corresponde a *adverb* (advérbio);
- **AUX** corresponde a *auxiliary* (verbo auxiliar);
- **CCONJ** corresponde a *coordinating conjunction* (conjunção coordenativa);
- **DET** corresponde a *determiner* (determinante);
- **NOUN** corresponde a *noun* (substantivo);
- **VERB** corresponde a *verb* (verbo);

} : PoS tags

3.3 Léxicos de Sentimento

Neste trabalho, os léxicos de sentimentos utilizados foram: AffectPT-br (CARVALHO; SANTOS; GUEDES, 2018), EmoLex (MOHAMMAD; TURNEY, 2010, 2013), LeIA (ALMEIDA, 2018), LIWC2007pt (BALAGE FILHO; PARDO; ALUÍSIO, 2013), Onto.PT (OLIVEIRA; GOMES, 2014; OLIVEIRA; SANTOS; GOMES, 2014), OpLexicon (SOUZA et al., 2011; SOUZA; VIEIRA, 2012), ReLi-Lex (FREITAS et al., 2012), SentiLex-PT (SILVA; CARVALHO; SARMENTO, 2012), SentiWordNet-PT-BR (BASTOS, 2023), UNILEX (SOUZA; PEREIRA; DALIP, 2017) e WordNetAffectBR (PASQUALOTTI; VIEIRA, 2008; PASQUALOTTI, 2015). A seguir foi realizado uma breve

descrição das características dos léxicos, levando em consideração os recursos utilizados para a criação dos LI e o método de atribuição da polaridade.

Os léxicos utilizados foram categorizados levando em consideração a forma de como foi realizada a atribuição de polaridade:

- **Léxicos de sentimentos criados manualmente:** Nessa categoria, enquadra-se somente o léxico ReLi-Lex.
  - No léxico ReLi-Lex, a lista de LI foi extraída de um corpus contendo *reviews* de livros e a atribuição de polaridade foi realizada de forma manual através de anotadores.
- **Léxicos de sentimentos criados automaticamente:** Nessa categoria, enquadra-se os léxicos AffectPT-br, Onto.PT, SentiWordNet-PT-BR e UNILEX.
  - No léxico AffectPT-br, a lista inicial de LI foi extraída do LIWC2015. Posteriormente, essa lista passou pelas etapas de adequação e expansão dos termos. A atribuição de polaridade foi realizada de forma automática utilizando as informações extraídas do LIWC2015.
  - No léxico Onto.PT, a lista de LI foi extraída de dicionários e *thesaurus*, a atribuição de polaridade foi realizada de forma automática utilizando as informações extraídas do SentiLex-PT.
  - No léxico SentiWordNet-PT-BR, tanto a criação da lista de LI quanto a atribuição de polaridade foram realizadas de forma automática, extraindo as informações de outros recursos.
  - No léxico UNILEX, a lista de LI foi extraída de *reviews* de redes sociais que estavam associados com o assunto política, a atribuição de polaridade foi realizada de forma automática utilizando o método de contagem de LI de outros léxicos.
- **Léxicos de sentimentos traduzidos:** Nessa categoria, enquadra-se os léxicos EmoLex, LeIA, LIWC2007pt e WordNetAffectBR.
  - No léxico EmoLex, a lista de LI foi extraída de dicionários e outros recursos, a atribuição de polaridade foi realizada de forma manual utilizando *crowdsourcing*.
  - No léxico LeIA, a lista de LI foi extraída de outros recursos, a atribuição de polaridade foi realizada de forma manual utilizando *crowdsourcing*.
  - No léxico LIWC2007pt, a lista de LI foi criada por avaliadores, a atribuição de polaridade foi realizada de forma automática utilizando métodos estatísticos.

  - No léxico WordNetAffectBR, parte da lista de LI foi extraída de dicionários e documentos científicos que tratam sobre a psicologia das emoções, o resto dos dados foram inseridas baseadas de forma intuitiva e arbitrária pelos autores. A atribuição de polaridade foi realizada de forma manual utilizando ferramenta de chat.

  Os recursos EmoLex, LeIA e LIWC2007pt foram traduzidos do inglês, enquanto que o WordNetAffectBR foi traduzido do inglês e italiano.

- **Léxicos de sentimentos híbridos:** Nessa categoria, enquadra-se os léxicos OpLexicon e SentiLex.

  - No léxico OpLexicon, a lista de LI foi extraída de corpus, dicionário e *thesaurus*, a atribuição de polaridade foi realizada de forma híbrida utilizando a união dos resultados resultantes de três métodos distintos (baseado em corpus, baseado em dicionário e baseado em *thesaurus*).
  - No léxico SentiLex-PT, a lista de LI foi extraída de corpus e outros recursos. A atribuição de polaridade foi realizada de forma híbrida utilizando a união dos resultados resultantes de dois métodos distintos (classificador estatístico e base de dados).

Para informações detalhadas sobre os SL consulte o Apêndice B.

4 TRABALHOS RELACIONADOS

Este capítulo aborda uma breve análise dos trabalhos que se dedicam à análise de sentimentos em diferentes contextos linguísticos, com ênfase especial naqueles que enfrentam desafios semelhantes ao presente estudo. A crescente necessidade de compreender as opiniões e sentimentos expressos pelos usuários em plataformas digitais impulsiona a pesquisa em diversas áreas, desde de técnicas baseadas em aprendizado de máquina até métodos mais centrados em aspectos linguísticos e semânticos. Nesse sentido, exploramos tanto os trabalhos que se beneficiam de recursos linguísticos abundantes, como os desenvolvidos para a língua inglesa, quanto aqueles que lidam com idiomas menos favorecidos em termos de recursos, como o português brasileiro.

Dentro deste escopo, uma atenção particular será dada aos estudos que empregam léxicos de sentimentos como uma ferramenta fundamental para a identificação e análise das polaridades das palavras em diferentes contextos e domínios. Este capítulo servirá como uma base para situar o presente trabalho no contexto mais amplo da pesquisa em análise de sentimentos, destacando as abordagens, metodologias e descobertas relevantes que informam e complementam os objetivos e contribuições desta dissertação.

4.1 Análise de Sentimento para Línguas de Muitos Recursos

Esta seção direciona seu foco para uma análise dos estudos voltados à compreensão de sentimentos na língua inglesa, conhecida por sua abundância de recursos linguísticos. Esses estudos exploram uma ampla gama de abordagens e técnicas para extrair e interpretar sentimentos expressos em diferentes contextos e domínios. Ao examinar esses trabalhos, nosso objetivo é não apenas destacar as metodologias eficazes empregadas para a língua inglesa, mas também identificar tendências e inovações que possam ter aplicações significativas em outras línguas e contextos linguísticos. Esta seção fornece uma base crucial para a compreensão das estratégias utilizadas na análise de sentimentos em línguas com uma abundância de recursos

linguísticos, preparando o terreno para a discussão de abordagens em línguas com recursos mais limitados posteriormente neste capítulo.

4.1.1 *Semantic Orientation Calculation* – SO-CAL (OSGOOD; SUCI; TANNENBAUM, 1957)

O trabalho "*Semantic Orientation Calculation*" (SO-CAL), de Osgood; Suci; Tannenbaum (1957), foi um marco na área da psicologia social e da ciência da computação. Esse estudo propôs um método inovador para quantificar o significado das palavras, lançando as bases para pesquisas e aplicações subsequentes em diversas áreas.

O método SO-CAL se baseia na avaliação de três dimensões principais que definem o significado de uma palavra:

1. **Avaliação**: Refere-se ao julgamento positivo ou negativo que uma pessoa faz da palavra. Em outras palavras, a avaliação indica se a palavra evoca sentimentos agradáveis ou desagradáveis.

2. **Potência**: Esta dimensão se refere à força ou intensidade do significado da palavra. Uma palavra com alta potência é aquela que transmite uma mensagem clara e forte, enquanto uma palavra com baixa potência pode ser mais vaga ou ambígua.

3. **Atividade**: A atividade indica o nível de ação ou dinamismo associado à palavra. Palavras com alta atividade evocam imagens de movimento e energia, enquanto palavras com baixa atividade são mais estáticas e passivas.

Para quantificar essas três dimensões, Osgood; Suci; Tannenbaum realizaram um estudo com 50 participantes. Os participantes avaliaram 200 adjetivos em escalas de sete pontos, variando de -3 (extremo negativo) a +3 (extremo positivo).

Os resultados do estudo mostraram que as três dimensões – avaliação, potência e atividade – eram suficientes para explicar a maioria da variância no significado das palavras. Isso significa que, ao conhecer as avaliações de uma palavra em cada uma dessas dimensões, é possível ter uma boa compreensão do seu significado geral.

O trabalho de Osgood; Suci; Tannenbaum é considerado um marco na área de análise de sentimento. Ao propor um método para quantificar o significado das palavras através das dimensões de avaliação, potência e atividade, lançou as bases para pesquisas e aplicações subsequentes em diversas áreas. O SO-CAL é considerado o precursor da análise de sentimento, pavimentando o caminho para o desenvolvimento de técnicas mais sofisticadas que hoje são utilizadas em diversas áreas como marketing, análise de opinião pública e até mesmo na área da saúde (TABOADA et al., 2011). Apesar de suas limitações, a importância do SO-CAL como precursor da análise de sentimento é inegável, inspirando pesquisas e inovações que continuam a aprimorar a compreensão das emoções e do comportamento humano.

4.1.2 Método de Análise de Sentimento de Turney (2002)

Turney (2002) propõe um sistema de recomendação baseado em um algoritmo simples de aprendizado não supervisionado para a língua inglesa. É utilizado o algoritmo PMI-IR para estimar a orientação semântica de uma frase. O PMI-IR consiste em utilizar Informações Mútuas Pontuais (do inglês, *Pointwise Mutual Information* — PMI) e Recuperação de Informação (do inglês, *Information Retrieval* — IR) para calcular a similaridade de pares de palavras ou frases.

O sistema proposto consiste em 3 etapas: extrair frases contendo adjetivos ou advérbios; estimar a orientação semântica de cada frase; classificar a revisão com base na orientação semântica média das frases.

Na primeira etapa acontece a PoS tags, através de um marcador de classe gramatical, e a extração de frases que contenham adjetivos e advérbios. Depois as palavras das frases são agrupadas seguindo os seguintes padrões:

- <adjetivo> + <substantivo> + <qualquer tipo de palavra>
- <advérbio> + <adjetivo> + <qualquer tipo de palavra, exceto substantivo>
- <adjetivo> + <adjetivo> + <qualquer tipo de palavra, exceto substantivo>
- <substantivo> + <adjetivo> + <qualquer tipo de palavra, exceto substantivo>
- <advérbio> + <verbo> + <qualquer tipo de palavra>

Embora o adjetivo isolado possa indicar a subjetividade, o contexto pode influenciar o resultado da orientação semântica. Por exemplo, o adjetivo “imprevisível” pode ter uma orientação semântica negativa em uma avaliação de um automóvel, como na frase “direção imprevisível”, mas, em uma avaliação de um filme, a frase “enredo imprevisível” pode ter uma orientação semântica positiva. Dessa forma, o bigrama (conjunto de palavras agrupadas em sequência de duas palavras) inicial será composto por um adjetivo ou advérbio e a palavra após ele fornecerá o contexto. Bigramas seguidos de substantivos foram excluídos para que os nomes dos itens da revisão não influenciem o resultado da classificação.

Na segunda etapa, é estimado o valor da orientação semântica das frases extraídas. O valor da orientação semântica é obtido através do algoritmo PMI-IR, onde a PMI é utilizado para medir a força de associação semântica entre duas palavras. A PMI entre duas palavras, $word_1$ e $word_2$, é definida pela Equação 1.

$$PMI\left(word_1, word_2\right) = log_2\left(\frac{P(word_1 \wedge word_2)}{P(word_1 \cdot word_2)}\right) \tag{1}$$

Onde: $P(word_1 \wedge word_2)$ e $P(word_1 \cdot word_2)$ representam, respectivamente, a probabilidade das palavras coocorrerem, com e sem dependência estatística. Como a razão

entre $P(word_1 \wedge word_2)$ e $P(word_1 \cdot word_2)$ representa uma medida do grau de dependência estatística entre as palavras, temos que $PMI(word_1, word_2)$ representará a quantidade de informação que adquirimos sobre a presença de uma das palavras quando observamos a outra, ou seja, $PMI(word_1, word_2)$ representa uma medida que quantifica o grau de associação entre duas palavras.

Dessa forma, se o valor de $PMI(word_1, word_2)$ é alto, isso sugere que saber da presença de uma das palavras nos dá indícios da provável presença da outra palavra. Em contrapartida, se o valor de $PMI(word_1, word_2)$ é baixo, isso sugere que as palavras não estão fortemente associadas no corpus estudado.

A Orientação Semântica (do inglês, *Semantic Orientation* — SO) de uma frase é definida pela Equação 2.

$$SO\,(phrase) = PMI\,(phrase, \text{“}excellent\text{”}) - PMI\,(phrase, \text{“}poor\text{”}) \tag{2}$$

A escolha das palavras “$excellent$” e “$poor$” foi em razão do uso frequente dessa palavras no sistema de avaliação de cinco estrelas, sendo “$excellent$” equivalente a cinco estrelas e “$poor$” equivalente a uma estrela. Caso a SO seja positiva, a frase está mais fortemente associada a “$excellent$”. Caso contrário, ela está associada a “$poor$”.

Na terceira etapa, é realizada o cálculo da SO média das frases de cada avaliação. Caso a média da SO seja positiva, a avaliação será classificada com recomendada. Caso contrário, ela será classificada com não recomendada.

O sistema proposto por Turney (2002) foi testado em quatro domínios diferentes (avaliações de automóveis, bancos, filmes e destinos de viagens). A acurácia para os domínios foram de 84%; 80%; 65,83%; 70,53%; respectivamente, atingindo uma acurácia média de, aproximadamente, 75%.

4.1.3 Método de Análise de Sentimento de Pang; Lee; Vaithyanathan (2002)

Pang; Lee; Vaithyanathan (2002) propuseram um sistema de AS baseado em técnicas de aprendizado de máquina para o domínio de avaliações de filmes. É utilizado os algoritmos de *Maximum Entropy*, *Support Vector Machines* e *Naïve Bayes* para obter o resultado da análise.

O *dataset* utilizado no estudo foi construído selecionando aleatoriamente 700 avaliações classificadas com sentimento positivo e 700 avaliações classificadas com sentimento negativo. Em seguida, esses dados foram divididos em três partes iguais e, quando possível, foi realizada a validação cruzada.

Também foi utilizado um conjunto predefinido de *features* que podem aparecer em uma avaliação, esse conjunto foi separado em dois subconjuntos: unigrama e bigrama. O subconjunto unigrama contém unigramas (como *“still”*) e o subconjunto

bigrama contém bigramas (como *"really stinks"*).

Ademais, foi testado a técnica de adicionar a tag NOT_ entre uma palavra de negação (*"not"*, *"isn't"*, *"didn't"*, etc.) e a primeira marca de pontuação após a palavra de negação. Resultados preliminares mostram que a técnica teve um efeito insignificante, mas, que na média representou uma perda ligeiramente prejudicial.

O estudo foi baseado em unigramas (com marcação de negação) e bigramas. A marcação de negação não foi utilizada em bigramas, pois, os bigramas já fornecem uma maneira única e independente de entender o contexto em que as palavras estão inseridas.

Como o tempo de treinamento com *Maximum Entropy* se torna altamente custoso com o aumento de *features*, este número foi limitado a 16.165 *features*. Foram realizados dois testes: utilizando os unigramas que aparecem pelo menos quatro vezes; utilizando bigramas que ocorrem pelo menos sete vezes. Ainda, foi testada a influência da inserção de POS tags, da posição de palavras e o uso somente de adjetivos.

A inserção de PoS tags funciona como desambiguação do sentido das palavras, auxiliando na distinção dos usos diferentes das palavras. Como exemplo podemos citar a palavra *"love"* nas avaliações *"I **love** this movie"* e *"This is a **love** story"*. Enquanto que na primeira frase *"love"* tenha um sentido orientação positiva de sentimento, *"love"* na segunda frase representa um sentido orientação neutra. No entanto, o efeito da inserção de PoS tags se mostrou inconclusivo, pois, se por um lado a acurácia para *Naïve Bayes* melhora, por outro lado, para *Support Vector Machines* piora e para *Maximum Entropy* permanece inalterado.

O resultado do uso somente de adjetivos mostrou-se ineficiente. A lista de 2633 adjetivos fornece informações menos úteis do que a lista de unigramas. Na verdade, uma lista dos unigramas mais frequentes, contendo o mesmo número de *features* (2633 unigramas), produz um resultado melhor. Sendo comparável a lista de presença de todos os unigramas, que contém a lista dos 16165 unigramas mais frequentes.

O sistema proposto por Pang; Lee; Vaithyanathan (2002) teve o melhor desempenho utilizando *Support Vector Machines*, enquanto que *Naïve Bayes* teve o pior desempenho, embora a diferença dos resultados não sejam muito significativa. A lista de *features* que mostrou ser mais eficiente foi a presença dos unigramas.

A acurácia média dos resultados da validação cruzada para os modelos de *Support Vector Machines* e *Naïve Bayes*, utilizando a presença dos unigramas, foram, respectivamente, 82,9% e 81%.

4.1.4 Análise de Sentimento Baseada em Aspecto de Rani; Jain (2023)

Rani; Jain (2023) propuseram um modelo de ABSA chamado de Modelo LSTM Bidirecional Duplo baseado em Aprendizagem Multitarefa (do inglês, *Multi-task Learning based Dual Bidirectional LSTM Model* — MLDBM). Esse modelo foi desenvolvido para

a área médica, baseado no relato de pacientes encontrados em redes sociais médicas e nos fóruns de saúde. Esses relatos podem fornecer informações importantes, pois podem destacar vários aspectos dos medicamentos, incluindo seus efeitos colaterais, benefícios, eficácia, dosagem recomendada, impacto em condições médicas específicas, custos e experiência geral do usuário.

Como os modelos de Memória de Longo e Curto Prazo (do inglês, *Long Short-Term Memory* — LSTM) são projetados principalmente para capturar dependências sequenciais entre entradas e não são explicitamente treinados para identificar e extrair aspectos ou características específicas de um texto, os modelos LSTM não são capazes de detectar adequadamente as palavras ou frases importantes que são relevantes para o ABSA.

Baseando-se no trabalho desenvolvido por Wang et al. (2016), Rani; Jain (2023) propuseram um novo modelo BiLSTM duplo construído com duas camadas BiLSTM (*Bidirectional LSTM*) independentes, que são projetadas para capturar mais informações contextuais e aprimorar a capacidade geral do modelo.

O primeiro BiLSTM recebe os *embeddings* de sequência de entrada gerados pelo BERT, enquanto que o segundo BiLSTM recebe o resultado da camada de Autoatenção com Múltiplas Cabeças (do inglês, *Multi-Head Self Attention* — MHSA). O MHSA é utilizado para codificar as palavras contextuais significativas necessárias para identificar o aspecto, extrair aspectos importantes e gerar representações específicas do aspecto.

O resultado do BiLSTM duplo passa então por uma camada de concatenação que realiza a soma elemento a elemento das representações de cada BiLSTM. As representações resultantes contêm as informações combinadas dos dois BiLSTM.

O resultado da camada de concatenação passa então por uma camada de atenção que auxilia o modelo na compreensão das porções significativas da entrada que são cruciais para a captura de informações relacionadas ao sentimento. Especificamente, a camada de atenção concentra-se nas representações específicas do aspecto que foram capturadas pelas camadas anteriores. A camada de atenção atribui peso a cada elemento, indicando a importância desse elemento para capturar o sentimento geral da frase.

A camada final é uma camada SoftMax, que é usada para fazer previsões sobre a polaridade de sentimento da sentença de entrada. A camada SoftMax pega a soma ponderada dos vetores ocultos da camada de atenção como entrada e produz uma distribuição de probabilidade.

4.2 Análise de Sentimento para Línguas de Poucos Recursos

Línguas de poucos recursos são aquelas em que os recursos linguísticos (como: ferramentas ou *datasets*) são escassos. Isso está intimamente relacionado com a complexidade da língua analisada.

Um desses casos é a língua Hindi e o Português. O Hindi é a quarta linguagem mais popular no mundo, por possuir uma complexidade elevada torna-se custoso desenvolver recursos linguísticos que contemplem toda a complexidade da língua Hindi (GUPTA et al., 2021).

4.2.1 Análise de Sentimento Baseada em Aspecto de Singh et al. (2020)

Singh et al. (2020) desenvolveram um sistema de detecção de sentimentos abusivos baseados em aspectos para a língua Nepali. Esse modelo foi desenvolvido utilizando um *dataset* de mídia social.

Treinou-se dois tipos diferentes de *embeddings*, Monolíngue e Multilíngue. Para realizar a extração dos aspectos foram testados BiLSTM+CRF e BERT multilíngue pré-treinado, e para fazer a classificação da polaridade do sentimento foram utilizados SVM, CNN, BiLSTM and BERT multilíngue pré-treinado.

Os Campos Aleatórios Condicionais (do inglês, *Conditional Random Fields*) são modelos estatísticos utilizados para prever sequências de rótulos para sequências de dados de entrada. Eles são particularmente úteis em tarefas de processamento de linguagem natural, como reconhecimento de entidades nomeadas e marcação de partes da fala, onde o contexto é crucial para determinar a etiqueta correta. Ao contrário de modelos mais simples que fazem previsões para cada ponto de dados de forma independente, os CRFs levam em consideração a estrutura de dependência entre as etiquetas, permitindo fazer previsões mais precisas baseadas no contexto. Memória de Longo Prazo de Curto Prazo.

LSTM é um tipo especial de Rede Neural Recorrente (do inglês, *Recurrent Neural Network* — RNN) projetada para lidar com o problema do desaparecimento do gradiente, que afeta as RNNs tradicionais. As LSTMs são capazes de aprender dependências de longo prazo e são amplamente utilizadas em tarefas que envolvem sequências de dados, como tradução automática, geração de texto e reconhecimento de fala. Elas funcionam mantendo um estado interno, ou "memória", que permite que a rede retenha informações ao longo do tempo, tornando-as eficazes para processar sequências de dados onde o contexto temporal é importante.

Rede Neural Convolucional (do inglês, *Convolutional Neural Network* — CNN) é um tipo de rede neural profunda especialmente projetadas para processar dados com uma estrutura de grade, como imagens. As CNNs utilizam operações de convolução, que aplicam filtros aos dados de entrada para extrair características importantes. Isso

as torna extremamente eficientes para tarefas de visão computacional, como reconhecimento de imagens, detecção de objetos e classificação de imagens. As CNNs podem aprender hierarquias de características, onde características de baixo nível são usadas para construir características de alto nível, facilitando a aprendizagem de representações complexas dos dados.

Para a extração de aspecto o modelo utilizando o BERT multilíngue pré-treinado teve um desempenho um pouco melhor que em comparação com BiLSTM + CRF. Singh et al. (2020) salientam que caso o modelo BERT fosse treinado do zero com base no corpus inglês e nepalês teria um desempenho muito superior ao BERT multilíngue pré-treinado. Entretanto, para a classificação da polaridade do sentimento o modelo BiLSTM teve melhor desempenho em comparação ao modelo BERT multilíngue.

4.2.2 Análise de Sentimento Baseada em Aspecto de Rani; Kumar (2021)

Rani; Kumar (2021) propõe um sistema de ABSA para a língua Hindi. É utilizado o HDP[1] (*Hindi Dependency Parser*) (AMBATI, 2011) junto com o HSWN[2] (Hindi SentiWordNet (JOSHI et al., 2010), com o intuito de encontrar as associações entre os aspectos e as palavras de uma frase, e encontrar a polaridade expressa por um determinado aspecto ou pela frase.

O sistema proposto consiste em 8 etapas: fase de extração de dados; fase de pré-processamento; extração de nós de sentimento; fase de extração de aspecto; criação de vetor de aspecto; fase de geração do grafo de dependência; fase de manipulação de negação e intensificadores; fase de atribuição de polaridade.

Dentre todas as etapas citadas, a fase de manipulação de negação e intensificadores não constam no projeto desenvolvido neste artigo. Essa etapa tem o intuito de aprimorar o sistema buscando advérbios de negação ou intensidade, que podem inverter ou aumentar/diminuir a polaridade da frase, tornando o sistema mais confiável e eficiente.

O sistema proposto por Rani; Kumar (2021) foi testado em um *corpus* de resenhas de filmes, que é composto por 2.247 *reviews* de filmes escritas em hindi e extraídas dos sites Aaj Tak e Jagran. O sistema de Rani; Kumar (2021) atingiu uma precisão de 83,2%, fazendo com que o modelo possa ser utilizado em outras áreas, em razão do seu excelente resultado.

4.2.3 Análise de Sentimento Baseada em Aspecto para Língua Portuguesa

Freitas (2015)propõe um sistema de ABSA para a língua Portuguesa utilizando

[1]O recurso Hindi Dependency Parser pode ser obtido através do link `https://sivareddy.in/downloads/index.html`.

[2]O recurso Hindi SentiWordNet pode ser obtido através do formulário disponível em `https://www.cfilt.iitb.ac.in/resources/senti/HSWN_downloaderInfo.php`.

diferentes PoS *taggers*[3] , léxicos de sentimento e ontologia de domínio. Segundo Freitas (2015), “o uso de ontologias tem o potencial de refinar e melhorar o processo de análise de sentimentos por meio da identificação de propriedades e relações entre conceitos”. Dessa forma, a ontologia é utilizada para encontrar os aspectos que estão relacionadas com determinada área, refinando os aspectos encontrados, ou seja, estamos centralizando as buscas nos aspectos mais importantes relacionados com a área de interesse. Neste trabalho foi utilizada a ontologia do domínio de acomodações denominada Hontology (CHAVES; FREITAS; VIEIRA, 2012).

A profundidade máxima da ontologia Hontology é cinco, quando mais profundo for a ontologia maior é a complexidade do grafo e maior é quantidade de conceitos utilizados na estrutura da ontologia (CHAVES; FREITAS; VIEIRA, 2012). Por essa razão, no trabalho desenvolvido por Freitas (2015) são utilizados somente os três primeiros níveis da ontologia, totalizando 77 aspectos, sendo 73 formados por palavras simples e 4 formados por palavras compostas.

Além disso, a metodologia de Freitas (2015) utiliza: (1) os léxicos de sentimento, OpLexicon (SOUZA et al., 2011; SOUZA; VIEIRA, 2012), SentiLex(SILVA; CARVALHO; SARMENTO, 2012), LIWC2007pt (BALAGE FILHO; PARDO; ALUÍSIO, 2013) e synsets com polaridades do Onto.PT (OLIVEIRA; GOMES, 2014; OLIVEIRA; SANTOS; GOMES, 2014); (2) os PoS *taggers*, TreeTagger[4] (SCHMID, 1999, 2013), FreeLing[5] (CARRERAS et al., 2004; ATSERIAS et al., 2006; PADRÓ et al., 2010; PADRÓ, 2011; PADRÓ; STANILOVSKY, 2012) e CitiusTagger[6] (GAMALLO et al., 2014; GARCIA; GAMALLO, 2015); (3) as regras linguísticas que levam em consideração a negação e a posição do adjetivo em relação ao aspecto analisado. Ainda, foram utilizadas onze configurações diferentes, levando em consideração os léxicos de sentimento, os PoS *taggers* e as regras linguísticas. Os melhores resultados obtidos foram utilizando o Onto.PT, o TreeTagger e a posição dos adjetivos com 79% de medida-f para positivo e 47% de medida-f para negativo, sendo a média igual a 63,2%, considerando os aspectos ‘quarto’, ‘localização’, ‘atendimento’, ‘limpeza’ e ‘custo-benefício’.

[3]Os *taggers* são ferramentas de etiquetagem morfossintática e morfológica. Dessa forma, os *taggers* atuam como ferramentas computacionais que atribuem automaticamente o PoS e as etiquetas morfológicas para cada uma das palavras presentes em uma frase (FINATTO et al., 2024). Como exemplo de *tagger*, podemos citar o PoS *tagger*, que é um etiquetador morfossintático.

[4]O recurso TreeTagger pode ser obtido através do link `https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`.

[5]O recurso FreeLing pode ser obtido através do link `https://github.com/TALP-UPC/FreeLing/releases`.

[6]O recurso CitiusTagger pode ser obtido através do link `https://gramatica.usc.es/pln/tools/CitiusTools.html`.

5 METODOLOGIA

5.1 Dataset

O *dataset* utilizado no presente trabalho foi o mesmo *dataset* usado na competição ABSAPT-2022 (SILVA et al., 2022)[1]. O *dataset* foi construído juntando os *datasets* desenvolvidos anteriormente por Freitas (2015)[2] e Corrêa (2021)[3]. Ambos os recursos foram anotados seguindo o protocolo de anotação proposto por Freitas (2015). Como dependência deste protocolo de anotação temos a necessidade de uma ontologia de domínio de onde são retirados os aspectos a serem analisados, em ambos os casos foi utilizada a Hontology (CHAVES; FREITAS; VIEIRA, 2012).

O *dataset* da competição ABSAPT-2022 contém *reviews* de usuários do TripAdvisor[4], sendo que cada *review* pode, ou não, conter uma ou mais opiniões. O *dataset* é composto por 3.111 registros, provenientes de 847 avaliações sobre o setor hoteleiro. Sendo 2.112 registros com polaridade positiva, 527 registros com polaridade negativa e 472 registros com polaridade neutra. Dos 3.111 registros, 2.856 contém aspectos simples[5] e 255 contém aspectos compostos[6].

As informações do *dataset* da competição ABSAPT-2022 são distribuídas em 5 colunas: *review*, *polarity*, *aspect*, *start_position* e *end_position*. A coluna *review* contém a *review* avaliada, a coluna *polarity* contém a polaridade do aspecto, a coluna *aspect* contém o aspecto encontrado na *review*, a coluna *start_position* contém a posição ini-

[1] O *dataset* usado na competição ABSAPT-2022 (SILVA et al., 2022) pode ser baixado entrando em contato com absapt2022@inf.ufpel.edu.br. Após o contato você receberá por e-mail com o corpus compactado e a senha necessária para descompactá-lo. Mais informações podem ser obtidas pelo link `https://sites.google.com/inf.ufpel.edu.br/absapt2022/`.

[2] O *dataset* criado e usado no estudo de Freitas (2015) está disponível publicamente, através do link `https://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/tripadvisor/`.

[3] O *dataset* criado e usado por Corrêa (2021) era privado até ser integrado no dataset do ABSAPT-2022.

[4] O TripAdvisor é o maior site de viagens do mundo, possuindo mais de 859 milhões de avaliações e opiniões sobre 8,6 milhões de acomodações, restaurantes, experiências, companhias aéreas e cruzeiros. Dessa forma, ele se torna uma importante fonte de informações para treinamento de modelos de SA para serviços de hospedagens.

[5] Chamamos de Aspectos simples aqueles formados por uma única palavra.

[6] Aspectos compostos são aspectos formados pela combinação de duas ou mais palavras.

cial do aspeto na *review* e a coluna *end_position* contém a posição final do aspeto na *review*.

5.2 Aspectos

A Hontology apresenta diversos níveis e, quanto mais profundo for o nível de uma ontologia, maior será a sua complexidade e a quantidade de conceitos que serão empregados durante a etapa de busca de aspectos. Por essa razão, assim como no estudo realizado por Freitas (2015), somente os três primeiros níveis da ontologia serão utilizados. No entanto, diferentemente do trabalho de Freitas (2015), optou-se exclusivamente pelos aspectos simples da lista de aspectos que compõem os três primeiros níveis da Hontology, resultando em um total de 73 aspectos.

A lista de aspectos utilizada utilizada durante a etapa de busca de aspectos é composta pelos seguintes aspectos: ‘aeroporto’, ‘almoço’, ‘apartamento’, ‘aquecimento’, ‘atendimento’, ‘banheira’, ‘cafeteira’, ‘calefação’, ‘cama’, ‘carpete’, ‘casal’, ‘cassino’, ‘chuveiro’, ‘cidade’, ‘colchão’, ‘conforto’, ‘corredor’, ‘cortina’, ‘cozinha’, ‘custo-benefício’, ‘ducha’, ‘elevador’, ‘escada’, ‘estabelecimento’, ‘estacionamento’, ‘família’, ‘farmácia’, ‘frigobar’, ‘funcionários’, ‘garagem’, ‘gerente’, ‘gerência’, ‘horário’, ‘hotel’, ‘iluminação’, ‘instalações’, ‘internet’, ‘jantar’, ‘jardim’, ‘lavanderia’, ‘limpeza’, ‘localização’, ‘lojas’, ‘luxo’, ‘motel’, ‘museu’, ‘móveis’, ‘padrão’, ‘parque’, ‘pia’, ‘piscina’, ‘porteiro’, ‘praia’, ‘praça’, ‘preço’, ‘quarto’, ‘recepção’, ‘rodoviária’, ‘rua’, ‘secador’, ‘serviço’, ‘shopping’, ‘suíte’, ‘tapete’, ‘teatro’, ‘telefone’, ‘televisão’, ‘toalha’, ‘tomada’, ‘torneira’, ‘travesseiro’, ‘turista’, ‘tv’.

5.3 Tratamento dos Léxicos de Sentimento

Cada léxico de sentimento adota uma estrutura de organização própria, resultando na ausência de um padrão comum entre os léxicos. Além disso, o uso de separadores distintos entre eles dificulta a padronização do código. Por essa razão, foi necessário realizar o tratamento dos léxicos. Em tese, esse tratamento consiste em uma padronização de cada léxico, de forma a facilitar o acesso à informação e a implementação do sistema proposto. Para tal, gerou-se um dicionário para cada léxico utilizado, cujas chaves são compostas pelos LI (por exemplo, palavras, *stemming* de palavras, gírias, expressões ou elementos de linguagem (*hashtags*, *emoticon*)) utilizados pelo recurso léxico e os valores das chaves correspondem à polaridade do LI utilizado na chave. No entanto, como nem todos os léxicos possuem a coluna de polaridade, ou não utilizavam “1” para positivo e “-1” para negativo, foi necessário realizar a transformação dos dados. Por exemplo:

- Para recursos como o WordNetAffectBR, que possuem símbolos para represen-

tar cada polaridade, foi criado uma nova coluna chamada de ”polaridade” e atribuído a polaridade correspondente ao símbolo utilizado. Para o símbolo “+” foi atribuído “1” e para símbolo “-” foi atribuído “-1”.

- Para recursos como o SentiWordNet-PT-BR, que possuem uma porcentagem de polaridade positiva ou negativa, foi criado uma nova coluna chamada de “polaridade” e atribuído “1” ou “-1” dependendo do maior valor de polaridade. Se o valor de polaridade positiva for maior é atribuído o valor “1”, caso contrário, é atribuído o valor “-1”.

- Para recursos como o Reli-Lex, que possuem listas para cada polaridade, foi criado uma nova coluna chamada de ”polaridade” e atribuído a polaridade correspondente ao *dataset* utilizado. Para *datasets* com sentimento positivo foi atribuído “1” e para sentimento negativo foi atribuído “-1”.

Cada léxico de sentimentos possui sua própria forma de armazenar os LI. Dessa forma, em alguns casos foi necessário realizar a normalização *Unicode*, que remove sinais diacríticos[7] das palavras. Além disso, em alguns léxicos, também foi necessário corrigir erros relacionados à codificação, enquanto em outros foi necessário corrigir erros relacionados ao separador, pois utilizavam um tipo diferente de separador ou mais de um separador.

Os léxicos Reli-Lex, LeIA, WordNetAffectBR utilizavam a codificação “ISO-8859-1”, por outro lado, os demais léxicos utilizavam a codificação “UTF-8”. O léxico WordnetAffectBR empregava o separador “;”, enquanto que os léxicos OntoPT, OpLexicon, LIWC utilizavam o separador “,”. Já os léxicos AffectPT-br, Emolex, LeIA, SentiWordNet utilizavam o separador “\t”. Os léxicos UNILEX e SentiLex utilizavam dois separadores: “,” e “\t”, “.” e “;”, respectivamente.

5.4 Análise de Dependência Sintática

O DP do spaCy utiliza uma ampla gama de DEP tags, sendo que cada rótulo possui suas características únicas, determinadas pela configuração estrutural da frase. Contudo, uma vez que *reviews* frequentemente seguem padrões sintáticos similares, o número de DEP tags que devem ser analisados é reduzido.

Para realizar essa investigação, o *dataset* foi examinado para identificar os DEP tags utilizados, dando especial atenção àqueles que apresentavam alguma relação com adjetivos. Os DEP tags encontrados nos textos opinativos incluem:

- **acl** corresponde a *adnominal clause* (cláusula adnominal – modificador oracional do substantivo);

[7] Sinais diacríticos são sinais gráficos que serve para diferençar letras ou palavras, como acentos, til e cedilha

- **amod** corresponde a *adjectival modifier* (modificador adjetivo);
- **advcl** corresponde a adverbial *clause modifier* (modificador de cláusula adverbial);
- **advmod** corresponde a *adverbial modifier* (modificador adverbial);
- **nmod** corresponde a *nominal modifier* (modificador nominal);
- **conj** corresponde a *conjunct* (conjunção);
- **nsubj** corresponde a *nominal subject* (sujeito nominal).

Depois de identificar os DEP tags foi necessário reconhecer qual estrutura sintática utilizada por cada um deles, pois, dependendo da relação utilizada a relação entre os *tokens* também mudará. Por exemplo, nas relações de dependência 'nsubj' e 'amod', o adjetivo é o *head token*, enquanto que nas relações de dependência 'acl' e 'conj', o adjetivo é o *child token*. Além disso, como o DP do spaCy costuma identificar erroneamente o POS tag, foi necessário procurar, além dos adjetivos, por substantivos e advérbios.

5.5 Correção Ortográfica

Ao explorar mais minuciosamente o *dataset* foi constatado que algumas *reviews* apresentavam erros ortográficos. Inicialmente, esses erros passaram despercebidos, pois só se constatou esse problema quando observou-se uma amostra maior do *dataset*, em razão de que os erros ortográficos não apareciam nas primeiras linhas do *dataset*.

Depois de constatado esse problema, verificou-se se os erros representavam uma grande significância. Foram encontrados várias palavras as quais apresentava problemas de ortografia, muitas delas estavam relacionadas aos aspectos ou adjetivos. Como exemplos de erros ortográficos, podemos citar:

- **custo-benefício:** o aspecto custo-benefício estava representado por: custo/benefício, custo/beneficio, custoXbenefício, custoXbeneficio, custo-beneficio, custo benefício ou custo beneficio;
- **recepção** o aspecto recepção estava representado por: receçao, receção, recepcao, recepcão ou recepçao;
- **amei:** o adjetivo amei estava representado por: ameei, ameii, ameeeiiiii, ...

Os métodos de verificação ortográfica funcionam, geralmente, em nível lexical (por exemplo, autocorrect[8] e pyspellchecker[9]), fornecendo correções ortográficas, ou em nível contextual e sintático (por exemplo, LanguageTool[10]), possibilitando, além das correções ortográficas, a detecção de erros gramaticais (VOGHOEI et al., 2023).

Para corrigir os erros ortográficos, foram testados tanto a correção ortográfica manual quanto as bibliotecas de correção ortográfica autocorrect, LanguageTool e pyspellchecker. Enquanto a correção manual requer a criação de regras específicas para cada erro ortográfico identificado, limitando-se a um *dataset* específico e não abrangendo todos os erros presentes em outros *datasets*, bibliotecas de correção ortográfica podem ser aplicados a qualquer *dataset*.

5.6 Agrupamento de adjetivos

O agrupamento de adjetivos consiste em uma sequência de adjetivos que estão ligados a um mesmo aspecto por meio de uma ou mais conjunções (**conj**) ou através do uso de vírgula. O uso do agrupamento de adjetivos visa auxiliar os recursos que tem poucos LI, tendo em vista que, em situações onde um aspecto é descrito por adjetivos interligados, quando o primeiro adjetivo não for identificado, é possível recorrer à polaridade dos adjetivos subsequentes.

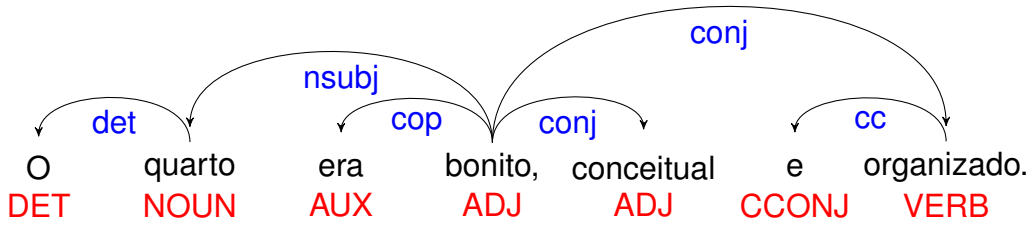

Figura 8 – Exemplo de Análise de Dependência Sintática onde ocorre o agrupamento de adjetivos. Fonte: Autoria Própria.

A Figura 8 apresenta o grafo de dependência semântica gerado pelo spaCy para um *review* onde ocorre o agrupamento de adjetivos. Como podemos observar, os adjetivos “bonito”, “conceitual” e “organizado” fazem referência ao substantivo “quarto”. Podemos observar que as relações de dependência geradas pelo spaCy possibilitam a conexão do aspecto “quarto” com o adjetivo “bonito”, que é o primeiro adjetivo ligado ao aspecto, e que as conjunções (**conj**) ligam o adjetivo “bonito” aos outros adjetivos (“conceitual” e “organizado”). No entanto, podemos notar que a marcação de parte

[8] A documentação do recurso autocorrect pode ser encontrada em `https://github.com/filyp/autocorrect`.

[9] A documentação do recurso pyspellchecker pode ser encontrada em `https://pyspellchecker.readthedocs.io/en/latest/`

[10] A documentação do recurso LanguageTool pode ser encontrada em `https://pypi.org/project/language-tool-python/`.

do discurso (do inglês, *Part-of-Speech* — PoS) do adjetivo “organizado” está incorreta, uma vez que o *parser* anotou este *token* como um verbo (**VERB**) e não como um adjetivo (**ADJ**). Este erro de marcação interfere diretamente na execução de uma metodologia baseada em léxico, visto que, na etapa de Identificação das Polaridades é realizada a busca pelos adjetivos disponíveis nos léxicos de sentimento.

Para atribuir a polaridade a um agrupamento, inicialmente agrupa-se os adjetivos em uma lista e calcula-se a polaridade de cada adjetivo pertencente ao agrupamento, utilizando um dos léxicos da lista de léxicos. Adjetivos que não foram encontrados nos léxicos de sentimentos são classificados com polaridade neutra e, portanto, não interferem significamente nos resultados. A polaridade final é determinada pela moda das polaridades obtidas. Caso, nenhum dos adjetivos pertencente ao agrupamento forem encontrados o aspecto será classificado como neutro.

Para o cálculo da moda, empregou-se o $st.mode$. Caso haja um número igual de ocorrências de polaridades positivas e negativas em um agrupamento de adjetivos, a moda será determinada pelo primeiro valor de polaridade que aparecer. A escolha do $st.mode$ deve-se ao fato de que, geralmente, a polaridade de um aspecto é influenciada principalmente pela polaridade do primeiro termo opinativo encontrado.

O uso do agrupamento de adjetivos visa auxiliar os recursos que tem poucos LI. Em situações onde um aspecto é descrito por adjetivos interligados, quando o primeiro adjetivo não for identificado, é possível recorrer à polaridade de sentimento dos adjetivos subsequentes. As métricas de avaliação nas subseções seguintes mostram os resultados quando nosso método de ABSA é utilizado juntamente com o agrupamento de adjetivos.

5.7 Abordagem Proposta

Neste trabalho, propomos analisar a possibilidade de utilização de analisadores de dependência sintática na etapa de identificação de polaridade. Chamamos nossa abordagem de ABSAuDA (do Inglês *Aspect-Based Sentiment Analysis using Dependency Analysis*). Diferentemente do trabalho de Freitas (2015), que usava uma janela de vizinhança centrada nos aspectos, nosso trabalho busca maior exatidão ao utilizar um recurso que informe diretamente qual termo opinativo se relaciona com o aspecto. A Figura 9 apresenta a metodologia utilizada.

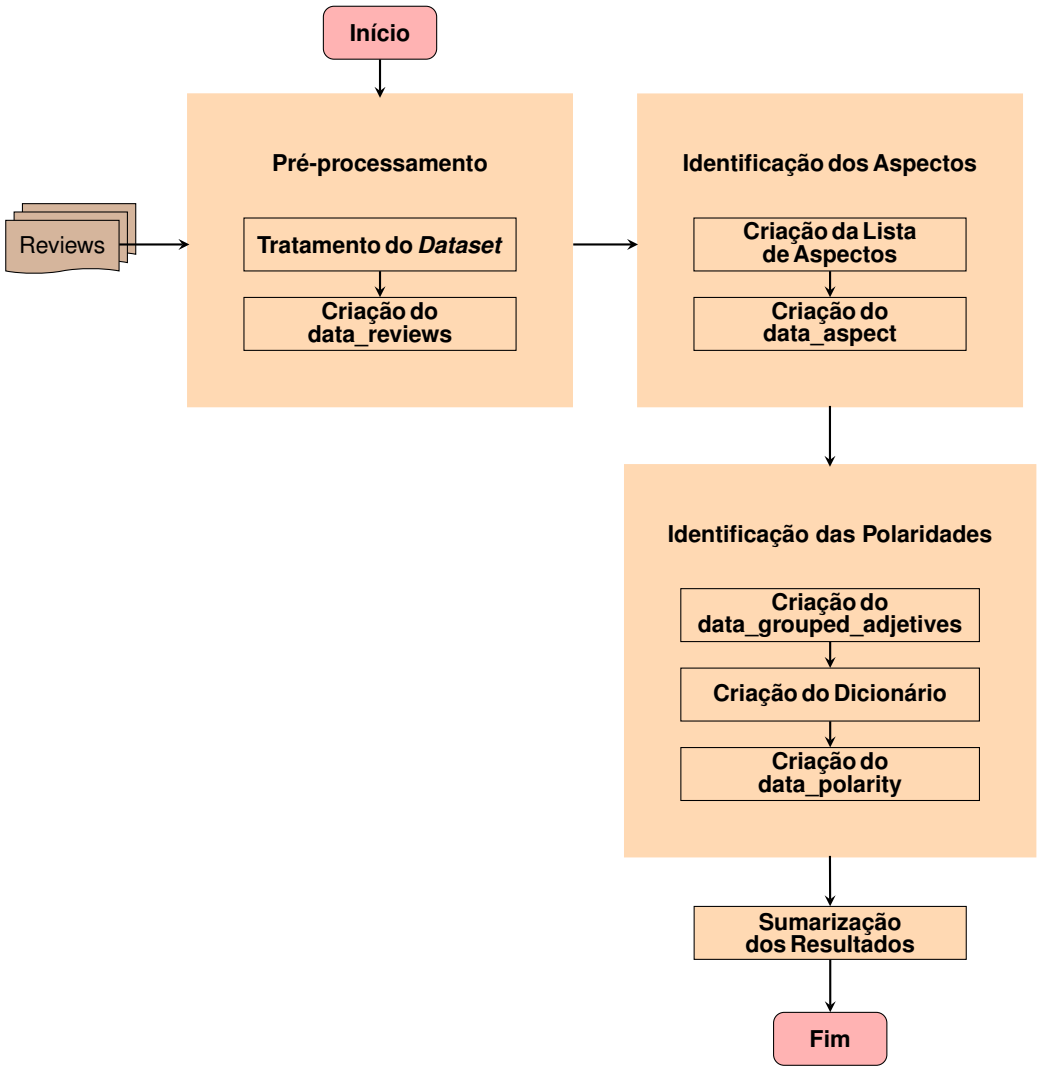

Figura 9 – Fluxograma da metodologia proposta.

A metodologia utilizada no sistema proposto está descrito a seguir.

5.7.1 Pré-processamento do *Dataset*

A etapa de pré-processamento consiste nas seguintes etapas: tratamento do *dataset* e criação do data_reviews.

1. **Tratamento do *dataset*:** consiste em realizar o tratamento das *reviews*. Nessa

etapa foi realizado a correção ortográfica e padronização do texto (remoção de espaços em branco duplos, padronização de valores monetários e conversão do texto em caixa baixa).

2. **Criação do data_reviews:** consiste em criar um *DataFrame*[11] contendo apenas registros que apresentam aspectos simples. O objetivo inicial deste trabalho é focar em aspectos simples para desenvolver o entendimento sobre a ABSA, pois, trabalhar com aspectos compostos, exigiria a aplicação de tratamentos mais complexos e o desenvolvimento de modelos capazes de reconhecer esses aspectos complexos.

5.7.2 Identificação dos Aspectos

Essa tarefa divide-se nas seguintes etapas: criação da lista de aspectos e criação do *dataset* data_aspect.

1. **Criação da lista de aspectos:** consiste em criar uma lista com todos aspectos presentes no *dataset*. Na composição desta lista, realizou-se a lematização de todos os substantivos. A escolha de lematizar apenas os substantivos deve-se ao fato de que as contrações estavam sendo desfeitas em suas palavras originais, resultando na formação de novos tokens (por exemplo, o token 'do' tornando-se 'de o').

2. **Criação do data_aspect:** consiste em criar, a partir do *DataFrame* data_reviews e da lista de aspectos, um novo *DataFrame* contendo os *tokens*, tanto lematizados quanto não lematizados, referentes aos aspectos. Esses *tokens* foram identificados utilizando o método *PhraseMatcher* da biblioteca spaCy.

   O método *PhraseMatcher* é capaz de encontrar, em uma frase, *tokens* ou *spans* (sequências de tokens) a partir de uma lista predefinida de palavras. Uma característica importante do *PhraseMatcher* é que ele pode auxiliar na exclusão de *tokens* que estão contidos dentro de um *span*. Isso é possível porque o resultado do *match*, que é a classe responsável por retornar as correspondências encontradas, fornece as posições de início e fim das correspondências na frase. Utilizando este método, evita-se a inclusão de registros errados no *DataFrame*, que representam correspondências que estejam contidas em outra correspondência.

[11] O *DataFrame* é uma estrutura de dados fornecida pela biblioteca pandas, que organiza os dados em uma tabela bidimensional de linhas e colunas, como uma planilha, permitindo a manipulação eficiente de grandes volumes de dados com diferentes tipos de valores (numéricos, *string*, booleano, etc.), além de oferecer rótulos tanto para as linhas quanto para as colunas, facilitando o acesso e a análise estruturada da informação (MCKINNEY, 2022).

Portanto, se um token inicia na mesma posição que um *span* e termina antes do final deste *span*, ou inicia após o início do *span* e termina antes do seu fim, ou ainda inicia após o início do *span* e termina na mesma posição que o fim do *span*, considera-se que o *token* está contido dentro do *span*. Isso ocorre, por exemplo, com 'casa' e 'casa de banho', 'da' e 'centro da cidade', 'cama' e 'roupa de cama', entre outros.

Cada registro, além dos aspectos, também armazenará as informações referentes as polaridades relacionadas aos aspectos, que foram extraídas do *dataset* usado na competição ABSAPT-2022. A polaridade dos aspectos será utilizada posteriormente para comparar com os resultados encontrados utilizando a metodologia proposta. Foi atribuído a polaridade neutro aos aspectos que foram encontrados, utilizando o método *PhraseMatcher*, e não constavam no *dataset* usado na competição ABSAPT-2022.

5.7.3 Identificação da Polaridade

Essa tarefa divide-se nas seguintes etapas: criação do data_grouped_adjetives, criação do dicionário e criação do data_polarity.

1. **Criação do data_grouped_adjetives:** consiste em criar, a partir do *DataFrame* data_aspect, um novo *DataFrame* contendo todos os adjetivos ligados aos aspectos. Para encontrar os adjetivos ligados aos aspectos utiliza-se uma base de regras baseado na SDA.

2. **Criação do dicionário:** consiste em criar um dicionário destinado a armazenar, em uma lista, as informações obtidas de cada um dos léxicos de sentimentos.

   Esse dicionário armazenará, de cada um dos léxicos de sentimento, todos itens linguísticos classificados como adjetivos. Caso isso não seja possível, em razão de não haver POS tags no recurso, é selecionado todos os itens linguísticos do recurso. Além disso, também será armazenado a polaridade de cada um dos itens linguísticos selecionados.

3. **Criação do data_polarity:** consiste em criar, a partir do *DataFrame* data_grouped_adjetives e do dicionário, um novo *DataFrame* contendo a polaridade dos adjetivos ligados aos aspectos. Caso nenhum dos adjetivos seja encontrado na lista de LI dos léxicos, ou não tenha nenhum adjetivo ligado ao aspecto, será atribuído a polaridade neutro ao aspecto.

5.7.4 Sumarização dos Resultados

Consiste em criar, a partir do *dataset* data_polarity, um novo *dataset* com os resultados das métricas: acurácia, precisão, revocação e medida-f.

5.8 Métricas de Avaliação de Modelos

Modelos de inteligência artificial e de aprendizado de máquina tem seu desempenho avaliado por um conjunto bem definido de métricas. Para entendermos melhor como cada métrica funciona, primeiramente é necessário entendermos alguns conceitos.

A Tabela 1 mostra uma matriz de confusão. Uma matriz de confusão é uma tabela que evidencia os erros e acertos de um modelo, sendo empregada para confrontar o resultado predito pelo modelo com o resultado esperado.

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

Tabela 1 – Matriz de Confusão.

Levando em consideração erros e acertos podemos ter os seguintes rótulos:

- **VP:** VP indica a quantidade de verdadeiro positivo, denotando os casos corretamente classificados como pertencentes à classe **Positivo**;
- **VN:** VN indica a quantidade de verdadeiro negativo, denotando os casos corretamente classificados como pertencentes à classe **Negativo**;
- **FP:** FP indica a quantidade de falso positivo, denotando os casos que o modelo previu um exemplo como pertencente a classe **Positivo** quando, na verdade, o exemplo pertencia a classe **Negativo**. Na estatística, esse erro é considerado **Erro Tipo I**;
- **FN:** FN indica a quantidade de falso negativo, denotando os casos que o modelo previu um exemplo como pertencente a classe **Positivo** quando, na verdade, o exemplo pertencia a classe **Positivo**. Na estatística, esse erro é considerado **Erro Tipo II**;

A partir das métricas básicas, contidas na matriz de confusão, podemos calcular as métricas de avaliação tipicamente utilizadas para classificadores: Acurácia, Precisão, Revocação, e Medida-f.

- **Acurácia:** métrica que avalia a razão do total de acertos sobre o somatório dos valores dos rótulos da matriz de confusão. Ela pode ser obtida com base na equação (3).

$$ACC = \frac{VP + VN}{VP + VN + FP + FN} \tag{3}$$

- **Precisão:** métrica que avalia a razão da quantidade de verdadeiros positivos sobre o somatório de todos os valores positivos. Ela pode ser obtida com base na equação (4).

$$P = \frac{VP}{VP + FP} \tag{4}$$

- **Revocação:** métrica que avalia a capacidade do método de detectar com sucesso resultados classificados como positivo. Ela pode ser obtida com base na equação (5).

$$R = \frac{VP}{VP + FN} \tag{5}$$

- **Medida-f:** média harmônica calculada com base na precisão e na revocação. Ela pode ser obtida com base na equação (6).

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{6}$$

As métricas básicas, e por consequência as derivadas, foram propostas para problemas de classificação binária. Quando temos problemas multiclasse é necessário agregar os resultados dos erros para cada uma das classes em um único valor que represente o desempenho do modelo.

Nos nossos experimentos isso ocorre quando adicionamos a classe de orientação de sentimento neutro, assim a classificação de polaridade conta com exemplos positivos, negativos e neutros.

Cada métrica, exceto a acurácia, pode ser agregada através de diferentes operações: micro, macro e ponderada. Para exemplificar as fórmulas será utilizada a métrica precisão e um modelo com três classes (Cl_1, Cl_2, Cl_3), com N amostras (N_1, N_2, N_3).

- **micro:** é calculada levando em consideração os valores totais. Na média micro, todas as amostras contribuem igualmente para a métrica média final. Ela pode ser obtida com base na equação (7):

$$P_{micro} = \frac{VP_{Cl_1} + VP_{Cl_2} + VP_{Cl_3}}{VP_{Cl_1} + VP_{Cl_2} + VP_{Cl_3} + FP_{Cl_1} + FP_{Cl_2} + FP_{Cl_3}} \tag{7}$$

- **macro:** é calculada através da média da métrica para todas as classes. Na média macro, todas as classes contribuem igualmente para a métrica média final. Ela pode ser obtida com base na equação (8):

$$P_{macro} = \frac{1}{3} \times (P_{Cl_1} + P_{Cl_2} + P_{Cl_3}) \tag{8}$$

- **ponderada:** é calculada através da média ponderada da métrica para todas as classes. Na média ponderada, a contribuição de cada classe para a média é ponderada pelo seu suporte[12]. Ela pode ser obtida com base na equação (9):

$$P_{ponderada} = \frac{P_{Cl_1} \times N_1 + P_{Cl_2} \times N_2 + P_{Cl_3} \times N_3}{N_1 + N_2 + N_3} \tag{9}$$

Para avaliar a abordagem proposta, utilizou-se a métrica medida-f, uma vez que esta métrica estabelece uma relação entre duas métricas importantes: precisão e revocação. A fim de comparar os resultados alcançados através da metodologia proposta com a metodologia de Freitas (2015), adotou-se a média micro. Por outro lado, para a comparação com os resultados da competição ABSAPT-2022 (SILVA et al., 2022), utilizou-se a média macro. Essa diferenciação ocorreu devido ao fato de que a tabela de classificação dos resultados da competição ABSAPT-2022 (SILVA et al., 2022) considera a média macro, enquanto que a média micro foi selecionada para comparação com os resultados de Freitas (2015), pois essa média reflete maior variação entre os resultados, facilitando a comparação entre os léxicos. Através da média micro, cada amostras contribui igualmente para a média final da métrica analisada, auxiliando na visualização da precisão geral do modelo, onde cada previsão correta, independentemente da classe, é igualmente importante.

[12]Suporte é a contagem de exemplos que pertencem a esta classe.

6 RESULTADOS E DISCUSSÃO

Neste capítulo apresentamos os resultados obtidos nos nossos experimentos discutidos no Capítulo 5. Apresentamos os resultados obtidos com variações do algoritmo de análise de sentimento em nível de aspectos utilizando diferentes metodologias de correção ortográfica dos textos, bem como diferentes léxicos de sentimento. As Tabelas 2 e 3 apresentam, respectivamente, as abordagens utilizadas na metodologia proposta e na metodologia de Freitas (2015), onde cada abordagem é representada por um código único.

Tabela 2 – Tabela de Abordagens.

Código	**Abordagem**
sa/sco	sem agrupamento/sem correção ortográfica
ca/sco	com agrupamento/sem correção ortográfica
ca/m	com agrupamento/com correção ortográfica (manual)
ca/ac	com agrupamento/com correção ortográfica (autocorrect)
ca/psc	com agrupamento/com correção ortográfica (pyspellchecker)
ca/LT	com agrupamento/com correção ortográfica (LanguageTool)
ca/ac+m	com agrupamento/com correção ortográfica (autocorrect + manual)
ca/psc+m	com agrupamento/com correção ortográfica (pyspellchecker + manual)
ca/LT+m	com agrupamento/com correção ortográfica (LanguageTool + manual)

Tabela 3 – Tabela de Abordagens da metodologia de Freitas (2015).

Código	**Abordagem**
sco	sem correção ortográfica
m	com correção ortográfica (manual)
ac	com correção ortográfica (autocorrect)
psc	com correção ortográfica (pyspellchecker)
LT	com correção ortográfica (LanguageTool)

A fim de verificar a eficiência de um *ensemble* dos resultados dos léxicos, realizou-se também a inclusão do EnsembleLex, *ensemble* dos resultados obtidos com LeIA, Onto.PT, OpLexicon, ReLi-Lex e SentiLex-PT, aos resultados da metodologia proposta. O EnsembleLex foi gerado utilizando os quatro léxicos que obtiveram os melhores resultados (LeIA, Onto.PT, OpLexicon e SentiLex-PT) e acrescentou-se os resultados

obtidos pelo ReLi-Lex, pois, esse léxico demonstrou bons resultados mesmo tendo tão poucos itens linguísticos e em alguns casos se sobressaiu sobre os demais léxicos.

6.1 Comparação com os resultados da competição ABSAPT-2022 (SILVA et al., 2022)

Nessa seção apresentaremos os resultados obtidos através da metodologia proposta e compararemos com os melhores resultados obtidos na competição ABSAPT-2022 (SILVA et al., 2022). Para realizar essa comparação foi utilizado a medida-f com média macro. As Tabelas 4 e 5 apresentam, respectivamente, os resultados obtidos pelos competidores da competição ABSAPT-2022 (SILVA et al., 2022) e obtidos através da metodologia proposta, para cada um dos tipos de abordagem descritos na Tabela 5, utilizando *reviews* com polaridade neutra.

Tabela 4 – Resultados de medida-f (macro) obtidos pelos competidores da competição ABSAPT-2022 (SILVA et al., 2022).

Time	**Resultado**
Team Deep Learning Brasil	0,818
Team PiLN	0,775
Team UFSCar	0,612
Team PeAm	0,612
Team UFPR	0,612
Team Owl	0,573

Tabela 5 – Resultados de medida-f (macro) obtidos através da metodologia proposta (utilizando *reviews* com polaridade neutra).

Nome do Léxico	**sa/sco**	**ca/sco**	**ca/m**	**ca/ac**	**ca/psc**	**ca/LT**	**ca/ac+m**	**ca/psc+m**	**ca/LT+m**
AffectPT-br	0,271	0,281	0,283	0,275	0,283	0,281	0,276	0,285	0,282
AffectPT-br c/WE	0,330	0,344	0,348	0,330	0,354	0,348	0,332	0,355	0,348
EmoLex	0,270	0,279	0,283	0,282	0,290	0,277	0,284	0,290	0,281
LeIA	0,349	0,362	0,367	0,352	0,374	0,363	0,354	0,375	0,365
LIWC2007pt	0,290	0,298	0,301	0,287	0,299	0,300	0,289	0,301	0,302
Onto.PT	0,385	0,396	0,402	0,401	0,410	0,399	0,406	0,412	0,397
OpLexicon	**0,400**	**0,416**	**0,423**	**0,409**	**0,429**	**0,419**	**0,411**	**0,430**	**0,418**
ReLi-Lex	0,295	0,305	0,308	0,302	0,311	0,304	0,305	0,313	0,303
SentiLex-PT	0,366	0,381	0,385	0,380	0,390	0,381	0,384	0,392	0,379
SentiWordNet-PT-BR	0,346	0,352	0,359	0,354	0,363	0,354	0,356	0,366	0,356
UNILEX	0,308	0,316	0,320	0,293	0,326	0,324	0,293	0,325	0,322
WordNetAffectBR	0,106	0,109	0,109	0,129	0,112	0,108	0,130	0,113	0,110
EnsembleLex	**0,433**	**0,448**	**0,454**	**0,437**	**0,461**	**0,453**	**0,439**	**0,460**	**0,450**

Embora o uso do EnsembleLex melhorou os resultados obtidos pelo léxico OpLexicon, o léxico que obteve o melhor resultado na metologia proposta, ainda sim não

foi suficiente para superar os resultados obtidos pelos competidores da competição ABSAPT-2022 (SILVA et al., 2022). O resultado de medida-f (macro) obtido pelo OpLexicon foi 0,429; enquanto que o EnsembleLex obteve 0,461; em ambos os casos o melhor resultado foi obtido como uso da biblioteca de correção ortográfica pyspellchecker. O resultado obtido pelo EnsembleLex ficaria em sétimo lugar, já que o sexto competidor obteve 0,573 de medida-f (macro).

O uso da correção manual após a aplicação da correção através das bibliotecas de correção ortográfica não resultou em uma grande elevação dos resultados, significando que essas bibliotecas foram capazes de encontrar e corrigir a maioria dos erros ortográficos. Por exemplo, no EnsembleLex, o uso das bibliotecas de correção ortográfica pyspellchecker e LanguageTool resultou em um resultado inferior, enquanto que o uso da biblioteca de correção ortográfica autocorrect junto com a correção ortográfica manual resultou em um leve acréscimo dos resultados.

6.2 Comparação com os resultados obtidos através da metodologia de Freitas (2015)

Nessa seção apresentaremos os resultados obtidos através da metodologia proposta e compararemos com os resultados obtidos através da metodologia de Freitas (2015). Para realizar essa comparação foi utilizado a medida-f com média micro.

6.2.1 Resultados obtidos sem utilizar *reviews* com polaridade neutra

As Tabelas 6 e 7 apresentam, respectivamente, os resultados obtidos através da metodologia de Freitas (2015), por meio das abordagens descritas na Tabela 3, e através da metodologia proposta, por meio das abordagens descritas na Tabela 2.

Tabela 6 – Resultados de medida-f (micro) obtidos através da metodologia de Freitas (2015) (sem utilizar *reviews* com polaridade neutra).

Nome do Léxico	***sco***	***m***	***ac***	***psc***	***LT***
AffectPT-br	0,808	0,809	0,795	0,813	0,811
AffectPT-br c/WE	0,816	0,817	0,802	**0,821**	0,820
EmoLex	0,806	0,807	0,786	0,812	0,812
LeIA	0,812	0,813	0,797	0,818	0,817
LIWC2007pt	0,809	0,811	0,794	0,815	0,813
Onto.PT	0,778	0,778	0,757	0,784	0,771
OpLexicon	0,817	0,818	0,800	**0,824**	0,820
ReLi-Lex	0,815	0,815	0,802	**0,821**	0,819
SentiLex-PT	**0,822**	**0,823**	**0,811**	**0,829**	**0,824**
SentiWordNet-PT-BR	0,740	0,737	0,716	0,737	0,731
UNILEX	0,795	0,796	0,776	0,801	0,798
WordNetAffectBR	0,804	0,805	0,789	0,809	0,808

Tabela 7 – Resultados de medida-f (micro) obtidos através da metodologia proposta (sem utilizar *reviews* com polaridade neutra).

Nome do Léxico	sa/sco	ca/sco	ca/m	ca/ac	ca/psc	ca/LT	ca/ac+m	ca/psc+m	ca/LT+m
AffectPT-br	0,239	0,253	0,258	0,214	0,261	0,259	0,215	0,262	0,258
AffectPT-br c/WE	0,302	0,320	0,329	0,262	0,336	0,331	0,264	0,336	0,332
EmoLex	0,172	0,180	0,184	0,163	0,190	0,175	0,164	0,191	0,180
LeIA	0,334	0,352	0,363	0,299	**0,371**	0,358	0,300	0,373	0,363
LIWC2007pt	0,235	0,246	0,251	0,203	0,254	0,250	0,205	0,254	0,250
Onto.PT	0,338	0,356	0,366	0,334	**0,375**	0,358	0,337	0,374	0,358
OpLexicon	**0,393**	**0,417**	**0,428**	**0,357**	**0,435**	**0,420**	**0,358**	**0,435**	**0,421**
ReLi-Lex	0,250	0,262	0,268	0,219	0,272	0,263	0,222	0,274	0,262
SentiLex-PT	0,314	0,332	0,341	0,300	**0,344**	0,334	0,303	0,346	0,334
SentiWordNet-PT-BR	0,284	0,295	0,305	0,270	0,310	0,297	0,272	0,313	0,301
UNILEX	0,263	0,274	0,282	0,198	0,289	0,282	0,198	0,288	0,281
WordNetAffectBR	0,009	0,010	0,011	0,013	0,012	0,010	0,013	0,012	0,011
EnsembleLex	**0,428**	**0,454**	**0,465**	**0,392**	**0,473**	**0,461**	**0,394**	**0,471**	**0,461**

Na metodologia de Freitas (2015), os quatro léxicos que obtiveram os melhores resultados foram AffectPT-br c/WE, OpLexicon, ReLi-Lex e SentiLex-PT. Enquanto que, na metodologia proposta, os quatro léxicos que obtiveram os melhores resultados foram LeIA, Onto.PT, OpLexicon e SentiLex-PT. Em ambas as abordagens os melhores resultados foram obtidos utilizando a biblioteca de correção ortográfica pyspellchecker. Os melhores resultados obtidos na metodologia de Freitas (2015) foram, respectivamente, 0,821; 0,824; 0,821 e 0,829; enquanto que, na metodologia proposta, os melhores resultados foram, respectivamente, 0,371; 0,375; 0,435 e 0,344. Na metodologia de Freitas (2015), o léxico que obteve o melhor resultado foi o SentiLex-PT. Por outro lado, na metodologia proposta, o léxico que obteve o melhor resultado foi o OpLexicon.

Na metodologia proposta, o uso do EnsembleLex melhorou o resultado obtido pelo léxico OpLexicon. O melhor resultado obtido com EnsembleLex foi 0,473; resultante do uso da biblioteca de correção ortográfica pyspellchecker. O uso da correção manual após a aplicação da correção através das bibliotecas de correção ortográfica não resultou em uma grande diferença entre os resultados, significando que essas bibliotecas foram capazes de encontrar e corrigir a maioria dos erros ortográficos. Por exemplo, no EnsembleLex, o uso das bibliotecas de correção ortográfica pyspellchecker e autocorrect resultaram, respectivamente, em um leve decréscimo e em um leve acréscimo dos resultados, enquanto que o uso da biblioteca de correção ortográfica LanguageTool junto com a correção ortográfica manual não alterou o resultado.

6.2.2 Resultados obtidos utilizando *reviews* com polaridade neutra

As Tabelas 8 e 9 apresentam, respectivamente, os resultados obtidos através da metodologia de Freitas (2015), por meio das abordagens descritas na Tabela 3, e através da metodologia proposta, por meio das abordagens descritas na Tabela 2.

Tabela 8 – Resultados de medida-f (micro) obtidos através da metodologia de Freitas (2015) (utilizando *reviews* com polaridade neutra).

Nome do Léxico	sco	cco(m)	cco(ac)	cco(psc)	cco(LT)
AffectPT-br	0,391	0,392	0,390	0,399	0,377
AffectPT-br c/WE	0,469	0,478	0,449	**0,487**	0,473
EmoLex	0,297	0,298	0,298	0,304	0,295
LeIA	0,419	0,420	0,405	0,427	0,405
LIWC2007pt	0,376	0,376	0,372	0,382	0,358
Onto.PT	0,457	0,468	0,435	**0,468**	0,460
OpLexicon	**0,504**	**0,515**	**0,480**	**0,517**	**0,506**
ReLi-Lex	0,429	0,438	0,407	0,440	0,440
SentiLex-PT	0,466	0,474	0,436	**0,477**	0,471
SentiWordNet-PT-BR	0,411	0,420	0,397	0,419	0,413
UNILEX	0,365	0,366	0,358	0,373	0,350
WordNetAffectBR	0,157	0,157	0,186	0,159	0,155

Tabela 9 – Resultados de medida-f (micro) obtidos através da metodologia proposta (utilizando *reviews* com polaridade neutra).

Nome do Léxico	sa/sco	ca/sco	ca/m	ca/ac	ca/psc	ca/LT	ca/ac+m	ca/psc+m	ca/LT+m
AffectPT-br	0,352	0,363	0,365	0,352	0,369	0,368	0,355	0,371	0,368
AffectPT-br c/WE	0,405	0,419	0,426	0,391	0,432	0,429	0,395	0,434	0,430
EmoLex	0,294	0,300	0,304	0,031	0,308	0,297	0,312	0,310	0,300
LeIA	0,431	0,444	0,452	0,418	**0,460**	0,450	0,421	0,462	0,454
LIWC2007pt	0,348	0,356	0,359	0,341	0,361	0,359	0,344	0,362	0,359
Onto.PT	0,428	0,439	0,446	0,435	**0,452**	0,440	0,439	0,453	0,440
OpLexicon	**0,476**	**0,493**	**0,502**	**0,460**	**0,508**	**0,496**	**0,462**	**0,509**	**0,498**
ReLi-Lex	0,359	0,367	0,372	0,353	0,375	0,369	0,357	0,378	0,368
SentiLex-PT	0,412	0,425	0,431	0,416	**0,435**	0,427	0,420	0,438	0,427
SentiWordNet	0,384	0,389	0,397	0,385	0,401	0,390	0,388	0,404	0,393
UNILEX	0,367	0,373	0,379	0,331	0,384	0,381	0,333	0,384	0,380
WordNetAffectBR	0,160	0,161	0,161	0,193	0,162	0,161	0,195	0,164	0,162
EnsembleLex	**0,504**	**0,521**	**0,531**	**0,486**	**0,536**	**0,528**	**0,488**	**0,536**	**0,528**

Na metodologia de Freitas (2015), os quatro léxicos que obtiveram os melhores resultados foram AffectPT-br c/WE, Onto.PT, OpLexicon e SentiLex-PT. Enquanto que, na metodologia proposta, os quatro léxicos que obtiveram os melhores resultados foram LeIA, Onto.PT, OpLexicon e SentiLex-PT. Em ambas as abordagens os melhores

resultados foram obtidos utilizando a biblioteca de correção ortográfica pyspellchecker. Os melhores resultados obtidos na metodologia de Freitas (2015) foram, respectivamente, 0,487; 0,468; 0,517 e 0,477; enquanto que, na metodologia proposta, os melhores resultados foram, respectivamente, 0,460; 0,452; 0,508 e 0,435. Tanto na metodologia de Freitas (2015) quanto na metodologia proposta o léxico que obteve o melhor resultado foi o OpLexicon.

Na metodologia proposta, o uso do EnsembleLex melhorou o resultado obtido pelo léxico OpLexicon. O melhor resultado obtido com EnsembleLex foi 0,536; resultante do uso da biblioteca de correção ortográfica pyspellchecker. Os resultados da metologia proposta, utilizando *reviews* com polaridade neutra, obtidos através EnsembleLex, superaram os resultados obtidos através da metodologia de Freitas (2015). O efeito de superar a metologia de Freitas (2015), utilizando *reviews* com polaridade neutra, obtidos através EnsembleLex, pode ser resultado da metologia escolhida, onde os aspectos que não tinham adjetivos relacionados ou que os adjetivos relacionados não eram encontrados recebiam a polaridade neutro.

O uso da correção manual após a aplicação da correção através das bibliotecas de correção ortográfica não resultou em uma grande diferença entre os resultados, significando que essas bibliotecas foram capazes de encontrar e corrigir a maioria dos erros ortográficos. Por exemplo, no EnsembleLex, o uso das bibliotecas de correção ortográfica pyspellchecker e autocorrect resultaram, respectivamente, em um leve decréscimo e em um leve acréscimo dos resultados, enquanto que o uso da biblioteca de correção ortográfica LanguageTool junto com a correção ortográfica manual não alterou o resultado.

6.3 Investigando os resultados obtidos na metodologia proposta

Com relação à qualidade da correção ortográfica realizada pelas bibliotecas em Python, observa-se que as bibliotecas pyspellchecker e LanguageTool apresentaram resultados semelhantes, com uma leve vantagem para pyspellchecker, enquanto a biblioteca autocorrect demonstrou o pior desempenho entre as três bibliotecas. Na maioria dos casos, pyspellchecker superou a correção manual, LanguageTool mostrou-se próxima à correção manual, e autocorrect foi inferior à correção manual. Quanto ao tempo de execução da correção ortográfica pelas bibliotecas em Python, baseando-se no *dataset* inteiro, excluindo-se os registros que possuem aspectos compostos, os tempos foram, respectivamente, de 24 minutos para autocorrect, 39 minutos para pyspellchecker e 9 minutos para LanguageTool. Portanto, as bibliotecas que oferecem o melhor custo-benefício, considerando o tempo de execução e a qualidade da correção ortográfica, são a LanguageTool, seguida pela pyspellchecker.

Idealmente, se considerarmos apenas os registros que contêm aspectos relacio-

nados a termos opinativos, obteremos uma melhor compreensão da precisão do modelo desenvolvido. As Tabelas 10 e 11 apresentam os resultados de medida-f (micro) obtidos através da metodologia proposta, excluindo os registros que não possuem adjetivos relacionados aos aspectos.

Tabela 10 – Resultados de medida-f (micro) obtidos através da metodologia proposta (sem utilizar *reviews* com polaridade neutra e excluindo os registros que não possuem adjetivos relacionados aos aspectos).

Nome do Léxico	sa/sco	ca/sco	ca/m	ca/ac	ca/psc	ca/LT	ca/ac+m	ca/psc+m	ca/LT+m
AffectPT-br	0,468	0,460	0,457	0,423	0,456	0,466	0,425	0,463	0,465
AffectPT-br c/WE	0,591	0,582	0,583	0,518	0,587	0,569	0,521	0,593	0,598
EmoLex	0,336	0,328	0,329	0,323	0,331	0,316	0,323	0,337	0,324
LeIA	0,655	0,640	0,643	0,590	0,648	0,644	0,592	0,657	0,654
LIWC2007pt	0,459	0,447	0,445	0,401	0,443	0,450	0,403	0,448	0,451
Onto.PT	0,663	0,648	0,649	0,659	0,654	0,644	0,665	0,660	0,646
OpLexicon	**0,769**	**0,758**	**0,761**	**0,705**	**0,759**	**0,756**	**0,706**	**0,767**	**0,759**
ReLi-Lex	0,490	0,476	0,476	0,433	0,474	0,473	0,437	0,483	0,472
SentiLex-PT	0,616	0,604	0,604	0,593	0,601	0,600	0,597	0,610	0,602
SentiWordNet	0,555	0,537	0,541	0,534	0,542	0,534	0,536	0,552	0,542
UNILEX	0,516	0,499	0,501	0,391	0,504	0,507	0,390	0,508	0,506
WordNetAffectBR	0,018	0,019	0,019	0,026	0,021	0,018	0,026	0,022	0,020
EnsembleLex	**0,839**	**0,826**	**0,827**	**0,775**	**0,825**	**0,830**	**0,777**	**0,831**	**0,831**

Tabela 11 – Resultados de medida-f (micro) obtidos através da metodologia proposta (utilizando *reviews* com polaridade neutra e excluindo os registros que não possuem adjetivos relacionados aos aspectos).

Nome do Léxico	sa/sco	ca/sco	ca/m	ca/ac	ca/psc	ca/LT	ca/ac+m	ca/psc+m	ca/LT+m
AffectPT-br	0,484	0,479	0,475	0,451	0,476	0,485	0,450	0,481	0,481
AffectPT-br c/WE	0,603	0,594	0,595	0,539	0,600	0,609	0,540	0,605	0,605
EmoLex	0,356	0,350	0,352	0,355	0,357	0,341	0,355	0,361	0,361
LeIA	0,659	0,645	0,648	0,599	0,653	0,650	0,600	0,662	0,662
LIWC2007pt	0,476	0,465	0,462	0,475	0,460	0,467	0,426	0,465	0,465
Onto.PT	0,653	0,634	0,635	0,637	0,638	0,630	0,641	0,644	0,644
OpLexicon	**0,757**	**0,744**	**0,747**	**0,692**	**0,747**	**0,743**	**0,691**	**0,755**	**0,755**
ReLi-Lex	0,501	0,489	0,488	0,454	0,488	0,486	0,456	0,495	0,495
SentiLex-PT	0,619	0,607	0,607	0,594	0,604	0,603	0,596	0,613	0,613
SentiWordNet	0,555	0,534	0,537	0,524	0,538	0,529	0,526	0,548	0,548
UNILEX	0,518	0,500	0,501	0,403	0,506	0,509	0,401	0,508	0,508
WordNetAffectBR	0,060	0,068	0,067	0,092	0,072	0,067	0,091	0,071	0,071
EnsembleLex	**0,819**	**0,802**	**0,803**	**0,751**	**0,803**	**0,806**	**0,751**	**0,807**	**0,807**

Como podemos notar, os resultados obtidos nas Tabelas 10 e 11 mantém o mesmo padrão obtido nas Tabelas 7 e 9, com exceção do EnsembleLex, no qual a correção

ortográfica com LanguageTool teve resultados parecidos ou melhores que a correção ortográfica com pyspellchecker. Além disso, os resultados demonstram que, se os termos opinativos forem encontrados, há uma grande chance de o modelo acertar corretamente a polaridade do aspecto.

As Figuras 10 e 11 apresentam os resultados da matriz de confusão obtidos através do léxico OpLexicon e do ensemble EnsembleLex, sem utilizar reviews com polaridade neutra. A Figura 10 apresenta os resultados incluindo os registros que não possuem adjetivos relacionados aos aspectos, enquanto a Figura 11 apresenta os resultados caso fosse excluído os registros que não possuem adjetivos relacionados aos aspectos.

		Predito		
		negativo	neutro	positivo
Real	negativo	68	307	36
	neutro	0	0	0
	positivo	23	886	891

(a) OpLexicon (pyspellcheker).

		Predito		
		negativo	neutro	positivo
Real	negativo	87	283	41
	neutro	0	0	0
	positivo	26	819	955

(b) EnsembleLex (pyspellchecker).

		Predito		
		negativo	neutro	positivo
Real	negativo	68	334	35
	neutro	0	0	0
	positivo	27	938	895

(c) OpLexicon (LanguageTool).

		Predito		
		negativo	neutro	positivo
Real	negativo	87	310	40
	neutro	0	0	0
	positivo	29	861	970

(d) EnsembleLex (LanguageTool).

Figura 10 – Matriz de Confusão da metodologia proposta obtidos através do léxico OpLexicon e do ensemble EnsembleLex (sem utilizar *reviews* com polaridade neutra).

		Predito		
		negativo	neutro	positivo
Real	negativo	68	94	36
	neutro	0	0	0
	positivo	23	151	891

(a) OpLexicon (pyspellcheker).

		Predito		
		negativo	neutro	positivo
Real	negativo	87	70	41
	neutro	0	0	0
	positivo	26	84	955

(b) EnsembleLex (pyspellchecker).

		Predito		
		negativo	neutro	positivo
Real	negativo	68	90	35
	neutro	0	0	0
	positivo	27	159	895

(c) OpLexicon (LanguageTool).

		Predito		
		negativo	neutro	positivo
Real	negativo	87	66	40
	neutro	0	0	0
	positivo	29	82	970

(d) EnsembleLex (LanguageTool).

Figura 11 – Matriz de Confusão da metodologia proposta obtidos através do léxico OpLexicon e do ensemble EnsembleLex (sem utilizar *reviews* com polaridade neutra e excluindo os registros que não possuem adjetivos relacionados aos aspectos).

Como podemos observar, nas Figuras 10 e 11, ao considerarmos somente registros de *reviews* que não possuem polaridade neutra obtemos um total de 2.211 *revi-*

ews, sendo 1.800 positivas e 411 negativas. Das 2.211 *reviews*, ao excluirmos os registros que não possuem adjetivos relacionados aos aspectos, ficamos somente com 1.263 *reviews*, das quais 1.065 são positivas e 198 são negativas. Significando que, aproximadamente, somente 57,12% dos termos opinativos que estavam relacionados com os aspectos foram identificados. Entretanto, embora isso ocorra, a metodologia proposta demonstrou bons resultados, como pode ser observado na Tabela 10.

As Figuras 12 e 13 apresentam os resultados da matriz de confusão obtidos através do léxico OpLexicon e do ensemble EnsembleLex, utilizando *reviews* com polaridade neutra. A Figura 12 apresenta os resultados incluindo os registros que não possuem adjetivos relacionados aos aspectos, enquanto a Figura 13 apresenta os resultados caso fosse excluído os registros que não possuem adjetivos relacionados aos aspectos.

		Predito		
		negativo	neutro	positivo
Real	negativo	68	307	36
	neutro	14	370	20
	positivo	23	885	891

(a) OpLexicon (pyspellcheker).

		Predito		
		negativo	neutro	positivo
Real	negativo	87	283	41
	neutro	16	361	27
	positivo	26	818	955

(b) EnsembleLex (pyspellchecker).

		Predito		
		negativo	neutro	positivo
Real	negativo	68	334	35
	neutro	12	385	22
	positivo	27	938	895

(c) OpLexicon (LanguageTool).

		Predito		
		negativo	neutro	positivo
Real	negativo	87	310	40
	neutro	15	376	28
	positivo	29	861	970

(d) EnsembleLex (LanguageTool).

Figura 12 – Matriz de Confusão da metodologia proposta obtidos através do léxico OpLexicon e do ensemble EnsembleLex (utilizando *reviews* com polaridade neutra).

		Predito		
		negativo	neutro	positivo
Real	negativo	68	94	36
	neutro	14	36	20
	positivo	23	151	891

(a) OpLexicon (pyspellcheker).

		Predito		
		negativo	neutro	positivo
Real	negativo	87	70	41
	neutro	16	27	27
	positivo	26	84	955

(b) EnsembleLex (pyspellchecker).

		Predito		
		negativo	neutro	positivo
Real	negativo	68	90	35
	neutro	11	33	22
	positivo	27	159	895

(c) OpLexicon (LanguageTool).

		Predito		
		negativo	neutro	positivo
Real	negativo	87	66	40
	neutro	14	24	28
	positivo	29	82	970

(d) EnsembleLex (LanguageTool).

Figura 13 – Matriz de Confusão da metodologia proposta obtidos através do léxico OpLexicon e do ensemble EnsembleLex (utilizando *reviews* com polaridade neutra e excluindo os registros que não possuem adjetivos relacionados aos aspectos).

Como podemos observar, nas Figuras 12 e 13, ao considerarmos todas as *reviews* obtemos um total de 2.614 *reviews*, sendo 1.799 positivas, 411 negativas e 404 neutras. Das 2.614 *reviews*, ao excluirmos os registros que não possuem adjetivos relacionados aos aspectos, ficamos somente com 1.333 *reviews*, das quais 1065 são positivas, 198 são negativas e 70 são neutras. Significando que, aproximadamente, somente 50,99% dos termos opinativos que estavam relacionados com os aspectos foram identificados. Entretanto, embora isso ocorra, a metodologia proposta demonstrou bons resultados, como pode ser observado na Tabela 11.

A diferença entre os resultados das matrizes de confusão está fundamentada principalmente na quantidade aspectos que são caracterizados como neutro, dado que a metodologia proposta atribuiu a uma grande quantidade de aspectos a polaridade neutro, em razão da ausência de termos opinativos ligados a esses aspectos. Afim de verificar os resultados obtidos realizou-se uma inspeção nas entradas que continham *reviews* as quais continham aspectos que não estavam relacionados com termos opinativos. As informações encontradas foram sintetizadas nas seções 6.3.1 e 6.3.2.

6.3.1 Influência da marcação errada dos Rótulos de Dependência na Análise de Dependência Sintática

A marcação dos DEP tags de uma SDA é influenciada pela marcação do Analisador de Parte do Discurso (do inglês, *Part-of-Speech parser* - PoS parser) utilizado durante a SDA. Isso ocorre em razão de que o PoS parser realiza a marcação dos PoS tags de uma frase, o DP analisa essa sequência de PoS tags e realiza a marcação dos DEP tags, levando em consideração as informações aprendidas durante o seu treinamento. Quando ocorre do PoS parser errar a marcação de PoS tags, esse erro também influenciará o DP, modificando os resultando da marcação dos DEP tags.

O PoS parser do spaCy apresentou algumas inconsistências relacionadas com a marcação errada dos PoS tags que impactaram os resultados, uma vez que a abordagem proposta depende da marcação correta dos DEP tags. Nas seções a seguir serão discutidos os erros de marcação encontrados.

6.3.1.1 Problemas de marcação: palavras homônimas

O primeiro problema de marcação está relacionado com palavras homônimas, que são palavras que podem possuir a mesma pronúncia e/ou grafia e que naturalmente na língua podem assumir uma classe gramatical diferente dependendo do contexto com que a palavra está sendo empregada. Por exemplo, na frase “Jogo bilhar todos os sábados!”, a palavra “jogo” é classificado como verbo, enquanto que na frase “O jogo pode levar ao vício.”, a palavra “jogo” é classificada como substantivo.

As Figuras 14, 15 e 16 apresentam, respectivamente, as SDA para as *reviews* “Os quartos possuíam cadeiras estragadas.”, “Os quartos possuíam tomadas estragadas.”

e “Os quartos possuíam cadeiras velhas.”. Como é possível notar, essas *reviews* possuem a mesma estrutura sintática. No entanto, nessas *reviews* ocorre um contraste entre as marcações dos PoS tags, causado pela marcação errada dos PoS tags, o que gerou a alteração dos DEP tags.

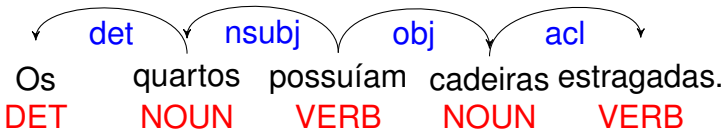

Figura 14 – Primeiro exemplo do erro de marcação do PoS parser. Fonte: Autoria Própria.

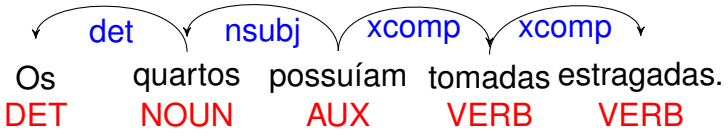

Figura 15 – Segundo exemplo do erro de marcação do PoS parser. Fonte: Autoria Própria.

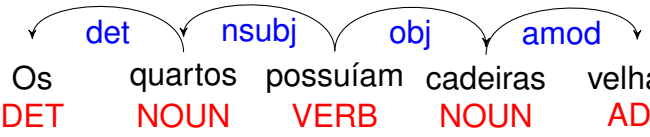

Figura 16 – Terceiro exemplo do erro de marcação do PoS parser. Fonte: Autoria Própria.

Na Figura 14, ocorre a marcação correta do PoS tag do *token* “cadeira” como substantivo (**NOUN**), enquanto que atribui-se verbo (**VERB**) para o PoS tag do token “estragadas”, sendo que o token também pode ser considerado como adjetivo (**ADJ**). Já na Figura 15, o *token* “possuíam”, que é raiz (**Root**) da frase, passou de verbo (**VERB**) para verbo auxiliar (**AUX**), enquanto que os *tokens* “tomadas” e “estragadas” foram considerados como verbo (**VERB**), quando na verdade deveriam ser considerados, respectivamente, como substantivo (**NOUN**) e adjetivo (**ADJ**), como pode ser observado na Figura 16, onde ocorre a marcação correta dos PoS tags.

A diferença entre as marcações dos PoS tags das *reviews* das Figuras 14, 15 e 16 deve-se ao fato das palavras utilizadas nas *reviews* serem classificadas como palavras homônimas. Desse modo, quando o DP do spaCy realiza a marcação dos PoS tags pode ocorrer dele realizar a classificação errada, em razão de que o DP atribui o PoS tag levando em consideração as informações aprendidas no seu treinamento. Se durante o treinamento, o *dataset* utilizado conter poucos exemplos da utilização de uma palavra, que em diferentes contextos assume PoS tag diferentes, quando o DP for classificá-la novamente utilizará um dos PoS tag aprendidos, mesmo que não tenha aprendido exemplos da utilização da palavra com o mesmo contexto do texto analisado, como é o caso de “tomadas”, na Figura 15, que era um substantivo, e foi marcado como passado do verbo “tomar”.

6.3.1.2 Problemas de marcação: palavras com erros ortográficos

As Figuras 17 e 18 apresentam, respectivamente, as SDA para as *reviews* “O hotel e muito bom.” e “O hotel é muito bom.”. Como é possível notar, nessas *reviews* ocorre um contraste entre as marcações dos PoS tags, causado pela marcação errada dos PoS tags, o que gerou a alteração dos DEP tags.

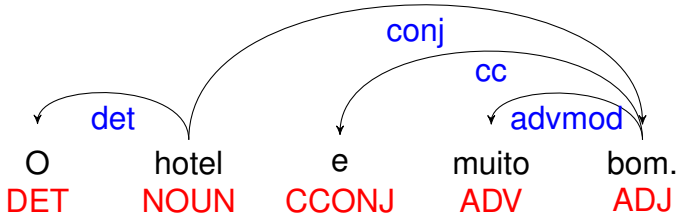

Figura 17 – Quarto exemplo do erro de marcação do PoS parser. Fonte: Autoria Própria.

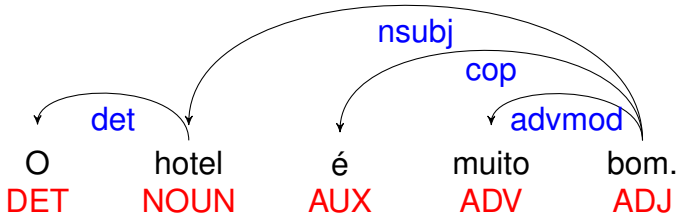

Figura 18 – Quinto exemplo do erro de marcação do PoS parser. Fonte: Autoria Própria.

As Figuras 19 e 20 apresentam, respectivamente, as SDA para as *reviews* “Os quartos sao confortáveis.” e “Os quartos são confortáveis.”. Como é possível notar, nessas *reviews* não ocorre um contraste entre as marcações dos PoS tags, entretanto o DP identificou com uma estrutura diferente, o que gerou a alteração dos DEP tags.

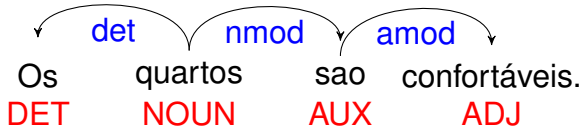

Figura 19 – Sexto exemplo do erro de marcação do PoS parser. Fonte: Autoria Própria.

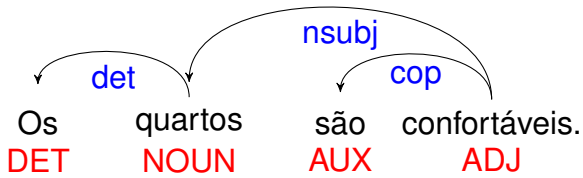

Figura 20 – Sétimo exemplo do erro de marcação do PoS parser. Fonte: Autoria Própria.

Na Figura 17, ocorre a marcação errada do PoS tag do *token* “e” como conjunção coordenativa (**CCONJ**), quando na verdade o *token* deveria ser marcado como verbo auxiliar (**AUX**), como é mostrado na Figura 18. Enquanto que nas Figuras 19 e 20, embora exista palavras com erros ortográficos, não ocorreu mudança nos PoS tags.

A diferença entre as marcações dos PoS tags das *reviews* das Figuras 17 e 18 deve-se ao fato de que palavras que possuem erros ortográficos são confundidas com outras palavras, assumindo assim os seus PoS tags. Entretanto, nas Figuras 19 e 20, embora as palavras tenham sido marcadas com os mesmos PoS tags, foram interpretadas pelo DP como possuindo uma estrutura diferente, como pode ser observado pela diferença entre os DEP tags das *reviews*.

6.3.1.3 Problemas de marcação: palavras que são acompanhadas por símbolos

O terceiro problema de marcação está relacionado com palavras são acompanhadas por símbolos gráficos, como $\#, @, \$, \&$ ou aspas (‘’). As Figuras 21 e 22 apresentam, respectivamente, as SDA para as *reviews* “Bons quartos com aquecimento e Internet liberada.” e “Bons quartos com aquecimento e ‘Internet’ liberada.”. Como é possível notar, nessas *reviews* não ocorre um contraste entre as marcações dos PoS tags, entretanto o DP identificou com uma estrutura diferente, o que gerou a alteração dos DEP tags.

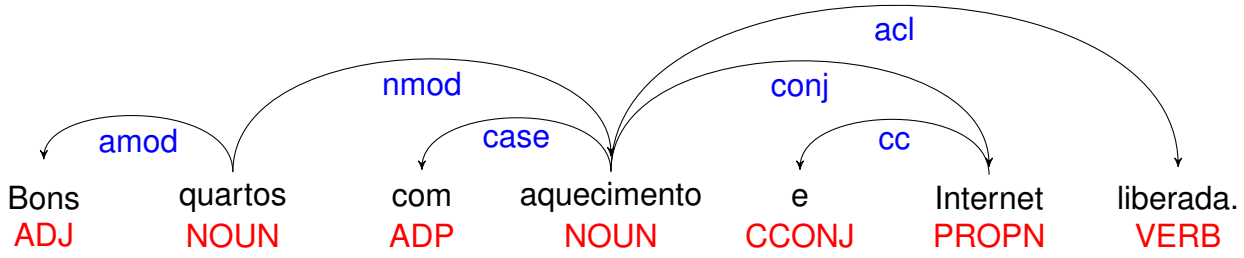

Figura 21 – Oitavo exemplo do erro de marcação do PoS parser. Fonte: Autoria Própria.

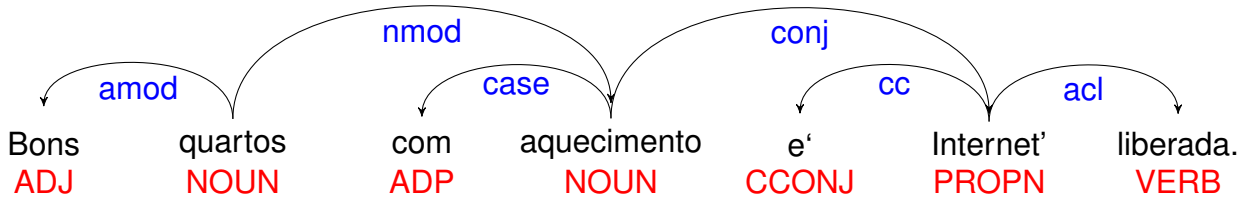

Figura 22 – Nono exemplo do erro de marcação do PoS parser. Fonte: Autoria Própria.

A diferença entre as *reviews* das Figuras 21 e 22 é presença do token “Internet” entre aspas simples, como demonstrado na Figura 22. Isso fez com que o DP identificasse uma nova estrutura sintática, mesmo que os PoS tags das duas *reviews* tenham sido identificados como os mesmos.

6.3.2 Influência da linguagem na metodologia proposta

Uma frase pode ser reescrita de diversas formas, utilizando palavras diferentes ou reestruturando a frase. Quando é utilizado palavras diferentes, como sinônimos e intensificadores, pode ser modificado desde o sentido até a intensidade da frase. Em contrapartida, quando a frase é reestruturada modifica-se a sua estrutura sintática.

Ao realizar a ABSA de *reviews* que diferem da estrutura usual é possível encontrar desafios que podem impactar os resultados, uma vez que a abordagem proposta depende desde a estrutura da frase até as palavras que a compõe. Nas seções a seguir serão discutidos desafios encontrados relacionados com a influência causada pelo uso da linguagem.

6.3.2.1 Influência da linguagem: aspectos ausentes na ontologia

As Figuras 23, 24 e 25 apresentam, respectivamente, as SDA para as *reviews* “O quarto possui Internet Wi-Fi cortesia.”, “O quarto possui Wi-Fi cortesia.” e “O quarto possui Internet cortesia.”. Como é possível notar, as *reviews* das Figuras 23 e 24 podem ser consideradas como formas reescritas da *review* da Figura 25.

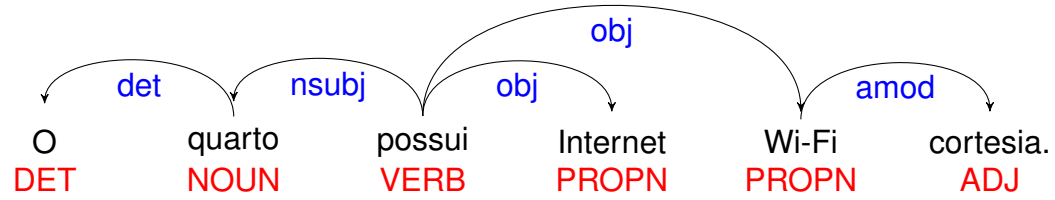

Figura 23 – Primeiro exemplo da influência da linguagem. Fonte: Autoria Própria.

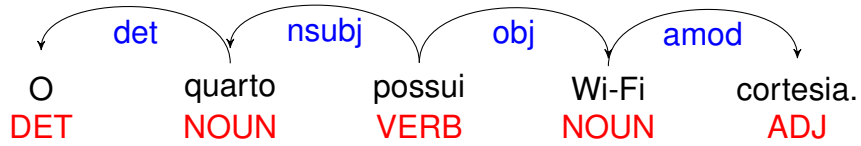

Figura 24 – Segundo exemplo da influência da linguagem. Fonte: Autoria Própria.

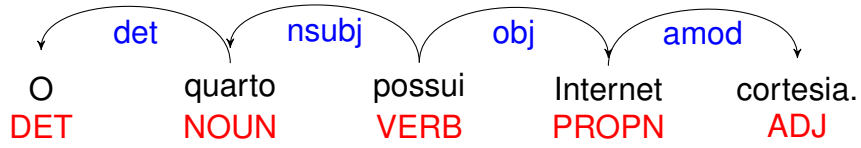

Figura 25 – Terceiro exemplo da influência da linguagem. Fonte: Autoria Própria.

As *reviews* das Figuras 23, 24 e 25 mantêm uma estrutura semelhante, entretanto a diferença dessas *reviews* para a metologia proposta está na presença do aspecto Wi-Fi. Como o aspecto Wi-Fi não está presente na ontologia utilizada, as *reviews* nos quais os termos opinativos estão relacionados somente com o aspecto Wi-Fi são

desconsiderados, pois, para a metologia proposta somente é analisado os aspectos explícitos. Para contornar esse problema pode-se substituir os aspectos “Internet Wi-Fi” e “Wi-Fi” pelo aspecto “Internet”.

6.3.2.2 Influência da linguagem: palavras estrangeiras

Outro fator que vale destacar é o uso de palavras estrangeiras durante a escrita de uma *review*. Como palavras escritas em outros idiomas não estão presentes nos léxicos de sentimentos, os termos opinativos de *reviews* que estão escritas em outros idiomas são desconsiderados, o que resultaria em um aspecto com polaridade neutra, mesmo que aquele termo opinativo tenha uma polaridade. Para contornar esse problema pode-se acrescentar a metologia proposta uma etapa de tradução dos termos opinativos.

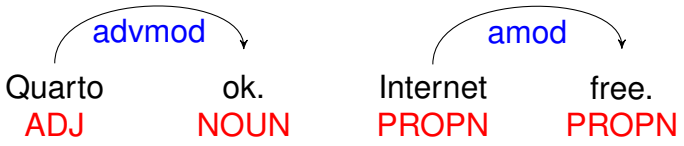

Figura 26 – Quarto exemplo da influência da linguagem. Fonte: Autoria Própria.

A Figura 26 apresenta a SDA para a *review* “Quarto ok. Internet free.”, um exemplo de *review* no qual os termos opinativos estão escritas em outra língua. Na *review* as palavras estrangeiras identificadas foram “ok” e “free”, que podem ser facilmente substituídas, respectivamente, por “bom” e “grátis”. A substituição das palavras estrangeiras contribuiriam positivamente para a sumarização dos resultados.

6.3.2.3 *Influência da linguagem: enumeração*

Uma *review* também pode ser estruturada através de uma lista composta de informações detalhadas, onde a cada aspecto destacado na lista é atribuído a mesma polaridade da lista. A Figura 27 apresenta a SDA para a *review* “Pontos negativos: não limpam o quarto, não tem elevador. Ponto positivo: localização!!!”, um exemplo de *review* no qual os aspectos destacados estão dispostos em listas.

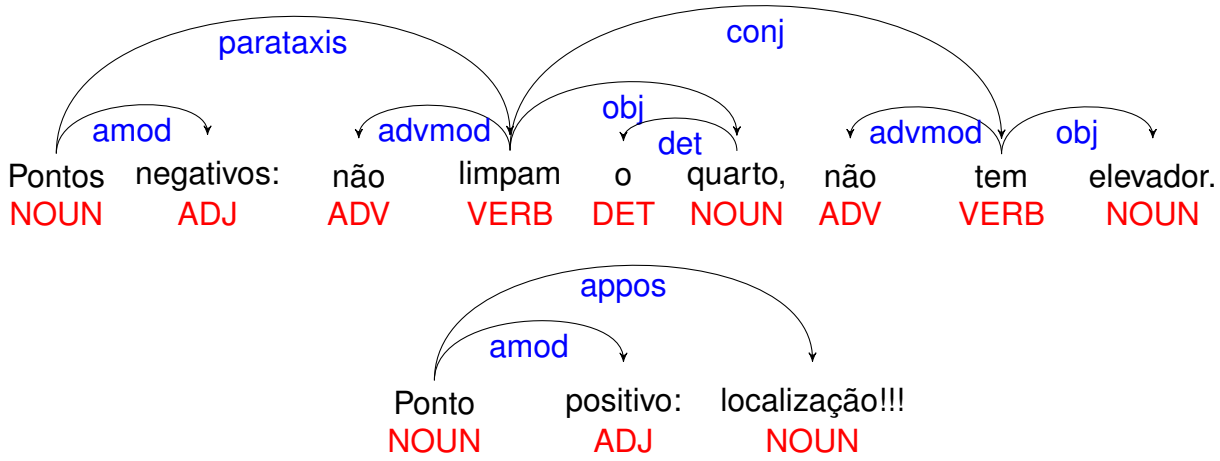

Figura 27 – Quinto exemplo da influência da linguagem. Fonte: Autoria Própria.

Reviews estruturadas em listas são mais difíceis de identificar a polaridade dos aspectos com modelos que utilizam SDA, pois na estruturação do DP desse tipo de *review* o *token* que caracteriza a polaridade fica muito afastado do aspecto e acaba não estando relacionado com os aspectos, como é demonstrado na Figura 28.

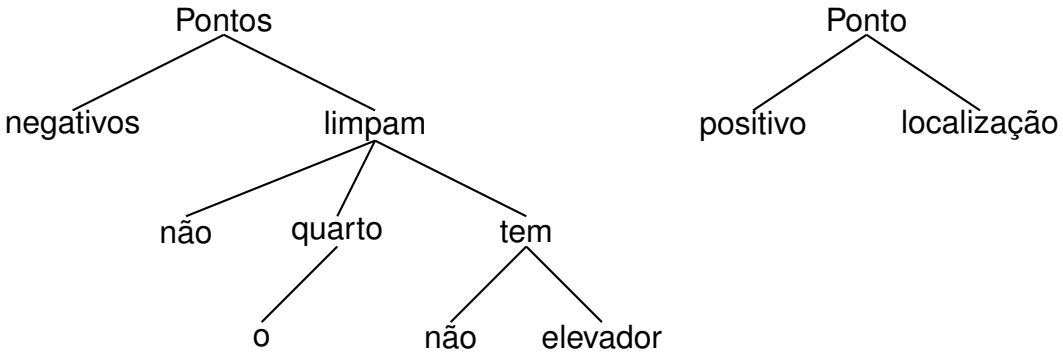

Figura 28 – Representação da Gramática de Dependência para a sentença “Pontos negativos: não limpam o quarto, não tem elevador. Ponto positivo: localização!!!”.

Na *review* da Figura 27 é possível identificar três aspectos: limpeza, elevador e localização. O aspecto limpeza é implícito e é caracterizado pela oração “não limpam o quarto”, enquanto que os aspectos elevador e localização estão explícitos na *review*. Como o aspecto “elevador” está presente na oração “não tem elevador” e o *token* que está relacionado com ele, que é o *token* “tem”, não está presente nos SL, precisaríamos percorrer o grafo em busca do termo opinativo que o caracteriza, mas, como

demonstrado na Figura 28, o termo opinativo que o caracteriza encontra-se em um ramo diferente, fazendo com que seja difícil encontrá-lo. Esse problema é melhor demonstrado em “Ponto positivo: localização!!!”, onde o aspecto “localização” e o token “positivo” ficam em ramos opostos, como demonstrado na Figura 28.

6.3.3 Influência da linguagem: expressões multi-vocabulares e palavras compostas

Embora o DP faça a marcação correta dos tokens de uma *review*, isso não quer dizer que a polaridade do aspecto será identificada, pois a polaridade depende de dois fatores: a identificação dos termo opinativos e a extração da polaridade dos SL. Para que a extração da polaridade ocorra de maneira satisfatória é preciso que o termo opinativo esteja presente em algum dos SL. No entanto, nem todas os termos opinativos são encontrados nos SL, como é o caso de MWE (por exemplo: expressões idiomáticas e expressões regionais) e palavras compostas.

As Figuras 29, 30 e 31 apresentam, respectivamente, as SDA para as *reviews* “O hotel é bomzinho. Custo-benefício: vale a pena!”, “Os quartos são superconfortáveis. O serviço é de primeiríssima.” e “o hotel é superbem localizado. Atendimento e serviço nota 10!”. Como é possível notar, cada *review* possui a sua própria estrutura sintática, sendo que o ponto em comum entre elas é o uso de MWE e palavras compostas.

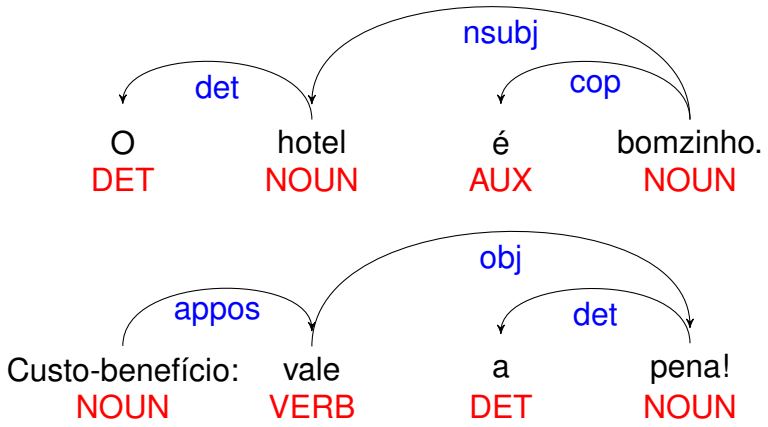

Figura 29 – Sexto exemplo da influência da linguagem. Fonte: Autoria Própria.

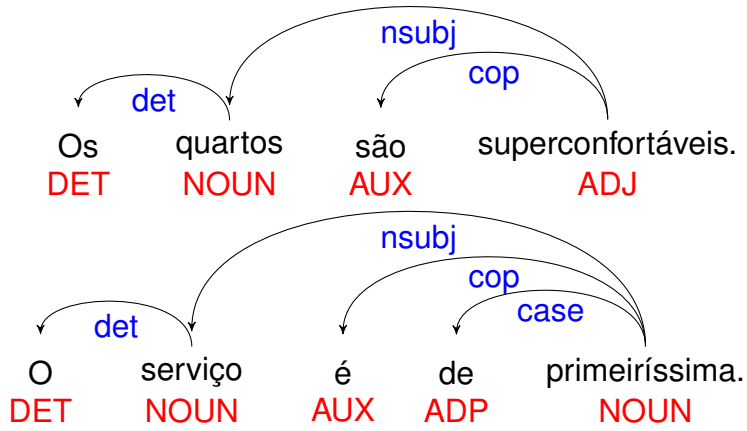

Figura 30 – Sétimo exemplo da influência da linguagem. Fonte: Autoria Própria.

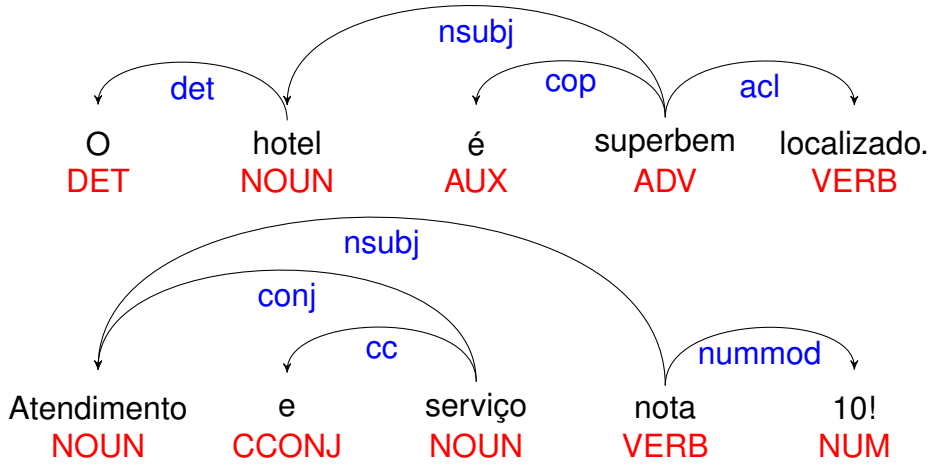

Figura 31 – Oitavo exemplo da influência da linguagem. Fonte: Autoria Própria.

Nas *reviews* das Figuras 29, 30 e 31 é possível identificar as seguintes MWE e palavras compostas: “bomzinho”, “vale a pena”, “superconfortáveis”, “de primeiríssima”, “superbem” e “nota 10”. O problema de se utilizar MWE e palavras compostas é que elas são dificilmente encontradas nos léxicos de sentimentos, pois, não são usualmente utilizadas. Dessa forma, as *reviews* onde isso ocorre são atribuídos a polaridade neutra.

Para contornar esse problema é possível substituir as MWE por palavras que produzem efeitos semelhantes, por exemplo “vale a pena”, “de primeiríssima” e “nota 10” poderiam ser substituídas por “ótimo”. Enquanto que as palavras compostas poderiam ser reduzidas ao seu radical, por exemplo “bomzinho”, “superconfortáveis” e “superbem” poderiam ser substituídas, respetivamente, para “bom”, “confortáveis” e “bem”. A complexidade desse desafio está na substituição das MWE. Para realizar a substituição de MWE precisaríamos encontrar ou criar um dicionário de sinônimos, no qual cada expressão conhecida corresponderia a uma palavra mais genérica, que seria facilmente encontrada nos SL.

6.4 Influência do domínio, quantidade de itens linguísticos e tipo de léxico

A Tabela 12 apresenta o melhor resultado obtido na metodologia proposta sem utilizar *reviews* com polaridade neutra e excluindo os registros que não possuem adjetivos relacionados aos aspectos). O resultado foi obtido após realizar a correção ortográfica através da biblioteca pyspellchecker e utilizar o agrupamento de adjetivos.

Tabela 12 – Melhor resultado obtido com o modelo proposto.

Nome do Léxico	**Quantidade de Itens Linguísticos**	**medida-f (micro)**
AffectPT-br	1135	0,456
AffectPT-br c/WE	17617	0,587
EmoLex	11231	0,331
LeIA	5779	0,648
LIWC2007pt	27491	0,443
Onto.PT	14039	0,654
OpLexicon	**32119**	**0,759**
ReLi-Lex	**345**	**0,474**
SentiLex-PT	7010	0,601
SentiWordNet-PT-BR	6007	0,542
UNILEX	3845	0,504
WordNetAffectBR	289	0,021

Ao observarmos a Tabela 12 nota-se que os resultados obtidos pelos léxicos ReLi-Lex e OpLexicon são próximos. Isso levantou uma questão importante: a quantidade de itens linguísticos de um léxico influencia os resultados do modelo? Se considerarmos a razão da quantidade de termos de um léxico de sentimento pela métrica avaliada, obteremos uma relação de proporcionalidade. Assim, quanto menor a quantidade de termos de um léxico de sentimento e maior é o valor da métrica avaliada, melhor é aquele recurso em relação à métrica avaliada.

Observando, por exemplo, os léxicos de sentimento OpLexicon e ReLi-Lex podemos notar a discrepância entre os números de itens linguísticos. Enquanto que o OpLexicon possui 32.119 itens linguísticos, ReLi-Lex possui apenas 345, isso é, OpLexicon é, aproximadamente, 93 vezes maior que ReLi-Lex.

Podemos, por exemplo, calcular a razão de proporcionalidade entre a medida-f e o número de itens linguísticos. Para o OpLexicon, obtém-se um valor de $2,363 \times 10^{-5}$, enquanto que para o ReLi-Lex, o valor é $1,374 \times 10^{-3}$. Desse modo, quanto menor a razão, menor é a eficácia do léxico de sentimento para o domínio analisado. Embora o domínio do ReLi-Lex seja para livros, já que ele foi construído nesse contexto, ele mostrou-se bastante relevante, mesmo possuindo tão poucos termos.

Vale salientar que o processo de desenvolvimento do recurso linguístico influencia

significativamente os resultados obtidos. Por exemplo, um léxico de sentimento criado automaticamente ou por meio de tradução de uma língua estrangeira pode apresentar menor eficiência em comparação a um léxico desenvolvido manualmente. Isso ocorre porque esses recursos podem não ter sido submetidos a um processo de revisão, o qual possibilitaria a correção de eventuais erros. Adicionalmente, o domínio do léxico também pode afetar os resultados, uma vez que, dependendo do contexto em que é utilizado, a mesma palavra pode receber classificações de polaridade diferentes, considerando que uma palavra pode ser utilizada com diversos sentidos diferentes.

7 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

7.1 Conclusões

Este trabalho se propôs a avaliar a utilização de grafos de dependência sintática em um sistema de Análise de Sentimento Baseada em Aspectos para a Língua Portuguesa. A ideia básica é modificar o método proposto por Freitas (2015), substituindo a busca por proximidade pelas relações de dependência expressamente presentes no grafo de dependência sintática.

Além disso, realizamos experimentos com doze (12) léxicos de sentimento disponíveis para a Língua Portuguesa. Esse tipo de abordagem é de suma importância para abordagens baseadas em léxicos de sentimento, visto que os termos opinativos podem sofrer troca de orientação de sentimento em diferentes domínios. Por exemplo, comida **quente** geralmente é considerado bom, já cerveja **quente** geralmente é considerado ruim. Ou seja, o mesmo adjetivo possui polaridade diferente em dois domínios ligeiramente diferentes.

Nossos experimentos foram realizados utilizando o *corpus* da competição ABSAPT-2022 (SILVA et al., 2022), realizada no evento IBERLEF de 2022. Nesta competição a *task2* consistia da Identificação da Polaridade do Sentimento no *corpus* proposto. Os melhores resultados obtidos pelos ganhadores foram de 0,818 para a métrica F1.

Cabe destacar que os modelos que os trabalhos participantes da competição ABSAPT-2022 utilizaram aprendizado de máquina em suas soluções, todos eles dependem de modelos de linguagem enquanto que a abordagem de Freitas (2015) e a nossa modificação dependem de léxicos de sentimento e de uma ontologia de domínio.

No contexto de linguagens com poucos recursos, os grandes modelos de linguagem se apresentam como um recurso promissor para o desenvolvimento de novas soluções para as tarefas de PLN, pois, são modelos treinados utilizando grandes quantidades de dados e possuem uma infinidade de parâmetros. No entanto, esses modelos demandam um grande custo computacional para o seu treinamento, Meyer et al.

(2023) discute os custos financeiros e ambientais para o treinamento de Modelos de Linguagem. Os autores apresentam uma especulação sobre o custo de treinamento do ChatGPT pela OpenAI, que giraria em torno de US$5 milhões só pelas horas de processamento em 10.000 unidades gráficas NVIDIA V100. Outro ponto destacado é o grande consumo energético e a emissão de carbono necessários para o treinamento desses modelos.

Por outro lado, as abordagens baseadas em léxicos podem apresentar um desempenho aceitável consumindo muito menos recursos computacionais. Os modelos gerados por essas abordagens são do tipo **modelos baseados em conhecimento**, o que quer dizer que eles não necessitam de grandes estágios de treinamento, embora necessitem de um "engenheiro de conhecimento" para construir as regras necessárias para o funcionamento do sistema.

Como pudemos notar, o método de Freitas (2015) atingiu desempenho geral superior que os campeões da competição ABSAPT-2022, em termos de métrica F1. Isto indica que sistemas baseados em léxicos podem ter desempenho relevante mesmo consumindo muito menos recursos financeiros e ambientais, além de demandar uma quantidade de recursos muito inferior aos modelos dependentes de grandes modelos de linguagem.

Uma discussão importante é a capacidade de abordagens baseadas em modelos de linguagem serem reutilizados para diversas tarefas, o que, no geral, poderia diluir os custos citados anteriormente. No entanto, o tamanho do mercado das linguagens de baixos recursos, e das economias dos países que falam essas línguas, não tem sido suficientemente grande para que empresas invistam no desenvolvimento de novos modelos de linguagem exclusivos para o português, por exemplo.

Quanto aos resultados do ABSAuDA com relação ao modelo de Freitas (2015), cabe ressaltar que a abordagem da autora envolve estágios adicionais de processamento de regras linguísticas que não implementamos neste trabalho, como, por exemplo, a inversão de polaridade quando um termo de negação influencia uma opinião. Por exemplo: "Eu **não** gosto de cama deste hotel". Em outras palavras, o ABSAuDA ainda é um modelo mais simples, com menos regras que o modelo de Freitas (2015).

Como podemos ver nos resultados, o agrupamento de adjetivos e a correção ortográfica contribuem para as melhorias na abordagem ABSAuDA. O uso do EnsembleLex possibilitou para que fosse alcançado um resultado superior ao resultado do melhor léxico, contribuindo positivamente na abordagem ABSAuDA. Além disso, aparentemente, a limitação de qualidade das ferramentas de análise de dependência sintática e de marcação de parte do discurso da biblioteca spaCy prejudicaram o desenvolvimento da abordagem proposta. Um indício que suporta essa afirmação é que a abordagem de Freitas (2015) realizava uma busca de adjetivos por vizinhança, sem a garantia da relação de dependência entre o aspecto e o adjetivo (termo opinativo).

Porém, cabe salientar, que o método de Freitas (2015) também possuí outras regras linguísticas que podem ter influenciado nos resultados. Nossos experimentos carecem de uma análise por ablação (do Inglês, *ablation analysis*) no trabalho de Freitas (2015) para garantirmos o impacto da contribuição de cada etapa adicional do trabalho de Freitas (2015).

Consideramos que uma das contribuições deste trabalho foi a demonstração de problemas no *dataset* anotado, bem como a severa limitação dos anotadores de parte do discurso e de dependência sintática da biblioteca spaCy.

7.2 Trabalhos Futuros

Diversas frentes de trabalho se abrem a partir dos resultados obtidos no Capítulo 6. Discutiremos alguns deles nessa seção.

Acreditamos que a baixa qualidade do *Dependency Parser* disponível para português, no pacote spaCy, pode estar prejudicando o significativamente o desempenho dos modelos testados. Assim, nos parece trivial que a criação de um melhor analisador de dependência precisa ser desenvolvido para a língua portuguesa.

Os léxicos de sentimento utilizados tem, em sua generalidade, um direcionamento a domínios específicos. No contexto de construção de um método geral de análise de sentimento baseado em léxicos fica evidenciada a necessidade de lidar com a polissemia[1].

Em comparação com o trabalho de Freitas (2015), ainda existem outras regras linguísticas a serem implementadas que poderiam melhorar o desempenho do ABSAuDA.

A aplicação a outros domínios linguísticos faz-se necessária para um melhor entendimento das capacidades de generalização dos modelos baseados em léxicos. Como discutimos acima, a polissemia e o analisador de dependência podem ter impacto significativo na abordagem proposta neste trabalho. Assim, se trocarmos o domínio de aplicação, os resultados podem sofrer grande variação, para melhor ou pior.

[1] Polissemia é o nome dado ao fenômeno linguístico que ocorre quando uma mesma palavra ou expressão apresenta significados distintos dependendo do contexto em que aparece.

REFERÊNCIAS

AARTS, K. Parsimonious methodology. **Methodological Innovations Online**, [S.l.], v.2, n.1, p.2–10, 2007.

AIZAWA, A. An information-theoretic perspective of tf–idf measures. **Information Processing & Management**, [S.l.], v.39, n.1, p.45–65, 2003.

AKHTAR, M. S.; KUMAR, A.; EKBAL, A.; BHATTACHARYYA, P. A hybrid deep learning architecture for sentiment analysis. In: COLING 2016, THE 26TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS: TECHNICAL PAPERS, 2016. **Proceedings...** [S.l.: s.n.], 2016. p.482–493.

ALFIO GLIOZZO, C. S. a. **Semantic Domains in Computational Linguistics**. 1.ed. [S.l.]: Springer, 2009.

ALLEMANG, D.; HENDLER, J. **Semantic web for the working ontologist**: effective modeling in RDFS and OWL. [S.l.]: Elsevier, 2011.

ALMEIDA, R. J. A. **LeIA - Léxico para Inferência Adaptada**. [S.l.]: GitHub, 2018. `https://github.com/rafjaa/LeIA`.

ALTINOK, D. **Mastering spaCy**: An end-to-end practical guide to implementing NLP applications using the Python ecosystem. [S.l.]: Packt Publishing Ltd, 2021.

AMBATI, B. R. Hindi dependency parsing and treebank validation. **Master's**, [S.l.], 2011.

ANTONIOU, G.; HARMELEN, F. v. Web ontology language: Owl. **Handbook on ontologies**, [S.l.], p.91–110, 2009.

ARAÚJO, M.; GONÇALVES, P.; CHA, M.; BENEVENUTO, F. iFeel: a system that compares and combines sentiment analysis methods. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 23., 2014. **Proceedings...** [S.l.: s.n.], 2014. p.75–78.

ARAUJO, M. et al. iFeel 2.0: A Multilingual Benchmarking System for Sentence-Level Sentiment Analysis. In: INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA, 2016. **Proceedings...** [S.l.: s.n.], 2016. v.10, n.1, p.758–759.

ARAUJO, M.; REIS, J.; PEREIRA, A.; BENEVENUTO, F. An evaluation of machine translation for multilingual sentence-level sentiment analysis. In: ACM SYMPOSIUM ON APPLIED COMPUTING, 31., 2016. **Proceedings...** [S.l.: s.n.], 2016. p.1140–1145.

ARIEW, R. **OCKHAM'S RAZOR**: A HISTORICAL AND PHILOSOPHICAL ANALYSIS OF OCKHAM'S PRINCIPLE OF PARSIMONY. [S.l.]: University of Illinois at Urbana-Champaign, 1976.

ARP, R.; SMITH, B.; SPEAR, A. D. **Building ontologies with basic formal ontology**. [S.l.]: Mit Press, 2015.

ATSERIAS, J. et al. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. In: LANGUAGE RESOURCES AND EVALUATION (LREC 2006), 2006, Genoa, Italy. **Proceedings...** [S.l.: s.n.], 2006.

AYE MAR, A.; SHIRAI, K. Automatic Construction of an Annotated Corpus with Implicit Aspects. In: THIRTEENTH LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 2022, Marseille, France. **Proceedings...** European Language Resources Association, 2022. p.6985–6991.

BAADER, F. **The description logic handbook**: Theory, implementation and applications. [S.l.]: Cambridge university press, 2003.

BAADER, F.; HORROCKS, I.; LUTZ, C.; SATTLER, U. **Introduction to description logic**. [S.l.]: Cambridge University Press, 2017.

BACCIANELLA, S.; ESULI, A.; SEBASTIANI, F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC'10), 7., 2010, Valletta, Malta. **Proceedings...** European Language Resources Association (ELRA), 2010.

BAEZA-YATES, R. A.; RIBEIRO-NETO, B. **Modern Information Retrieval**. 1.ed. Harlow, England: Addison-Wesley, 1999.

BAEZA-YATES, R. A.; RIBEIRO-NETO, B. **Modern Information Retrieval**: The Concepts and Technology Behind Search. 2.ed. Harlow, England: Addison-Wesley, 2011.

BALAGE FILHO, P. P.; PARDO, T. A. S.; ALUÍSIO, S. M. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 9., 2013. **Proceedings...** [S.l.: s.n.], 2013.

BARION, E. C. N.; LAGO, D. Mineração de textos. **Revista de ciências exatas e tecnologia**, [S.l.], v.3, n.3, p.123–140, 2008.

BARTA, R. et al. Covering the semantic space of tourism: an approach based on modularized ontologies. In: WORKSHOP ON CONTEXT, INFORMATION AND ONTOLOGIES, 1., 2009. **Proceedings...** [S.l.: s.n.], 2009. p.1–8.

BASTOS, P. T. T. **GitHub - Pedro-Thales/SentiWordNet-PT-BR**: Projeto de um dicionário de sentimentos em português. Accessado em: [06 fev. 2023]. Disponível em: <https://github.com/Pedro-Thales/SentiWordNet-PT-BR>.

BATISTA, D.; SILVA, M. J. A statistical study of the WPT05 crawl of the Portuguese Web. In: FALA 2010 VI JORNADAS EN TECNOLOGÍA DEL HABLA AND II IBERIAN SLTECH WORKSHOP, VIGO, SPAIN, 2010. **Anais...** [S.l.: s.n.], 2010.

BENTIVOGLI, L.; FORNER, P.; MAGNINI, B.; PIANTA, E. Revising the wordnet domains hierarchy: semantics, coverage and balancing. In: OF THE WORKSHOP ON MULTILINGUAL LINGUISTIC RESOURCES, 2004. **Proceedings...** [S.l.: s.n.], 2004. p.94–101.

BERNARD, J. (Ed.). **The Macquarie Thesaurus**. Sydney, Australia: Macquarie Library, 1986.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific american**, [S.l.], v.284, n.5, p.34–43, 2001.

BICK, E. **The parsing system palavras**: Automatic grammatical analysis of Portuguese in a constraint grammar famework. Aarhus: Aarhus University Press, 2000.

BICK, E.; HABER, R.; SANTOS, D.; AFONSO, S. Floresta sintá (c) tica: um treebank para o português. In: **Actas do XVII Encontro da Associação Portuguesa de Linguística (Lisboa, 2-4 de Outubro de 2001)**. [S.l.]: Associação Portuguesa de Linguística (APL), 2002. p.533–545.

BOND, F.; PAIK, K. A survey of wordnets and their licenses. In: GLOBAL WORDNET CONFERENCE (GWC 2012), 6., 2012. **Proceedings...** [S.l.: s.n.], 2012. p.64–71.

BOND, F.; VOSSEN, P.; MCCRAE, J. P.; FELLBAUM, C. Cili: the collaborative interlingual index. In: GLOBAL WORDNET CONFERENCE (GWC), 8., 2016. **Proceedings...** [S.l.: s.n.], 2016. p.50–57.

BOYD, R. L. **MEH**: Meaning Extraction Helper (Version 2.3.02) [Software].

BRADLEY, M. M.; LANG, P. J. **Affective norms for English words (ANEW)**: Instruction manual and affective ratings. [S.l.]: Technical report C-1, the center for research in psychophysiology ..., 1999.

BRANCO, A. et al. **The Portuguese Language in the Digital Age**. [S.l.]: Springer, 2012.

BRANTS, T.; FRANZ, A. **Web 1T 5-gram Version 1**. Philadelphia: Linguistic Data Consortium, 2006. LDC2006T13, Web Download. Disponível em: <https://catalog.ldc.upenn.edu/LDC2006T13>.

BROWN, S.; FLORES, J. **The A to Z of Medieval Philosophy and Theology**. [S.l.]: Scarecrow Press, 2010. (The A to Z Guide Series).

BRUCKSCHEN, M. et al. Anotaçao lingüıstica em xml do corpus pln-br. **Série de relatórios do NILC, ICMC-USP**, [S.l.], 2008.

CAI, H. et al. **MEMD-ABSA**: A Multi-Element Multi-Domain Dataset for Aspect-Based Sentiment Analysis.

CAMBRIA, E.; SCHULLER, B.; XIA, Y.; HAVASI, C. New avenues in opinion mining and sentiment analysis. **IEEE Intelligent systems**, [S.l.], v.28, n.2, p.15–21, 2013.

CARRERAS, X.; CHAO, I.; PADRó, L.; PADRó, M. FreeLing: An Open-Source Suite of Language Analyzers. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC'04), 4., 2004. **Proceedings...** [S.l.: s.n.], 2004.

CARVALHO, F. et al. Evaluation of the Brazilian Portuguese version of linguistic inquiry and word count 2015 (BP-LIWC2015). **Language Resources and Evaluation**, [S.l.], p.1–20, 2023.

CARVALHO, F.; SANTOS, G.; GUEDES, G. P. AffectPT-br: an Affective Lexicon based on LIWC 2015. In: INTERNATIONAL CONFERENCE OF THE CHILEAN COMPUTER SCIENCE SOCIETY (SCCC), 2018., 2018. **Anais...** [S.l.: s.n.], 2018. p.1–5.

CARVALHO, P. C. Q. d. F. Análise e representação de construções adjectivais para processamento automático de texto: adjectivos intransitivos humanos. , [S.l.], 2007.

CARVALHO, P.; SILVA, M. J. SentiLex-PT: Principais características e potencialidades. **Oslo Studies in Language**, [S.l.], v.7, n.1, 2015.

CHAPELLE, C. A. (Ed.). **The concise encyclopedia of applied linguistics**. Hoboken, NJ: Wiley Blackwell, 2020.

CHAVES, M.; GOMES, R.; PEDRON, C. Decision making based on Web 2.0 data: the small and medium hotel management. In: 2018 37TH , 2012. **Anais...** [S.l.: s.n.], 2012.

CHAVES, M. S.; FREITAS, L.; VIEIRA, R. Hontology: A multilingual ontology for the accommodation sector in the tourism industry. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE ENGINEERING AND ONTOLOGY DEVELOPMENT (KEOD 2012), 4., 2012, Barcelona, Espanha. **Proceedings...** SciTePress, 2012. p.149–154.

CHAVES, M.; TROJAHN, C. Towards a multilingual ontology for ontology-driven content mining in social web sites. In: SCITEPRESS, 2010. **Anais...** [S.l.: s.n.], 2010.

CIERI, C.; MAXWELL, M.; STRASSEL, S.; TRACEY, J. Selection criteria for low resource language programs. In: TENTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC'16), 2016. **Proceedings...** [S.l.: s.n.], 2016. p.4543–4549.

COHN, M. A.; MEHL, M. R.; PENNEBAKER, J. W. Linguistic Markers of Psychological Change Surrounding September 11, 2001. **Psychological Science**, [S.l.], v.15, n.10, p.687–693, 2004. PMID: 15447640.

CORCHO, O.; FERNÁNDEZ-LÓPEZ, M.; GÓMEZ-PÉREZ, A. Methodologies, tools and languages for building ontologies. Where is their meeting point? **Data & knowledge engineering**, [S.l.], v.46, n.1, p.41–64, 2003.

CORRÊA, U. B. **Análise de sentimento baseada em aspectos usando aprendizado profundo**: uma proposta aplicada à língua portuguesa. 2021. 123p. Doutorado em Computação — Universidade Federal de Pelotas, Pelotas.

COSTA ARRAIS, L. da; GALUCIO, A. V. M. Predicados Nominais e Adjetivais em Línguas do Ramo Tupari da Família Tupí. **Revista Brasileira de Línguas Indígenas**, [S.l.], v.3, n.2, p.156–182, 2020.

COSTA, L.; SANTOS, D.; ROCHA, P. A. Estudando o português tal como é usado: o serviço AC/DC. In: IN THE 7 TH BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY (STIL 2009)(SÃO CARLOS 8-11 DE SETEMBRO 2009), 2009. **Anais...** [S.l.: s.n.], 2009.

COSTE, R. G. D.; GALLISON, R. Dicionário de didáctica das línguas. **Coimbra, Editora Almedina**, [S.l.], v.22, 1983.

CRUZ, I.; GELBUKH, A. F.; SIDOROV, G. Implicit Aspect Indicator Extraction for Aspect based Opinion Mining. **Int. J. Comput. Linguistics Appl.**, [S.l.], v.5, n.2, p.135–152, 2014.

DE MARNEFFE, M.-C.; NIVRE, J. Dependency grammar. **Annual Review of Linguistics**, [S.l.], v.5, p.197–218, 2019.

DE SMEDT, T.; DAELEMANS, W. Pattern for python. **The Journal of Machine Learning Research**, [S.l.], v.13, n.1, p.2063–2067, 2012.

DIJKSTRA, E. W. A note on two problems in connexion with graphs. **Numerische Mathematik**, [S.l.], v.1, n.1, p.269–271, 1959.

DONG, H.; YU, S.; JIANG, Y. A Protégé Plug-In for Modeling Dimensional Ontology. In: NINTH INTERNATIONAL CONFERENCE ON HYBRID INTELLIGENT SYSTEMS, 2009., 2009. **Anais...** [S.l.: s.n.], 2009. v.3, p.125–128.

EKMAN, P. An argument for basic emotions. **Cognition & emotion**, [S.l.], v.6, n.3-4, p.169–200, 1992.

ELEUTÉRIO, S.; RANCHHOD, E.; MOTA, C.; CARVALHO, P. Dicionários Electrónicos do Português. Características e Aplicações. In: VIII SIMPOSIO INTERNACIONAL DE COMUNICACIÓN SOCIAL, 2003. **Actas...** [S.l.: s.n.], 2003. p.636–642.

ELLIOTT, C. D. **The affective reasoner**: a process model of emotions in a multiagent system. [S.l.]: Northwestern University, 1992.

EMYGDIO, J. L.; ALMEIDA, M. B.; TEIXEIRA, L. M. D. ENSAIO SOBRE ONTOLOGIA APLICADA NA RECUPERAÇÃO DA INFORMAÇÃO PARA A CIÊNCIA DA INFORMAÇÃO. **PontodeAcesso**, [S.l.], v.15, n.3, 2021.

ESULI, A.; SEBASTIANI, F. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In: FIFTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC'06), 2006, Genoa, Italy. **Proceedings...** European Language Resources Association (ELRA), 2006.

FELLBAUM, C. **WordNet**: An electronic lexical database. [S.l.]: MIT press, 1998.

FINATTO, M. J. B.; CASELI, H. M.; LOPES, L.; RASSI, A. Sequência de caracteres e palavras: Morfologia e morfossintaxe. In: CASELI, H. M.; NUNES, M. G. V. (Ed.). **Processamento de Linguagem Natural**: Conceitos, Técnicas e Aplicações em Português. 2.ed. São Carlos: BPLN, 2024.

FRANCIS, W. N.; KUCERA, H. **Frequency analysis of English usage**: Lexicon and usage. [S.l.]: Houghton Mifflin, 1982.

FREITAS, C. Sobre a construção de um léxico da afetividade para o processamento computacional do português. **Revista Brasileira de Linguística Aplicada**, [S.l.], v.13, p.1031–1059, 2013.

FREITAS, C.; MOTTA, E.; MILIDIÚ, R.; CÉSAR, J. Vampiro que brilha... rá! desafios na anotaçao de opiniao em um corpus de resenhas de livros. **Encontro de Linguística de Corpus**, [S.l.], v.11, p.22, 2012.

FREITAS, C.; SANTOS, D. Blogs, Amazônia e a Floresta Sintá (c) tica: um corpus de um novo gênero? **quot; In Ana Maria T Ibaños; Livia Pretto Mottin; Tony Berber Sardinha; Simone Sarmento (ed) Pesquisas e Perspectivas em Lingüística de Corpus 2015; 2015 Campinas: Mercado de Letras**, [S.l.], 2015.

FREITAS, C.; SANTOS, D.; GONÇALVES, A. Perguntas já respondidas sobre o AC/DC: desde como começar até uso complexo de funcionalidades poderosas. , [S.l.], 2011.

FREITAS, L. A. d. **Feature-level sentiment analysis applied to brazilian portuguese reviews**. 2015. 94p. Doutorado em Ciência da Computação — Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre.

FRIDRICH, J. **Steganography in digital media**: principles, algorithms, and applications. [S.l.]: Cambridge University Press, 2009.

GAMALLO, P. et al. Análisis morfosintáctico y clasificación de entidades nombradas en un entorno Big Data. **Procesamiento del Lenguaje Natural**, [S.l.], n.53, September 2014.

GARCIA, M.; GAMALLO, P. Yet another suite of multilingual NLP tools. In: IV SYMPOSIUM ON LANGUAGES, APPLICATIONS AND TECNOLOGIES (SLATE'15), 2015, Madri, Espanha. **Anais...** Universidad Complutense, 2015.

GAŠEVIC, D.; DJURIC, D.; DEVEDŽIC, V. **Model driven engineering and ontology development**. [S.l.]: Springer Science & Business Media, 2009.

GENNARI, J. H. et al. The evolution of Protégé: an environment for knowledge-based systems development. **International Journal of Human-computer studies**, [S.l.], v.58, n.1, p.89–123, 2003.

GEORGE, C. E.; SCERRI, J. Web 2.0 and User-Generated Content: legal challenges in the new frontier. **Journal of Information, Law and Technology**, [S.l.], v.2, 2007.

GLIOZZO, A.; STRAPPARAVA, C.; DAGAN, I. Unsupervised and supervised exploitation of semantic domains in lexical disambiguation. **Computer Speech & Language**, [S.l.], v.18, n.3, p.275–299, 2004.

GOLDFARB-TARRANT, S.; ROSS, B.; LOPEZ, A. Cross-lingual Transfer Can Worsen Bias in Sentiment Analysis. **arXiv preprint arXiv:2305.12709**, [S.l.], 2023.

GROOTENDORST, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. **arXiv preprint arXiv:2203.05794**, [S.l.], 2022.

GROUP, W. O. W. et al. OWL 2 web ontology language document overview. **http://www. w3. org/TR/owl2-overview/**, [S.l.], 2009.

GUPTA, V. et al. Toward Integrated CNN-Based Sentiment Analysis of Tweets for Scarce-Resource Language—Hindi. **ACM Trans. Asian Low-Resour. Lang. Inf. Process.**, New York, NY, USA, v.20, n.5, jun 2021.

HAPKE, H.; HOWARD, C.; LANE, H. **Natural Language Processing in Action**: Understanding, analyzing, and generating text with Python. [S.l.]: Simon and Schuster, 2019.

HARMAN, D. et al. Information retrieval: the early years. **Foundations and Trends® in Information Retrieval**, [S.l.], v.13, n.5, p.425–577, 2019.

HITZLER, P. et al. OWL 2 web ontology language primer. **W3C recommendation**, [S.l.], v.27, n.1, p.123, 2009.

HONNIBAL, M.; MONTANI, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. **To appear**, [S.l.], v.7, n.1, p.411–420, 2017.

HONNIBAL, M.; MONTANI, I.; VAN LANDEGHEM, S.; BOYD, A. spaCy: Industrial-strength Natural Language Processing in Python. , [S.l.], 2020.

HORNBY, A.; WEHMEIER, S.; ASHBY, M. Oxford Advanced Learner's Dictionary: Oxford University Press. **NewYork City, US A**, [S.l.], 2000.

HU, M.; LIU, B. Mining and summarizing customer reviews. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2004. **Proceedings...** [S.l.: s.n.], 2004. p.168–177.

HUANG, C.-L. et al. The development of the Chinese linguistic inquiry and word count dictionary. **Chinese Journal of Psychology**, [S.l.], 2012.

HUANG, C.-R. et al. (Ed.). **Ontology and the Lexicon, A Natural Language Processing Perspective**. [S.l.]: Cambridge University Press, 2010. (Studies in Natural Language Processing).

HUTTO, C.; GILBERT, E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. **Proceedings of the International AAAI Conference on Web and Social Media**, [S.l.], v.8, n.1, p.216–225, May 2014.

IDE, N.; ROMARY, L. International standard for a linguistic annotation framework. **Natural Language Engineering**, [S.l.], v.10, n.3-4, p.211–225, 2004.

JEFFRIES, L. **Discovering Language**: The Structure of Modern English. [S.l.]: Bloomsbury Publishing, 2006. (Perspectives on the English Language).

JOSHI, A.; BALAMURALI, A.; BHATTACHARYYA, P. et al. A fall-back strategy for sentiment analysis in hindi: a case study. **Proceedings of the 8th ICON**, [S.l.], 2010.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3.ed. [S.l.: s.n.], 2023. Draft of January 7, 2023.

KACEWICZ, E. et al. Pronoun Use Reflects Standings in Social Hierarchies. **Journal of Language and Social Psychology**, [S.l.], v.33, n.2, p.125–143, 2014.

KAMPS, J. et al. Using WordNet to measure semantic orientations of adjectives. In: LREC, 2004. **Anais...** [S.l.: s.n.], 2004. v.4, p.1115–1118.

KENNEDY, C. Comparison and polar opposition. In: SEMANTICS AND LINGUISTIC THEORY, 1997. **Anais...** [S.l.: s.n.], 1997. v.7, p.240–257.

KENNEDY, C. Polar opposition and the ontology of 'degrees'. **Linguistics and philosophy**, [S.l.], v.24, p.33–70, 2001.

KENNEDY, C. On the monotonicity of polar adjectives. **Perspectives on negation and polarity items**, [S.l.], p.201–221, 2001.

KENNEDY, C. **Projecting the adjective**: The syntax and semantics of gradability and comparison. [S.l.]: Routledge, 2013.

KIPFER, B. A. **Roget's International Thesaurus**. 8th.ed. New York: Collins Reference, 2019.

KOCH, I. G. **O texto e a construção dos sentidos**. [S.l.]: Editora Contexto, 2000.

KRAUWER, S. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In: SPECOM, 2003. **Proceedings...** [S.l.: s.n.], 2003. v.2003, n.8, p.15.

KRISTINA TOUTANOVA DAN KLEIN, C. D. M.; SINGER, Y. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: HLT-NAACL, 2003. **Anais...** [S.l.: s.n.], 2003.

KUMAR, A.; GUPTA, D. Sentiment Analysis as a Restricted NLP Problem. In: PINARBASI, F.; TASKIRAN, M. N. (Ed.). **Natural Language Processing for Global and Local Business**. Hershey, PA: IGI Global, 2021. p.65–96. (Advances in Business Information Systems and Analytics (ABISA)).

LIU, B. **Sentiment Analysis and Opinion Mining**. San Rafael, California: Morgan Claypool Publishers, 2012. 1–167p. n.1. (Synthesis Lectures on Human Language Technologies, v.5).

LIU, B. **Sentiment analysis**: Mining opinions, sentiments, and emotions. 2.ed. Cambridge, United Kingdom: Cambridge university press, 2020. 1–448p. (Studies in Natural Language Processing).

MAGNINI, B.; CAVAGLIA, G. Integrating Subject Field Codes into WordNet. In: LREC, 2000. **Anais...** [S.l.: s.n.], 2000. v.1413.

MAGNINI, B.; STRAPPARAVA, C. User modelling for news web sites with word sense based techniques. **User Modeling and User-Adapted Interaction**, [S.l.], v.14, p.239–257, 2004.

MAGNINI, B.; STRAPPARAVA, C.; PEZZULO, G.; GLIOZZO, A. The role of domain information in word sense disambiguation. **Natural Language Engineering**, [S.l.], v.8, n.4, p.359–373, 2002.

MANNING, C. D. et al. The Stanford CoreNLP Natural Language Processing Toolkit. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL) SYSTEM DEMONSTRATIONS, 2014. **Anais...** [S.l.: s.n.], 2014. p.55–60.

MANNING, C.; SCHUTZE, H. **Foundations of statistical natural language processing**. [S.l.]: MIT press, 1999.

MÄNTYLÄ, M. et al. Mining valence, arousal, and dominance: possibilities for detecting burnout and productivity? In: OF THE 13TH INTERNATIONAL CONFERENCE ON MINING SOFTWARE REPOSITORIES, 2016. **Proceedings...** [S.l.: s.n.], 2016. p.247–258.

MAR, A. A.; SHIRAI, K.; KERTKEIDKACHORN, N. Weakly Supervised Learning Approach for Implicit Aspect Extraction. **Information**, [S.l.], v.14, n.11, 2023.

MARKOWITZ, D. M. The Meaning Extraction Method: An Approach to Evaluate Content Patterns From Large-Scale Language Data. **Frontiers in Communication**, [S.l.], v.6, 2021.

MARNEFFE, M.-C. de; MANNING, C. D.; NIVRE, J.; ZEMAN, D. Universal Dependencies. **Computational Linguistics**, Cambridge, MA, v.47, n.2, p.255–308, 07 2021.

MASINI, F. Multi-word expressions and morphology. In: **Oxford Research Encyclopedia of Linguistics**. [S.l.: s.n.], 2019.

MAZIERO, E. G.; PARDO, T. A.; DI FELIPPO, A.; SILVA, B. C. Dias-da. A base de dados lexical ea interface web do tep 2.0: thesaurus eletrônico para o português do brasil. In: COMPANION PROCEEDINGS OF THE XIV BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, 2008. **Anais...** [S.l.: s.n.], 2008. p.390–392.

MCGUINNESS, D. L.; VAN HARMELEN, F. et al. OWL web ontology language overview. **W3C recommendation**, [S.l.], v.10, n.10, p.2004, 2004.

MCKINNEY, W. **Python for Data Analysis**. 3.ed. Sebastopol, CA: O'Reilly Media, 2022.

MEHMOOD, K.; ESSAM, D.; SHAFI, K.; MALIK, M. K. Sentiment analysis for a resource poor language—Roman Urdu. **ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)**, [S.l.], v.19, n.1, p.1–15, 2019.

MEL'CUK, I. A. et al. **Dependency syntax**: theory and practice. [S.l.]: SUNY press, 1988.

MEYER, J. G. et al. ChatGPT and large language models in academia: opportunities and challenges. **BioData Mining**, [S.l.], v.16, n.1, p.20, 2023.

MIHALCEA, R.; BANEA, C.; WIEBE, J. Learning multilingual subjective language via cross-lingual projections. In: OF THE 45TH ANNUAL MEETING OF THE ASSOCIATION OF COMPUTATIONAL LINGUISTICS, 2007. **Proceedings...** [S.l.: s.n.], 2007. p.976–983.

MILGRAM, S. The small world problem. **Psychology today**, [S.l.], v.2, n.1, p.60–67, 1967.

MOHAMMAD, S. M.; TURNEY, P. D. Crowdsourcing a word–emotion association lexicon. **Computational intelligence**, [S.l.], v.29, n.3, p.436–465, 2013.

MOHAMMAD, S.; TURNEY, P. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In: NAACL HLT 2010 WORKSHOP ON COMPUTATIONAL APPROACHES TO ANALYSIS AND GENERATION OF EMOTION IN TEXT, 2010. **Proceedings...** [S.l.: s.n.], 2010. p.26–34.

MUSEN, M. The Protégé Project: A Look Back and a Look Forward. **AI Matters**, [S.l.], v.1, n.4, Jun 2015.

NABER, D. Openthesaurus: Building a thesaurus with a web community. **Retrieved January**, [S.l.], v.3, p.2005, 2004.

NET, E. Classification of hotel establishments within the EU. **The European Consumer Centres' Network**, [S.l.], 2009.

NEWHAM, C. **Learning the bash shell**: Unix shell programming. [S.l.]: "O'Reilly Media, Inc.", 2005.

NEWMAN, M. L.; PENNEBAKER, J. W.; BERRY, D. S.; RICHARDS, J. M. Lying Words: Predicting Deception from Linguistic Styles. **Personality and Social Psychology Bulletin**, [S.l.], v.29, n.5, p.665–675, 2003. PMID: 15272998.

NOY, N. F. et al. Creating semantic web contents with protege-2000. **IEEE intelligent systems**, [S.l.], v.16, n.2, p.60–71, 2001.

NUNES, M. d. G. V. et al. **Desafios na construção de recursos lingüístico-computacionais para o processamento automático do português do Brasil**. [S.l.]: Mercado de Letras, 2005.

OLIVEIRA, H. G. A Survey on Portuguese Lexical Knowledge Bases: Contents, Comparison and Combination. **Information**, [S.l.], v.9, n.2, 2018.

OLIVEIRA, H. G.; GOMES, P. Onto.PT: Construção Automática de uma Ontologia Lexical para o Português. In: LUíS, A. R. (Ed.). **Estudos de Linguística**. Coimbra: Imprensa da Universidade de Coimbra, 2011. v.1, p.161–180.

OLIVEIRA, H. G.; GOMES, P. ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. **Language Resources and Evaluation Journal**, [S.l.], v.48, n.2, p.373–393, 2014.

OLIVEIRA, H. G.; SANTOS, A. P.; GOMES, P. Assigning Polarity Automatically to the Synsets of a Wordnet-like Resource. In: SYMPOSIUM ON LANGUAGES, APPLICATIONS AND TECHNOLOGIES, 3., 2014, Dagstuhl, Germany. **Anais...** Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2014. p.169–184. (Open Access Series in Informatics (OASIcs), v.38).

OLIVEIRA, H. G.; SANTOS, D.; GOMES, P. Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e a sua avaliação. **Linguamática**, [S.l.], v.2, n.1, p.77–93, 2010.

OLTRAMARI, A.; VOSSEN, P.; QIN, L.; HOVY, E. (Ed.). **New Trends of Research in Ontologies and Lexical Resources**: Ideas, Projects, Systems. Berlin: Springer, 2013. (Theory and Applications of Natural Language Processing).

ORTONY, A.; CLORE, G. L.; COLLINS, A. **The Cognitive Structure of Emotions**. 2.ed. Cambridge: Cambridge university press, 2022.

OSGOOD, C. E.; SUCI, G. J.; TANNENBAUM, P. H. **The measurement of meaning**. [S.l.]: University of Illinois press, 1957. n.47.

OTHERO, G. d. Á. Lingüística Computacional: uma breve introdução. **Letras de hoje**, [S.l.], v.41, n.2, 2006.

OVERALL, S. E.; VALLEJOS, R.; GILDEA, S. Nonverbal predication in Amazonia. **Nonverbal predication in Amazonian languages. Amsterdam/Philadelphia: John Benjamins**, [S.l.], v.122, p.1–49, 2018.

PADRó, L. Analizadores Multilingües en FreeLing. **Linguamatica**, [S.l.], v.3, n.2, p.13–20, December 2011.

PADRó, L. et al. FreeLing 2.1: Five Years of Open-Source Language Processing Tools. In: LANGUAGE RESOURCES AND EVALUATION CONFERENCE (LREC'10), 7., 2010, La Valletta, Malta. **Proceedings...** [S.l.: s.n.], 2010.

PADRó, L.; STANILOVSKY, E. FreeLing 3.0: Towards Wider Multilinguality. In: LANGUAGE RESOURCES AND EVALUATION CONFERENCE (LREC 2012), 2012, Istanbul, Turkey. **Proceedings...** [S.l.: s.n.], 2012.

PAGANI, L. A. Duas Noções Fundamentais para Gramáticas de Dependência. , [S.l.], 2015.

PAIVA, V. d.; RADEMAKER, A.; MELO, G. d. OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning. In: COLING 2012: DEMONSTRATION PAPERS, 2012, Mumbai, India. **Proceedings...** The COLING 2012 Organizing Committee, 2012. p.353–360. Published also as Techreport http://hdl.handle.net/10438/10274.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up? Sentiment Classification using Machine Learning Techniques. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP 2002), 2002., 2002. **Proceedings...** Association for Computational Linguistics, 2002. p.79–86.

PANTEL, P. Inducing ontological co-occurrence vectors. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL'05), 43., 2005. **Proceedings...** [S.l.: s.n.], 2005. p.125–132.

PASQUALOTTI, P. R. **Reconhecimento de expressões de emoções na interação mediada por computador**. 2008. Computação Aplicada — Universidade do Vale do Rio do Sinos.

PASQUALOTTI, P. R. **WordNet Affect BR - uma base de expressões de emoção em Português**: Reconhecimento de expressões de emoções na interação mediada por computador. [S.l.]: Novas Edições Academicas, 2015.

PASQUALOTTI, P. R.; VIEIRA, R. WordnetAffectBR: uma base lexical de palavras de emoções para a língua portuguesa. **RENOTE**, [S.l.], v.6, n.1, 2008.

PATEL, A. A.; ARASANIPALAI, A. U. **Applied Natural Language Processing in the Enterprise**. [S.l.]: "O'Reilly Media, Inc.", 2021.

PENNEBAKER, J. W.; BOOTH, R. J.; FRANCIS, M. E. LIWC2007: Linguistic inquiry and word count. **Austin, Texas: liwc. net**, [S.l.], 2007.

PENNEBAKER, J. W.; BOYD, R. L.; JORDAN, K.; BLACKBURN, K. **The development and psychometric properties of LIWC2015**. [S.l.: s.n.], 2015.

PENNEBAKER, J. W.; FRANCIS, M. E.; BOOTH, R. J. Linguistic inquiry and word count: LIWC 2001. **Mahway: Lawrence Erlbaum Associates**, [S.l.], v.71, 2001.

PENNEBAKER, J. W. et al. **The development and psychometric properties of LIWC2007**. [S.l.: s.n.], 2007.

PENNEBAKER, J. W. et al. When Small Words Foretell Academic Success: The Case of College Admissions Essays. **PLOS ONE**, [S.l.], v.9, n.12, p.1–10, 12 2015.

PIOLAT, A. et al. La version française du dictionnaire pour le LIWC: modalités de construction et exemples d'utilisation. **Psychologie française**, [S.l.], v.56, n.3, p.145–159, 2011.

PLUTCHIK, R. A general psychoevolutionary theory of emotion. In: **Theories of emotion**. [S.l.]: Elsevier, 1980. p.3–33.

PLUTCHIK, R. **The emotions**. [S.l.]: University Press of America, 1991.

PLUTCHIK, R. **The psychology and biology of emotion**. [S.l.]: HarperCollins College Publishers, 1994.

PORTER, M. F. An algorithm for suffix stripping. **Program**, [S.l.], v.14, n.3, p.130–137, 1980.

Protégé. Accessado em: [06 jul. 2023], https://protege.stanford.edu/about.php. Disponível em: <https://protege.stanford.edu/about.php>.

Protégé. Accessado em: [06 jul. 2023], https://protege.stanford.edu/software.php. Disponível em: <https://protege.stanford.edu/software.php>.

RAMÍREZ-ESPARZA, N.; PENNEBAKER, J. W.; GARCÍA, F. A.; SURIÁ, R. La psicología del uso de las palabras: Un programa de computadora que analiza textos en español. **Revista mexicana de psicología**, [S.l.], v.24, n.1, p.85–99, 2007.

RANCHHOD, E.; MOTA, C.; BAPTISTA, J. A Computational Lexicon of Portuguese for Automatic Text Parsing. In: SIGLEX99: STANDARDIZING LEXICAL RESOURCES, 1999. **Anais...** [S.l.: s.n.], 1999.

RANI, S.; JAIN, A. Aspect-based sentiment analysis of drug reviews using multi-task learning based dual BiLSTM model. **Multimedia Tools and Applications**, [S.l.], p.1–29, 2023.

RANI, S.; KUMAR, P. Aspect-Based Sentiment Analysis Using Dependency Parsing. **ACM Trans. Asian Low-Resour. Lang. Inf. Process.**, New York, NY, USA, v.21, n.3, dec 2021.

REIS, J. et al. Uma abordagem multilíngue para análise de sentimentos. In: IV BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 2015. **Anais...** [S.l.: s.n.], 2015.

RIBEIRO, F. N. et al. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. **EPJ Data Science**, [S.l.], v.5, n.1, p.1–29, 2016.

RIBEIRO, P. L. V. Uma abordagem unificada para análise de sentimento de tweets com domínio específico. , [S.l.], 2015.

RODRIGUES, R. G.; GOMES, R. R.; RODRIGUES, K. T.; GUEDES, G. P. TATMaster: Psycholinguistic divergences in automatically translated texts. In: BRAZILLIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, 23., 2017. **Proceedings...** [S.l.: s.n.], 2017. p.205–208.

RUSSELL, S. J.; NORVIG, P. **Inteligência artificial**. [S.l.]: Elsevier, 2004.

SANTOS, D. A sintaxe do AC/DC: apresentaçao do CWB e das opçoes tomadas. , [S.l.], 2012.

SANTOS, D.; BICK, E. Providing Internet access to Portuguese corpora: the AC/DC project. In: IN MARIA GAVRILIDOU; GEORGE CARAYANNIS; STELLA MARKANTONATOU; STELIOS PIPERIDIS; GREGORY STAINHAUER (ED) PROCEEDINGS OF THE SECOND INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2000)(ATHENS 31 MAY-2 JUNE 2000), 2000. **Anais...** [S.l.: s.n.], 2000.

SCHMID, H. Improvements in part-of-speech tagging with an application to German. In: **Natural language processing using very large corpora**. [S.l.]: Springer, 1999. p.13–25.

SCHMID, H. Probabilistic part-of-speech tagging using decision trees. In: NEW METHODS IN LANGUAGE PROCESSING, 2013. **Anais...** [S.l.: s.n.], 2013. p.154–164.

SCOTT, M. **Wordsmith Tools**. Oxford: Oxford University Press, 1996.

SCOTT, M. Developing wordsmith. **International Journal of English Studies**, [S.l.], v.8, n.1, p.95–106, 2008.

SILVA, F. L. V. da et al. ABSAPT 2022 at IberLEF: Overview of the Task on Aspect-Based Sentiment Analysis in Portuguese. **Procesamiento del Lenguaje Natural**, [S.l.], v.69, n.0, p.199–205, 2022.

SILVA, M. J.; CARVALHO, P.; COSTA, C.; SARMENTO, L. Automatic expansion of a social judgment lexicon for sentiment analysis. , [S.l.], 2010.

SILVA, M. J.; CARVALHO, P.; SARMENTO, L. Building a sentiment lexicon for social judgement mining. In: COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE: 10TH INTERNATIONAL CONFERENCE, PROPOR 2012, COIMBRA, PORTUGAL, APRIL 17-20, 2012. PROCEEDINGS 10, 2012. **Anais...** [S.l.: s.n.], 2012. p.218–228.

SIMÕES, A.; IRIARTE SANROMÁN, Á.; ALMEIDA, J. J. Dicionário-Aberto: Construção semiautomática de uma funcionalidade codificadora. , [S.l.], 2013.

SIMÕES, A.; SANROMÁN, Á. I.; ALMEIDA, J. J. Dicionário-Aberto: A Source of Resources for the Portuguese Language Processing. In: COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, 2012, Berlin, Heidelberg. **Anais...** Springer Berlin Heidelberg, 2012. p.121–127.

SINGH, O. M.; TIMILSINA, S.; BAL, B. K.; JOSHI, A. Aspect based abusive sentiment detection in Nepali social media texts. In: IEEE/ACM INTERNATIONAL CON-

FERENCE ON ADVANCES IN SOCIAL NETWORKS ANALYSIS AND MINING (ASONAM), 2020., 2020. **Anais...** [S.l.: s.n.], 2020. p.301–308.

SMITH, T. W. **Schadenfreude**: The joy of another's misfortune. [S.l.]: Little, Brown Spark, 2018.

SOUZA, A. T. d. **Text chunking**: um método de shallow parsing para identificação de sintagmas nominais lexicais de textos em português do Brasil segundo o formalismo Universal Dependencies. 2023. Tese (Doutorado em Ciência da Computação) — Universidade de São Paulo.

SOUZA, K. F. de; PEREIRA, M. H. R.; DALIP, D. H. Unilex: Método léxico para análise de sentimentos textuais sobre conteúdo de tweets em português brasileiro. **Abakós**, [S.l.], v.5, n.2, p.79–96, 2017.

SOUZA, M. et al. Construction of a Portuguese Opinion Lexicon from multiple resources. In: IN 8TH BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY - STIL, MATO GROSSO, 2011. **Anais...** [S.l.: s.n.], 2011.

SOUZA, M.; VIEIRA, R. Sentiment Analysis on Twitter Data for Portuguese Language. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, 10., 2012, Berlin, Heidelberg. **Proceedings...** Springer-Verlag, 2012. p.241–247. (PROPOR'12).

SRINIVASA-DESIKAN, B. **Natural Language Processing and Computational Linguistics**: A practical guide to text analysis with Python, Gensim, spaCy, and Keras. [S.l.]: Packt Publishing Ltd, 2018.

STAAB, S.; STUDER, R. (Ed.). **Handbook on Ontologies**. 2.ed. Berlin, Heidelberg: Springer, 2009. (International Handbooks on Information Systems).

STONE, P. J.; DUNPHY, D. C.; SMITH, M. S.; OGILVIE, D. M. The general inquirer: A computer approach to content analysis. , [S.l.], 1966.

STRAPPARAVA, C.; VALITUTTI, A. et al. Wordnet affect: an affective extension of wordnet. In: LREC, 2004. **Anais...** [S.l.: s.n.], 2004. v.4, n.1083-1086, p.40.

STRAUSS, A.; CORBIN, J. Basics of qualitative research techniques. , [S.l.], 1998.

SUROWIECKI, J. **The wisdom of crowds**. [S.l.]: Anchor, 2005.

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

TAUSCZIK, Y. R.; PENNEBAKER, J. W. The psychological meaning of words: LIWC and computerized text analysis methods. **Journal of language and social psychology**, [S.l.], v.29, n.1, p.24–54, 2010.

TEIXEIRA, J.; SARMENTO, L.; OLIVEIRA, E. Comparing verb synonym resources for portuguese. In: COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE: 9TH INTERNATIONAL CONFERENCE, PROPOR 2010, PORTO ALEGRE, RS, BRAZIL, APRIL 27-30, 2010. PROCEEDINGS 9, 2010. **Anais...** [S.l.: s.n.], 2010. p.100–109.

TEMPLE, M. **Dicionário Oxford Escolar**: para estudantes brasileiros de inglês. [S.l.]: Oxford University Press, 2007.

TRAPPEY, A. J. C.; TRAPPEY, C. V.; HSU, F.-C.; HSIAO, D. W. A Fuzzy Ontological Knowledge Document Clustering Methodology. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, [S.l.], v.39, n.3, p.806–814, 2009.

TTU. **Wheel of Emotions**. Accessado em: [17 abr. 2024]. Disponível em: <https://www.depts.ttu.edu/rise/PDFs/wheelofemotions.pdf>.

TUDORACHE, T.; VENDETTI, J.; NOY, N. F. Web-Protege: A Lightweight OWL Ontology Editor for the Web. In: OWLED, 2008. **Anais...** [S.l.: s.n.], 2008. v.432, p.2009.

TURNEY, P. D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. **arXiv preprint cs/0212032**, [S.l.], 2002.

USCHOLD, M.; GRUNINGER, M. Ontologies: principles, methods and applications. **The Knowledge Engineering Review**, [S.l.], v.11, n.2, p.93–136, 1996.

VALITUTTI, A.; STRAPPARAVA, C.; STOCK, O. Developing affective lexical resources. **PsychNology J.**, [S.l.], v.2, n.1, p.61–83, 2004.

VAN DIJK, W. W.; OUWERKERK, J. W. **Schadenfreude**: Understanding pleasure at the misfortune of others. [S.l.]: Cambridge University Press, 2014.

VOGHOEI, S. et al. Decoding the alphabet soup of degrees in the united states post-secondary education system through hybrid method: Database and text mining. **arXiv preprint arXiv:2309.13050**, [S.l.], 2023.

WANG, Y.; HUANG, M.; ZHU, X.; ZHAO, L. Attention-based LSTM for aspect-level sentiment classification. In: OF THE 2016 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2016. **Proceedings...** [S.l.: s.n.], 2016. p.606–615.

WATSON, D.; CLARK, L. A.; TELLEGEN, A. Development and validation of brief measures of positive and negative affect: the PANAS scales. **Journal of personality and social psychology**, [S.l.], v.54, n.6, p.1063, 1988.

WEIß, G. Agent orientation in software engineering. **The knowledge engineering review**, [S.l.], v.16, n.4, p.349–373, 2001.

WOOLDRIDGE, M. Agent-based software engineering. **IEE Proceedings-software**, [S.l.], v.144, n.1, p.26–37, 1997.

WU, H. C.; LUK, R. W. P.; WONG, K. F.; KWOK, K. L. Interpreting TF-IDF term weights as making relevance decisions. **ACM Transactions on Information Systems (TOIS)**, [S.l.], v.26, n.3, p.1–37, 2008.

XU, X.; ZHANG, J.-D.; XIONG, L.; LIU, Z. **iACOS**: Advancing Implicit Sentiment Extraction with Informative and Adaptive Negative Examples.

Apêndices

APÊNDICE A – Ontologias de Domínio

O desenvolvimento de ontologias de domínio como um modelo de representação formal e estruturado do conhecimento exigiu que certos eventos ocorressem.

Nesse capítulo será abordado o que foi necessário para que ontologias de domínio pudessem ser efetivamente representadas. Além disso, este capítulo também apresenta uma aplicação que permite a visualização dessas ontologias.

A.1 Web Semântica

A grande maioria do conteúdo Web é projetado para humanos lerem e, por essa razão, programas de computador não conseguem a manipular de forma significativa. Os computadores conseguem analisar habilmente o layout das páginas ou realizar o processamento da sua estrutura (como a identificação de um cabeçalho ou de um link para outra página), mas geralmente, não possuem uma maneira confiável de processar a semântica (BERNERS-LEE; HENDLER; LASSILA, 2001).

A Web Semântica (SW — do inglês *Semantic Web*) vem com o intuito de trazer uma estrutura ao conteúdo significativo das páginas da Web, permitindo que as máquinas possam compreender documentos e dados semânticos, o que possibilita uma melhor interação entre pessoas, e, entre pessoas e computadores (BERNERS-LEE; HENDLER; LASSILA, 2001).

A SW não é outro tipo de Web, mas uma extensão da atual, sendo as ontologias os recursos centrais dessa arquitetura (BERNERS-LEE; HENDLER; LASSILA, 2001; EMYGDIO; ALMEIDA; TEIXEIRA, 2021). Enquanto que a Rede Mundial de Computadores (WWW — do inglês *World Wide Web*) estabelece um sistema interconectado de páginas da Web, a SW estabelece um sistema interligado formado pelo conteúdo (dados e informações) dessas páginas (ARP; SMITH; SPEAR, 2015).

A.2 Ontology Web Language

A Linguagem Web de Ontologia (OWL — do inglês *Web Ontology Language*), que representa uma família de linguagens usada na SW, foi projetada para suprir a necessidade de uma linguagem na Web capaz de expressar ontologias (ARP; SMITH; SPEAR, 2015; MCGUINNESS; VAN HARMELEN et al., 2004; HITZLER et al., 2009).

A OWL faz parte de um conjunto de recomendações do Consórcio da Rede Mundial de Computadores (W3C — do inglês *World Wide Web Consortium*) relacionadas à SW. A OWL é uma recomendação do W3C desde 2004, tendo a sua segunda versão aprimorado e expandido a especificação original (MCGUINNESS; VAN HARMELEN et al., 2004; GROUP et al., 2009).

A OWL foi desenvolvida pelo Grupo de Trabalho sobre OWL da W3C (W3C OWL *Working Group*) e baseada em duas tecnologias importantes: Linguagem de Marcação Extensiva (XML — do inglês *eXtensible Markup Language*) e Estrutura de Descrição de Recursos (RDF — do inglês *Resource Description Framework*) (MCGUINNESS; VAN HARMELEN et al., 2004; GROUP et al., 2009; BERNERS-LEE; HENDLER; LASSILA, 2001).

O XML é uma linguagem de marcação que permite a representação de dados em formato de texto estruturado, sendo frequentemente usada para armazenar e trocar dados em um formato legível por máquina e humanos. Entretanto o XML não impõe uma estrutura específica aos dados e restrições semânticas ao significado desses documentos, permitindo que os desenvolvedores definam sua própria estrutura usando tags personalizadas (MCGUINNESS; VAN HARMELEN et al., 2004; GROUP et al., 2009; BERNERS-LEE; HENDLER; LASSILA, 2001; ANTONIOU; HARMELEN, 2009).

No entanto, o Esquema XML (do inglês XML *Schema*), também chamado de Definição do Esquema XML (XSD — do inglês XML *Schema Definition*), restringe a estrutura de documentos XML, definindo regras de validação em documentos no formato XML. Além disso, estende a linguagem XML com tipos de dados (MCGUINNESS; VAN HARMELEN et al., 2004; GROUP et al., 2009; BERNERS-LEE; HENDLER; LASSILA, 2001; ALLEMANG; HENDLER, 2011; ANTONIOU; HARMELEN, 2009).

Já o RDF estabelece um modelo de dados para objetos (recursos) e as relações entre eles, fornecendo uma semântica simples para este modelo de dados, que podem ser representados em uma sintaxe XML (MCGUINNESS; VAN HARMELEN et al., 2004; GROUP et al., 2009; BERNERS-LEE; HENDLER; LASSILA, 2001; ALLEMANG; HENDLER, 2011; ANTONIOU; HARMELEN, 2009).

O Esquema RDF (RDFS — do inglês RDF *Schema*) representa uma extensão do RDF, sendo utilizado para estruturar vocabulários RDF. Ele fornece um vocabulário para descrever propriedades e classes de recursos RDF, incluindo uma semântica destinada às hierarquias de generalização dessas propriedades e classes (MCGUINNESS; VAN HARMELEN et al., 2004; GROUP et al., 2009; BERNERS-LEE; HENDLER; LASSILA, 2001; ALLEMANG; HENDLER, 2011; ANTONIOU; HARMELEN, 2009).

A OWL estende ainda mais o vocabulário do RDFS, trazendo novas descrições de propriedades e classes, como: relações entre classes (por exemplo, disjunção), cardinalidade (por exemplo, ”exatamente um”), igualdade, tipagem mais rica de propri-

edades, características de propriedades (por exemplo, simetria) e classes enumeradas (MCGUINNESS; VAN HARMELEN et al., 2004; GROUP et al., 2009; BERNERS-LEE; HENDLER; LASSILA, 2001; ALLEMANG; HENDLER, 2011; ANTONIOU; HARMELEN, 2009).

A OWL é uma linguagem de modelagem semântica mais avançada, pois, oferece maior expressividade e suporte para a definição de classes, propriedades, restrições e axiomas para representar informações de forma mais precisa, permitindo a definição de ontologias mais complexas.

OWL 2 adiciona novas funcionalidades em relação a sua primeira versão (GROUP et al., 2009). Alguns dos novos recursos são açúcar sintático[1] (por exemplo, união disjunta de classes), enquanto outros oferecem nova expressividade, incluindo:

- chaves;
- cadeias de propriedade;
- tipos de dados e intervalos de dados mais ricos;
- restrições de cardinalidade qualificada;
- propriedades assimétricas, reflexivas e disjuntas;
- capacidades de anotação aprimoradas.

A.3 Protégé

Ferramentas podem ajudar a adquirir, organizar e visualizar a KB antes, durante e após a criação e construção de ontologias de domínio. Ambientes gráficos para o desenvolvimento de ontologias incorporam um editor de ontologias, o qual é integrado a outras ferramentas, e, em geral, oferecem suporte a diversas linguagens de representação de ontologias, como RDF e OWL (GAŠEVIC; DJURIC; DEVEDŽIC, 2009). Eles visam fornecer suporte para todos os processos de desenvolvimento da ontologia e para a sua posterior utilização (CORCHO; FERNÁNDEZ-LÓPEZ; GÓMEZ-PÉREZ, 2003).

Protégé é a principal ferramenta de engenharia ontológica, fornecendo um ambiente para desenvolvimento e edição de ontologias (NOY et al., 2001; GAŠEVIC; DJURIC; DEVEDŽIC, 2009). O Protégé foi desenvolvido pelo Centro de Stanford para Pesquisa em Informática Biomédica (BMIR — do inglês *Stanford Center for Biomedical Informatics Research*), na Faculdade de Medicina da Universidade de Stanford (PROTÉGÉ, 2023a).

[1] Em ciência da computação, um açúcar sintático é uma sintaxe dentro da linguagem de programação que tem por finalidade tornar suas construções mais fáceis de serem lidas e expressas.

Inicialmente, o Protégé foi concebido para elaborar ontologias em áreas relacionadas com a medicina, como a medicina clínica e as ciências biomédicas (GENNARI et al., 2003). Contudo, ao longo dos anos, sua utilização expandiu-se para outros domínios, em virtude do rápido crescimento de sua base de usuários, que atualmente ultrapassa 360 mil (GENNARI et al., 2003; PROTÉGÉ, 2023a). E, atualmente, é o software mais popular para edição de ontologias e amplamente utilizado no mundo (GENNARI et al., 2003).

Segundo Protégé (2023b) e Dong; Yu; Jiang (2009), "Protégé é uma plataforma gratuita e de código aberto que fornece a uma comunidade crescente de usuários um conjunto de ferramentas para construir modelos de domínio e aplicações baseadas em conhecimento com ontologias".

Protégé também integra classificadores dedutivos que contribuem para a verificação da consistência dos modelos e a dedução de novas informações a partir da análise de uma ontologia. Além disso, possui uma estrutura que suporta a integração de vários *plugins* de outros projetos, parecido com o software Eclipse (TRAPPEY et al., 2009).

Segundo Protégé (2023b), Protégé está disponível em duas versões:

- **Protégé Desktop:** O Protégé Desktop[2] é um ambiente de edição de ontologia rico em recursos, com suporte total para a linguagem de ontologia para a OWL 2 e conexões diretas na memória para raciocinadores lógicos de descrição, como HermiT e Pellet. Protégé Desktop está disponível para Mac OS, Windows e Linux, e requer que o Ambiente de Execução Java (do inglês, Java Runtime Environment - JRE) esteja instalado no computador.

- **WebProtégé:** WebProtégé[3] é um ambiente de desenvolvimento de ontologias para a Web que facilita a criação, upload, modificação e compartilhamento de ontologias para visualização e edição colaborativas.

Embora o cliente *desktop* (Protégé Desktop) seja um sistema mais robusto e disponibilize uma ampla gama de recursos, o cliente baseado na web (WebProtégé) alcançou uma popularidade notável, tendo recentemente superado o cliente *desktop* em termos de utilização. Isso se deve ao seu design minimalista, que facilita o acesso a diversas tarefas. A capacidade de simplesmente apontar um navegador da web para um servidor apropriado e iniciar a edição revela-se extremamente útil (MUSEN, 2015).

[2]O Protégé Desktop (MUSEN, 2015) pode ser baixado através do site `https://protege.stanford.edu/`.

[3]O WebProtégé (TUDORACHE; VENDETTI; NOY, 2008) pode ser acessado através do navegador, pelo site `https://webprotege.stanford.edu/`.

APÊNDICE B – Léxicos de Sentimento

B.1 AffectPT-br

Carvalho; Santos; Guedes (2018) desenvolveram um dicionário afetivo para o Português Brasileiro, chamado AffectPT-br[1]. Este dicionário foi desenvolvido utilizando o LIWC2015, versão padrão de 2015 do dicionário de Investigação Linguística e Contagem de Palavras (do inglês, *Linguistic Inquiry and Word Count* — LIWC) (CARVALHO; SANTOS; GUEDES, 2018; PENNEBAKER et al., 2015). Usou-se o LIWC2015 como principal referência para informações referentes a estrutura hierárquica de categorias, palavras a serem incluídas e atribuição de categorias.

Os LI do AffectPT-br foram classificados com base na estrutura hierárquica *Affective processes* (Processos Afetivos) dos *Psychological Processes* (Processos Psicológicos) do LIWC2015, o qual foi desenvolvida a partir de conceitos psicológicos e com a colaboração de avaliadores (CARVALHO; SANTOS; GUEDES, 2018; RAMÍREZ-ESPARZA et al., 2007). Essa estrutura é organizada nas seguintes categorias e subcategorias:

- **affect:** representa a categoria dos *Affective processes*, os quais são caracterizados por conceitos que alteram o estado emocional, tais como: feliz e triste.
  - **posemo:** representa a categoria das *Positive Emotions* (Emoções Positivas), caracterizada por conceitos que promovem uma alteração positiva no estado emocional, tais como: fé e incentivo.
  - **negemo:** representa a categoria das *Negative Emotions* (Emoções Negativas), caracterizada por conceitos que promovem uma alteração negativa no estado emocional, tais como: vilipêndio e torpor.
    - **anx:** representa a categoria da *Anxiety* (Ansiedade), caracterizada por conceitos que promovem um estado de ansiedade, tais como: briga e constrangimento.
    - **anger:** representa a categoria da *Anger* (Raiva), caracterizada por conceitos que promovem um estado de raiva, tais como: assassinado e abusada.

[1] O recurso AffectPT-br (CARVALHO; SANTOS; GUEDES, 2018) pode ser baixado através do link: `https://github.com/LaCAfe/AffectPT-br`.

* **sad:** representa a categoria da *Sadness* (Tristeza), caracterizada por conceitos que promovem um estado de tristeza, tais como: depressivo e desilusão.

B.1.1 Processo de desenvolvimento do AffectPT-br

O processo de desenvolvimento do AffectPT-br incluiu as etapas de tradução, comparação e expansão de palavras originalmente presentes nas categorias do dicionário LIWC2015.

Na primeira etapa, as palavras das subcategorias “posemo” e “negemo” do LIWC foram extraídas e enviadas aos serviços de tradução online Google Tradutor (do inglês, *Google Translate* — GT) e Tradutor da Microsoft (do inglês, *Microsoft Translator* — MT). Segundo Carvalho; Santos; Guedes (2018), embora tanto o GT quanto o MT apresente pequenas diferenças psicolinguísticas quando comparados à tradução humana, o procedimento mostrou-se ainda assim satisfatório, conforme demonstrado por Reis et al. (2015) e Rodrigues et al. (2017).

Na segunda etapa, foi realizado a comparação e expansão de palavras. Inicialmente, utilizou-se o *Oxford Advanced Learner's Dictionary* (HORNBY; WEHMEIER; ASHBY, 2000) para aprimorar a compreensão dos possíveis significados de cada palavra em inglês extraída do LIWC2015 e para avaliar se as traduções fornecidas pelos serviços de tradução online GT ou MT eram boas ou adequadas. Para evitar repetições, as palavras em inglês foram consultadas no Dicionário Bilíngue Oxford Inglês-Português Brasileiro (TEMPLE, 2007), facilitando assim a visualização de uma lista de sinônimos em português.

Ao final de todas as etapas, o AffectPT-br resultou em um total de 1.139 LI atribuídos à categoria 'affect', sendo 479 da categoria 'posemo' e 661 da categoria 'negemo'. Dessa lista, uma pequena parte termina com o caractere curinga $*$', possibilitando a realização da Expansão de Curinga (do inglês, *Wildcard Expansion* — WE). A WE consiste em substituir o caractere curinga $*$' nos LI por qualquer número de caracteres (como letras, hifens ou números) (NEWHAM, 2005). Isso permite a expansão do LI para suas formas flexionadas em termos de conjugação, gênero ou grau (CARVALHO; SANTOS; GUEDES, 2018).

B.1.2 Avaliação do desempenho do AffectPT-br

Foram utilizados dois *datasets* abertos para avaliar o desempenho do AffectPT-br, denominados MQD60k e TAS-PT. Os *dataset* estão disponíveis, respectivamente, nos seguintes *links*: `https://github.com/LaCAfe/MQD60k` e `https://github.com/pauloemmilio/dataset`.

O *dataset* MQD60k contém dados coletados da rede social chamada Meu Querido Diário. Nessa rede social, os usuários podem escrever textos relacionados às suas ex-

periências cotidianas, semelhantes a registros em um diário pessoal, os quais podem conter pensamentos, experiências ou sentimentos. Cada registro pode ser associado a uma das seis emoções propostas por Ekman (1992) (ou seja, medo, raiva, nojo, tristeza, felicidade e surpresa). Foram recolhidas um total de 59.166 publicações, das quais 32.244 estão associadas à classe felicidade e 26.922 à classe tristeza. As publicações foram selecionadas aleatoriamente entre usuários com idades compreendidas entre 13 e 99 anos, sem distinção de gênero.

O *dataset* TAS-PT contém dados coletados da rede social chamada X e é utilizado para realizar tarefas de SA em português. O *dataset* é composto por dois arquivos, um contendo IDs de *tweets* de sentimentos positivos e outro contendo IDs de *tweets* de sentimentos negativo. Como os arquivos não continham o conteúdo textual das mensagens, foi necessário a utilização da API do X para recuperar o conteúdo das mensagens através dos IDs nos arquivos. O processo resultou na criação do *dataset* TAS-PT-60k, que abrange 59.260 registros, dos quais 28.853 são negativos e 30.407 são positivos.

O desempenho AffectPT-br foi comparado com o desempenho do LIWC2007pt, versão em Português Brasileiro de 2007 do LIWC (BALAGE FILHO; PARDO; ALUÍSIO, 2013; PENNEBAKER; BOOTH; FRANCIS, 2007). Os resultados indicam que o AffectPT-br supera o LIWC2007pt na tarefa de classificação com todos os algoritmos de classificação adotados por Carvalho; Santos; Guedes (2018).

B.2 EmoLex

Mohammad; Turney (2010, 2013) desenvolveram o EmoLex[2], Léxico de Associação de Palavra-Emoção do NRC (*NRC Word-Emotion Association Lexicon*), que foi financiada pelo Conselho Nacional de Pesquisa do Canadá (do inglês, *National Research Council Canada* — NRC).

B.2.1 Teorias emocionais

O EmoLex foi fundamentado nas teorias emocionais de Ekman (1992) e Plutchik (1980, 1991, 1994). Ekman (1992) defende a existência de seis emoções fundamentais: *joy* (alegria), *sadness* (tristeza), *anger* (raiva), *fear* (medo), *disgust* (nojo) e *surprise* (surpresa). Enquanto que Plutchik (1980, 1991, 1994) expande essa lista, ao incluir na sua teoria, além das emoções identificadas por Ekman (1992), as emoções *trust* (confiança) e *anticipation* (antecipação).

[2]O LR EmoLex pode ser acessado `http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`.

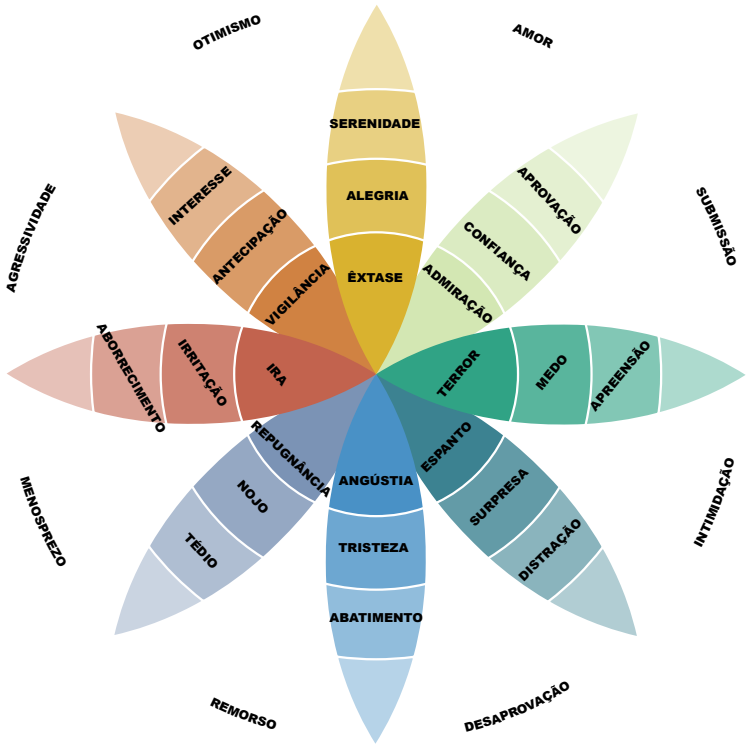

Figura 32 – Roda de emoções de Plutchik. Fonte: Adaptado de *Texas Tech University* (TTU, 2024)

Na Figura 32 pode-se observar a *Plutchik's wheel of emotions* (Roda de emoções de Plutchik), na qual Plutchik organiza as emoções. Nela, emoções semelhantes são colocadas uma ao lado da outra; emoções contrastantes são colocadas diametralmente opostas umas às outras; o raio indica intensidade; os espaços em branco entre as emoções básicas representam díades primárias – emoções complexas que são combinações de emoções básicas adjacentes.

B.2.2 Processo de desenvolvimento do EmoLex

Para gerar a lista de palavras associadas a emoções, iniciou-se a extração, a partir do *Macquarie Thesaurus* (BERNARD, 1986), de um conjunto de palavras e frases que pudessem ser empregadas na criação de um *dataset* de unigramas e bigramas. As categorias presentes no *Macquarie Thesaurus* foram empregadas, de maneira geral, como indicativas dos significados das palavras; assim, caso uma palavra aparecesse em mais de uma categoria, ela seria classificada como possuindo múltiplos sentidos.

O *Macquarie Thesaurus* possui uma lista com mais de 57.000 tipos de palavras em inglês comumente usadas, além de uma lista com mais de 40.000 frases comumente usadas. Desta lista, foram escolhidos os termos que ocorreram com mais frequência no corpus de n-gramas do Google[3] (BRANTS; FRANZ, 2006). Foram escolhidos os os 200 unigramas e os 200 bigramas mais frequentes de quatro PoS (substantivos,

[3]O corpus de n-gramas do Google pode ser acessado através do link `https://catalog.ldc.upenn.edu/LDC2006T13`.

verbos, advérbios e adjetivos). Termos que apareceram em mais de uma categoria do *Macquarie Thesaurus* foram excluídos da lista de termos selecionados. Dentre os 200 bigramas de advérbios analisados, 13 foram desconsiderados por não cumprirem o critério estabelecido, as demais categorias continuaram com 200 termos cada. Isso resultou em uma lista de 1587 termos.

Além do *Macquarie Thesaurus*, também foram explorados os LR WordNet Affect (VALITUTTI; STRAPPARAVA; STOCK, 2004; STRAPPARAVA; VALITUTTI et al., 2004) e General Inquirer[4] (STONE et al., 1966). O WordNet Affect possui centenas de palavras anotadas com várias categorias de emoções e sentimentos. Foram escolhidas somente as palavras que pertenciam as categorias descritas por Ekman (1992) e que possuíam no máximo dois sentidos, resultando em uma lista de 640 termos. Enquanto que o General Inquirer possui 11.788 palavras rotuladas em 182 categorias. Foram escolhidas somente as palavras que possuíam no máximo três sentidos, resultando em uma lista de 8.132 termos. A união dos termos selecionados resultou em uma lista com 10.359 termos.

Para realizar a anotação das emoções foi utilizado o *Amazon Mechanical Turk*, o serviço de *crowdsourcing* da Amazon. Em sistemas de *crowdsourcing*, os usuários solicitantes, como empresas e pesquisadores, enviam tarefas ao sistema para que outros usuários possam realizar as tarefas solicitadas. O solicitante pode dividir a tarefa em sub-tarefas que podem ser respondidas de forma independente, chamadas de Tarefas de Inteligência Humana (do inglês, *Human Intelligence Tasks* — HITs).

O solicitante especifica: palavras-chave relevantes para a tarefa; a remuneração que será paga pela resolução de cada HIT e o número de anotadores diferentes que devem resolver cada HIT. As pessoas que fornecem as respostas dos HITs são chamadas de *Turkers*. Os *Turkers* podem procurar tarefas inserindo palavras-chave as quais está interessado em responder e estipular uma remuneração mínima por cada HIT respondido.

Devido ao elevado custo e à complexidade envolvidos na solicitação de anotações de palavras com centenas de emoções, Mohammad; Turney (2010, 2013) optaram por requerer a anotação baseando-se exclusivamente nas oito emoções básicas propostas por Plutchik (1980, 1991, 1994).

Mohammad; Turney (2010, 2013) salientaram que existe uma série de benefícios em usar um sistema de *crowdsourcing*, como baixo custo, menor sobrecarga organizacional e tempo de resposta rápido. No entanto, ainda que existam esses benefícios, salientaram que existe desafios que são inerentes desse tipo de sistema, como controlar o controle de qualidade e encontrar o número suficiente de *Turkers* dispostos a responder as tarefas.

O controle de qualidade está correlacionado com *Turkers* que são atraídos pela

[4] O RL General Inquirer pode ser requisitado entrando em contado com rhhu@csail.mit.edu.

compensação de uma tarefa e podem inserir informações aleatórias ou, ainda, inserir deliberadamente informações incorretas. Já a dificuldade de encontrar o número suficiente de *Turkers* está relacionado com a dificuldade de responder uma tarefa ou com a baixa remuneração.

Além disso, enfrentam-se desafios relacionados com a anotação de emoções. Embora se exigisse dos anotadores apenas que fossem falantes nativos ou fluentes em inglês, com a justificativa de que essa competência os habilita eficientemente a identificar as emoções vinculadas às palavras, persiste o problema de palavras com múltiplos sentidos evocarem emoções distintas dependendo de como são empregadas.

Para superar esses desafios, eram apresentados aos anotadores um problema de escolha de palavras, antes de fazer perguntas sobre quais emoções estão associadas a um termo-alvo. O problema de escolha de palavras consiste em escolher o termo que está mais próximo da palavra-alvo entre um conjunto de quatro palavras. Três das quatro opções são distrações, enquanto que o termo restante é sinônimo de um dos sentidos da palavra-alvo.

Ao realizar essa pergunta, transmite-se o sentido da palavra para a qual se requerem as anotações, eliminando a necessidade de prover aos anotadores definições longas. Além disso, caso um anotador desconheça a palavra-alvo e forneça uma resposta ao acaso, existe uma probabilidade de 75% de que ele erre a resposta, o que permite descartar todas as suas respostas subsequentes associadas a esse termo específico.

Foi gerado um HIT separado para cada termo da união dos termos selecionados, contendo uma pergunta de escolha de palavras. Cada HIT tem um conjunto de cinco tarefas independentes (anotações) que devem ser respondidas pela mesma pessoa.

Com o objetivo de determinar o método mais eficaz para formular as questões, conduziram-se dois tipos distintos de testes de anotação. Cada teste envolveu um conjunto diferente de 2.100 termos. No primeiro conjunto de anotações, questionava-se se uma palavra estava vinculada a uma emoção específica, enquanto que no segundo, questionava-se se uma palavra provocava uma emoção determinada.

Descobriu-se que houve uma concordância significativamente maior entre os anotadores em relação às palavras associadas a uma emoção específica, em comparação com aquelas que evocavam uma emoção específica. Em vista disso, optou-se por realizar todas as anotações subsequentes utilizando-se de palavras que estivessem vinculadas a uma emoção específica.

Ao final de todos as etapas, o EmoLex resultou em um total de 10.000 LI atribuídos as oito emoções fundamentais de Plutchik (1980, 1991, 1994) (*joy*, *sadness*, *anger*, *fear*, *disgust*, *surprise*, *trust* e *anticipation*). Além disso, os LI são classificados segundo a polaridade da emoção, *positive* (positiva) e *negative* (negativa). Posteriormente, o EmoLex foi traduzido automaticamente para diversas línguas, como o

português.

B.3 LeIA

Almeida (2018) desenvolveu o LeIA[5] (Léxico para Inferência Adaptada). Segundo Almeida (2018), LeIA é um *fork* (derivação ou ramificação) do léxico e ferramenta para análise de sentimentos VADER[6] (*Valence Aware Dictionary and sEntiment Reasoner*) (HUTTO; GILBERT, 2014) adaptado para textos em português, com suporte para *emojis* e foco na análise de sentimentos de textos expressos em mídias sociais.

B.3.1 Característica do VADER

A abordagem de Hutto; Gilbert (2014) visa aproveitar das vantagens da *Parsimonious Rule-based Model*[7] (modelagem parcimoniosa baseada em regras) para construir um mecanismo computacional de análise de sentimento. O modelo proposto possui as seguintes características: funciona bem em textos de mídia social, mas possui a capacidade de generalização para diversos domínios; não requer dados de treinamento, mas é construído a partir de um léxico de sentimento de padrão ouro generalizável, baseado em valência e com curadoria humana; é rápido o suficiente para ser utilizado on-line com *streaming* de dados e não sofre gravemente com problemas de velocidade e desempenho.

B.3.2 Processo de desenvolvimento do VADER

O processo de desenvolvimento do VADER inclui as seguintes etapas:

- **Etapa 1 - Coleta de LI:** Coleta-se todos os LI presente em SL existentes, bem estabelecidos e validados por humanos, como LIWC (PENNEBAKER; FRANCIS; BOOTH, 2001; PENNEBAKER et al., 2007), Normas Afetivas para Palavras Inglesas (do inglês, *Affective Norms for English Words* — ANEW[8]) (BRADLEY; LANG, 1999) e Inquisitor Geral (do inglês, *General Inquirer* — GI) (STONE et al., 1966). Os LI coletados formam a lista preliminar de LI;

- **Etapa 2 - Enriquecimento dos dados:** A lista preliminar de LI é complementada

[5] O LR LeIA pode ser encontrado pelo link `https://github.com/rafjaa/LeIA`.

[6] O LR VADER pode ser encontrado pelo link `https://github.com/cjhutto/vaderSentiment`.

[7] Segundo Aarts (2007), *parsimony* (parcimônia) pode ser interpretada como simplicidade, mas também refere-se a ser econômico ou eficiente. Parsimonious Model (modelo parcimonioso) refere-se ao princípio descrito na *OCKHAM'S RAZOR* (Navalha de Ockham), também chamado de princípio da economia, que diz que entre explicações igualmente eficazes, a mais simples deve ser escolhida (ARIEW, 1976; BROWN; FLORES, 2010). Dessa forma, um modelo parcimonioso busca alcançar um equilíbrio entre simplicidade e capacidade de explicação, preferindo a solução mais simples que seja suficientemente eficaz.

[8] O RL ANEW pode ser baixado através do *link* `https://pdodds.w3.uvm.edu/teaching/courses/2009-08UVM-300/docs/others/everything/bradley1999a.pdf`.

com termos comumente usados para expressar sentimentos em textos de mídia social, formando a lista de LI enriquecida;

- **Etapa 3 - Avaliação Coletiva de Valência:** Utiliza-se a abordagem de *wisdom-of-the-crowd*[9] (sabedoria da multidão) para adquirir uma estimativa precisa da valência (intensidade do sentimento) de cada termo, independentemente do contexto. Utilizou-se o *Amazon Mechanical Turk* para determinar as estimativas pontuais de valência de sentimento para cada um dos mais de 9.000 LI presente na lista de LI enriquecida. Foram recolhidas avaliações de intensidade para cada LI utilizando dez avaliadores humanos independentes, totalizando mais de 90.000 avaliações. Os LI foram avaliados numa escala de -4 a +4, onde -4 representa um sentimento extremamente negativo e +4 um sentimento extremamente positivo, sendo o 0 considerado neutro.

  Para assegurar a qualidade dos resultados foram implementados quatro processos de controle de qualidade:

  1. cada anotador deveria ser passar por um teste de compreensão de leitura na língua inglesa. Para passar no teste o anotador deveria pontuar individualmente 80% ou mais em um teste padronizado de compreensão de leitura de nível universitário;
  2. cada anotador pré-selecionado teve que completar uma sessão de treinamento e orientação de classificação de sentimento on-line e pontuar 90% ou mais para corresponder à classificação de sentimento média conhecida (pré-validada) de LI que incluíam palavras individuais, *emoticons*, acrônimos, sentenças, postagens e trechos de texto (por exemplo, segmentos de sentenças ou frases);
  3. cada lote de 25 LI continha cinco “LI dourados”, LI os quais a distribuição de classificação de sentimento era conhecida (pré-validada). Se um anotador desviava-se em mais de um desvio padrão da média desta distribuição conhecida em três ou mais dos cinco “LI dourados”, todas as 25 classificações no conjunto deste anotador eram descartadas;
  4. implementou-se um programa de bônus para incentivar e recompensar o trabalho de mais alta qualidade. Por exemplo, solicitou-se aos trabalhadores que selecionassem a pontuação de valência que eles acreditavam que “a maioria das outras pessoas” escolheria para o LI fornecido.

[9] A abordagem de *wisdom-of-the-crowd* consiste no processo de reunir e sintetizar as opiniões de um grupo diversificado de pessoas para responder a uma pergunta (SUROWIECKI, 2005). Este método demonstrou ser tão eficaz quanto (e muitas vezes superior) as estimativas fornecidas por indivíduos isolados, inclusive especialistas. A eficácia desse processo baseia-se na ideia de que, ao agregar múltiplas perspectivas, é possível obter uma visão mais precisa ou uma resposta mais acurada a uma dada questão, devido à diversidade de conhecimentos e experiências dos participantes.

Testes revelaram que redigir as instruções dessa maneira resultava em um desvio padrão muito mais restrito, sem afetar significativamente a média da classificação de sentimento, permitindo que fosse alcançado resultados de maior qualidade (generalizados) de forma mais econômica.

- **Etapa 4 - Construção do Léxico Padrão-Ouro:** Para elaborar o léxico de sentimentos padrão-ouro, validado por humanos, manteve-se todos os LI com valência média diferente de zero e desvio padrão menor ou igual a 2,5. A lista de LI Padrão-Ouro é formada por mais de 7.500 LI.

- **Etapa 5 - Desenvolvimento de Heurísticas Generalizáveis:** Analisou-se uma amostra com cerca de 800 postagens, sendo 400 postagens negativas e 400 postagens positivas. Esta amostra foi selecionada de um conjunto com mais de 10.000 postagens aleatórias retiradas da rede social X. Foram selecionados as textos mais positivos e mais negativos dentro do conjunto de postagens, essa seleção foi realizada com o motor de análise de sentimentos Pattern.en[10]. O Pattern é um módulo de mineração da web para Python, e o módulo Pattern.en é uma ferramenta de NLP que utiliza o WordNet para pontuar o sentimento de acordo com os adjetivos em inglês usados no texto. Em seguida, dois especialistas humanos examinaram individualmente cada uma das 800 postagens e as pontuaram de forma independente de acordo com a intensidade de sentimento em uma escala de -4 a +4. Seguindo uma técnica de codificação indutiva baseada em dados, similar à *Grounded Theory approach* (abordagem da Teoria Fundamentada) proposta por Strauss; Corbin (1998), empregou-se subsequentemente técnicas de análise qualitativa para identificar propriedades e características do texto que influenciam a intensidade do sentimento percebido. A partir dessa análise qualitativa aprofundada foi possível encontrar cinco heurísticas generalizáveis, baseadas em pistas gramaticais e sintáticas, para indicar mudanças na intensidade do sentimento:

    1. **Pontuação:** percebeu-se que especificamente o ponto de exclamação (!), aumenta a magnitude da intensidade sem modificar a orientação semântica. Por exemplo, entre as frases “A comida aqui é boa!!!” e “A comida aqui é boa.”, a frase com ponto de exclamação é a que apresenta maior intensidade.
    2. **Capitalização:** percebeu-se que ao utilizar todas as letras maiúsculas para enfatizar uma palavra relevante ao sentimento na presença de outras palavras não capitalizadas, aumenta a magnitude da intensidade do sentimento

[10]Mais informações sobre o Pattern.en (DE SMEDT; DAELEMANS, 2012) podem ser obtidas em `https://github.com/clips/pattern` e `https://digiasset.org/html/pattern-en.html`.

sem afetar a orientação semântica. Por exemplo, entre as frases “A comida aqui é ÓTIMA!” e “A comida aqui é ótima!”, a frase que utiliza a palavras com letras maiúsculas é a que apresenta maior intensidade.

3. **Intensificadores:** percebeu-se que ao utilizar modificadores de grau (também chamados de intensificadores, palavras de reforço ou advérbios de grau) a intensidade do sentimento era afetada, aumentando ou diminuindo a intensidade. Por exemplo, entre as frases “O atendimento aqui é extremamente bom!”, “O atendimento aqui é bom!” e “O atendimento aqui é marginalmente bom!”, a frase que utiliza o intensificador “extremamente” é a que apresenta maior intensidade, enquanto que a frase que utiliza o intensificador “marginalmente” é apresenta menor intensidade.
4. **Conjunção contrastiva:** percebeu-se que ao utilizar a conjunção contrastiva “mas” sinaliza uma mudança na polaridade do sentimento, sendo o sentimento do texto após a conjunção dominante. Por exemplo, a “A comida aqui é ótima, mas o atendimento é horrível!” apresenta um sentimento misto, com a segunda parte dominando a avaliação geral.
5. **Negação:** percebeu-se que ao examinar o trigrama precedendo um LI carregado de sentimento, capturou-se quase 90% dos casos onde a negação altera a polaridade do texto. Por exemplo, uma frase negada seria “A comida daqui não é realmente tão boa!”

- **Etapa 6 - Avaliação do impacto gramatical:** A partir das combinações das heurísticas generalizáveis encontradas foram geradas novas variações de algumas postagens. Foram selecionadas 30 avaliações e essas geraram um total de 200 avaliações artificiais, que posteriormente foram inseridas em um novo conjunto com 800 postagens. Esse novo conjunto foi avaliado por 30 anotadores independentes que classificassem a intensidade do sentimento de todos as 1000 postagens, a fim de avaliar o impacto dessas características na intensidade do sentimento percebido.

- **Etapa 7 - Validação de Valência em Múltiplos Domínios:** Para realizar a validação das estimativas de valência de sentimentos foram utilizados corpora de diferentes domínios[11], utilizando dados agregados de avaliações humanas para estabelecer um consenso sobre a intensidade e a direção dos sentimentos expressos. Foram utilizados para essa etapa 20 avaliadores humanos e corporas do domínio de textos de mídia social, críticas de filmes, avaliações de produtos e artigos de notícias de opinião.

[11] Os corporas utilizados no desenvolvimento do VADER estão disponíveis para download através do *link* `https://github.com/cjhutto/vaderSentiment/tree/master`.

B.4 LIWC2007pt

O *software* de Investigação Linguística e Contagem de Palavras (LIWC — do inglês *Linguistic Inquiry and Word Count*)[12] é um *software* de análise de texto que calcula o grau do uso de diferentes categorias de palavras em uma ampla variedade de textos (PENNEBAKER; FRANCIS; BOOTH, 2001).

O núcleo do *software* LIWC é um recurso léxico, mais conhecido como dicionário LIWC (PENNEBAKER et al., 2015). Conforme descrito por (TAUSCZIK; PENNEBAKER, 2010), o dicionário LIWC não possui apenas POS tags e polaridade, como também possui classificações psicolinguísticas.

Inicialmente, a ideia era identificar um grupo de palavras que abordasse dimensões emocionais e cognitivas básicas frequentemente estudadas em campos da psicologia, tais como psicologia social, psicologia da saúde e psicologia da personalidade (PENNEBAKER et al., 2007). Naturalmente, o domínio das categorias de palavras expandiu-se consideravelmente (PENNEBAKER et al., 2007).

B.4.1 Processo de desenvolvimento do LIWC2007

O processo de desenvolvimento do LIWC2007 (PENNEBAKER et al., 2007) incluiu as etapas:

- **Etapa 1 - Coleta de LI:** Na concepção e desenvolvimento das escalas categóricas do LIWC, foram gerados conjuntos iniciais para cada escala categórica.

  Esse conjunto inicial de categorias foram baseados em palavras de diversas fontes. Pennebaker et al. (2007) basearam-se em escalas comuns de avaliação de emoções, como PANAS (WATSON; CLARK; TELLEGEN, 1988), Roget's International Thesaurus (KIPFER, 2019) e dicionários padrão de inglês. Após a criação da lista preliminar de LI das categorias, sessões de *brainstorming* entre 3 a 6 avaliadores foram realizadas, nas quais LI relevantes para cada categoria foram geradas e adicionadas às listas iniciais das categorias.

- **Etapa 2 - Parecer dos Avaliadores:** Uma vez que todas as palavras das listas foram reunidas, as palavras nas categorias *Psychological Processes* (Processos Psicológicos) e *Personal Concerns* (Preocupações Pessoais) e a maioria nas categorias Relativity (Relatividade) (excluindo tempo verbal) foram então avaliadas por três avaliadores independentes. No desenvolvimento do primeiro *software* LIWC, os juízes foram orientados a focar tanto na inclusão quanto na exclusão de LI em cada lista das categorias do dicionário LIWC. Na primeira fase de classificação, os avaliadores indicaram se cada palavra da lista de categorias deveria ou não ser incluída na categoria específica em questão. Eles também foram

[12] O *software* LIWC está disponível pelo *link* `https://www.liwc.app/`.

instruídos a incluir palavras adicionais que considerassem que deveriam ser incluídas na categoria. Todas as listas de LI das categorias foram atualizadas seguindo o seguinte conjunto de regras:

1. um LI permanece na lista de categorias se, e somente se, dois dos três avaliadores concordarem que ela deveria ser incluída;
2. um LI é excluída da lista de categorias se, e somente se, pelo menos dois dos três avaliadores concordarem que ela deveria ser excluída;
3. um LI é adicionada à lista de categorias se, e somente se, se dois dos três juízes concordassem que ela deveria ser incluída.

A segunda fase de classificação envolveu a discriminação dos LI de cada categoria. Os avaliadores receberam a lista de LI em ordem alfabética de cada categoria. Os avaliadores foram então instruídos a indicar em quais subcategorias de nível médio, se houver, o LI deveria ser incluído, como *Insight* (Entendimento) e *Causation* (Causalidade). Todas as listas de LI das subcategorias foram atualizadas seguindo o seguinte conjunto de regras:

1. um LI permanece na lista de subcategorias se, e somente se, dois dos três avaliadores concordarem que ela deveria ser incluída;
2. um LI é excluída da lista de subcategorias se, e somente se, pelo menos dois dos três avaliadores concordarem que ela deveria ser excluída.

- **Etapa 3 - Avaliação Psicométrica:** A avaliação inicial do LIWC ocorreu entre 1992 e 1994. Uma revisão significativa do LIWC foi realizada em 1997 para aprimorar o programa original e os dicionários. Arquivos de texto de várias dezenas de estudos, totalizando mais de 8 milhões de palavras, foram analisados usando a versão de 1997 do LIWC, bem como o Word-Smith, um programa avançado de contagem de palavras usado em análise do discurso (SCOTT, 2008, 1996). Categorias originais do LIWC que eram usadas em taxas muito baixas (menos de 0,3% das palavras constituíam a categoria) ou que apresentavam constantemente baixa confiabilidade ou validade foram omitidas. Várias categorias novas, incluindo *Social Processes* (Processos Sociais), várias categorias de *Personal Concern* (Preocupações Pessoais) e as dimensões de *Relativity* (Relatividade), foram adicionadas seguindo os mesmos procedimentos rigorosos baseados em avaliadores descritos anteriormente (incluindo ambas as etapas). Finalmente, uma vez que foi finalizado o novo dicionário LIWC, quaisquer palavras que não eram usadas pelo menos 0,005% do tempo em nossos arquivos de texto anteriores ou que não estavam listados em *Frequency Analysis of English Usage* (FRANCIS; KUCERA, 1982) foram excluídas.

- **Etapa 4 - Atualizações e Expansões:** A versão LIWC2007 envolveu atualização substancial dos dicionários e modificação na estrutura do dicionário. Baseando-se em mais de centenas de milhares de arquivos de texto compostos por centenas de milhões de palavras de amostras de linguagem escrita e falada, procurou-se identificar palavras comuns e categorias de palavras não capturadas nas versões anteriores do LIWC. Examinou-se as 2.000 palavras usadas com mais frequência, um grupo de quatro avaliadores realizou a análise dessas palavras e verificou quais novas palavras e novas categorias de palavras eram apropriadas para inclusão. Com base em estudos recentes que sugerem que palavras funcionais são particularmente relevantes para processos psicológicos, foi adicionado as categorias de *Conjunctions* (Conjunções), *Adverbs* (Advérbios), *Quantifiers* (Quantificadores), *Auxiliary Verbs* (Verbos Auxiliares), *Common Verbs* (Verbos Comumente Usados), *Impersonal Pronouns* (Pronomes Impessoais), *Total Function Words* (Palavras de Função Total) e *Total Relativity Words* (Palavras de Relatividade Total). Além disso, os pronomes de terceira pessoa foram divididos em 3ª pessoa do singular e 3ª pessoa do plural. Finalmente, um grande grupo de sinais de pontuação foi adicionado como categorias separadas.

B.4.2 Processo de desenvolvimento do LIWC2015

Já o processo de desenvolvimento do LIWC2015 (PENNEBAKER et al., 2015) incluiu as seguintes etapas:

- **Etapa 1 - Coleta de LI:** Na concepção e desenvolvimento das escalas categóricas do LIWC, foram gerados conjuntos de palavras para cada dimensão conceitual, utilizando o dicionário LIWC2007 como ponto de partida. Após a criação da lista preliminar de LI das categorias, sessões de *brainstorming* entre 2 a 6 avaliadores foram realizadas, nas quais LI relevantes para cada categoria foram geradas e adicionadas às listas iniciais das categorias.

- **Etapa 2 - Parecer dos Avaliadores:** Uma vez que todas as palavras das listas foram reunidas, cada palavra do dicionário foi examinada por um grupo de 4 a 8 avaliadores e avaliada qualitativamente em termos de “adequação” para cada categoria. Para que uma palavra permanecesse em determinada categoria, a maioria dos avaliadores teve que concordar com sua inclusão. Em casos de disputas, diversos corpora e fontes online foram consultados para determinar o uso comum, a inflexão e o significado de uma palavra. Palavras para as quais os avaliadores não conseguiram decidir a categoria adequada foram removidas do dicionário.

- **Etapa 3 - Análise da Frequência Básica:** Um vez que uma versão funcional do dicionário foi construída a partir das avaliações dos avaliadores, textos de di-

versas fontes foram analisados usando o Auxiliar de Extração de Significado (do inglês, *Meaning Extraction Helper* — MEH)[13] (BOYD, 2018) para determinar a frequência com que as palavras do dicionário foram usadas em vários contextos (MARKOWITZ, 2021). Essas fontes incluíam postagens em blogs, estudos de linguagem falada, Facebook e X, romances, escritos de estudantes e vários outros. Palavras do dicionário que não ocorreram pelo menos uma vez em vários corpora foram omitidas do dicionário.

- **Etapa 4 - Geração de Candidatos à Lista de LI:** Para ampliar a lista de LI do dicionário, foi explorado diversas fontes de linguagem para palavras de alta frequência que não haviam sido acrescentadas pelos avaliadores. Usando MEH, palavras de alta frequência foram quantificadas como uma porcentagem do total de palavras para centenas de milhares de arquivos de texto de vários estudos e fontes. Para diversas categorias linguísticas (por exemplo: verbos e adjetivos), o Stanford CoreNLP[14] (KRISTINA TOUTANOVA DAN KLEIN; SINGER, 2003; MANNING et al., 2014) foi usado em conjunto com o MEH para identificar palavras comuns. Todas as palavras candidatas foram então correlacionadas com todas as categorias do dicionário, a fim de detectar palavras comuns que ainda não estavam incluídas no dicionário. Palavras que se correlacionaram positivamente com as categorias do dicionário foram adicionadas a uma lista de palavras candidatas para possível inclusão. Em seguida, 4 a 8 avaliadores revisaram as listas de LI e verificaram: se as palavras deveriam ser incluídas no dicionário e se as palavras eram adequadas conceitualmente para categorias específicas do dicionário.

- **Etapa 5 - Avaliação Psicométrica:** Após a realização dos passos descritos anteriormente, cada categoria linguística foi dividida em suas palavras constituintes. Cada palavra foi, então, quantificada como uma porcentagem do total de palavras encontradas em 181.000 arquivos de texto, provenientes de 5 corpora, somando cerca de 231.000.000 de palavras. Todas as palavras de cada categoria foram tratadas como uma "resposta" e usadas para calcular estatísticas de consistência interna para cada categoria linguística como um todo. Um grupo de 2 a 8 avaliadores então revisou a lista de LI candidatas e votou sobre a inclusão das mesmas. LI para as quais não se estabeleceu uma maioria dos votos foram omitidas.

- **Etapa 6 - Fase de Refinamento:** Após a realização dos passos descritos ante-

[13] O *software Meaning Extraction Helper* está disponível pelo *link* `https://www.ryanboyd.io/software/meh/`.

[14] O *software Stanford CoreNLP* está disponível pelo *link* `https://stanfordnlp.github.io/CoreNLP/`.

riormente, eles foram repetidas novamente. Isso foi feito para detectar possíveis erros/descuidos que pudessem ter ocorrido durante o processo de criação do dicionário. Vale ressaltar que a avaliação psicométrica de cada categoria linguística mudou de forma insignificante durante cada fase de refinamento. Durante a última etapa da fase de refinamento final, dois avaliadores revisaram o dicionário em busca de possíveis erros.

- **Etapa 7 - Adição de Novas Categorias:** Uma grande mudança que ocorreu do LIWC2015 em relação com as suas verões anteriores foi a inclusão de quatro novas categorias: *Analytical thinking* (Pensamento analítico) (PENNEBAKER et al., 2015), *Clout* (Influência) (KACEWICZ et al., 2014), *Authentic* (Autêntico) (NEWMAN et al., 2003), and *Emotional tone* (Tom emocional) (COHN; MEHL; PENNEBAKER, 2004). Cada variável resumida foi derivada das descobertas encontradas e convertida em percentis com base em pontuações padronizadas de amplas amostras de comparação. Deve-se enfatizar que as variáveis resumidas são as únicas dimensões não transparentes na saída do LIWC2015.

B.4.3 A tradução do LIWC

O LIWC possui tradução para vários idiomas, como francês (PIOLAT et al., 2011), espanhol (RAMÍREZ-ESPARZA et al., 2007) e chinês (HUANG et al., 2012). LIWC2007pt (BALAGE FILHO; PARDO; ALUÍSIO, 2013) é a tradução para Português Brasileiro baseada na versão padrão do dicionário LIWC de 2007 (PENNEBAKER; BOOTH; FRANCIS, 2007).

O LIWC2007pt não apresenta resultados muito satisfatórios, isso deve-se ao fato de que o RL apresenta, segundo Carvalho; Santos; Guedes (2018), numerosos erros associados à ortografia, bem como à classificação e categorização inadequadas . No entanto, o LIWC2015pt, a tradução para Português Brasileiro baseada na versão padrão do dicionário LIWC de 2015, apresenta resultados muito superiores (CARVALHO et al., 2023). Isso ocorre visto que o RL aprimorou o seu processo de desenvolvimento, tornando-o muito mais rigoroso. Outro fator que pode ter impactado os resultado pode ter sido a utilização de métodos de tradução melhores ou ter realizado uma revisão mais rigorosa após a etapa de tradução.

Entretanto, embora o LIWC2015pt apresente resultados melhores que o LIWC2007pt, o SL LIWC2015pt não foi utilizado, pois, o recurso só é acessível mediante a compra da licença do *software* LIWC. Enquanto que a versão LIWC2007pt está disponível para *download*, de forma gratuita, em vários lugares.

B.5 Onto.PT

Oliveira; Gomes (2014) desenvolveram o Onto.PT[15], uma ontologia léxico-semântica, LR que têm ao mesmo tempo propriedades de um léxico e de uma ontologia (STAAB; STUDER, 2009; HUANG et al., 2010).

Ontologias léxico-semânticas são constituídas por uma Base de Conhecimento Léxico-Semântico (do inglês, *Lexical-Semantic Knowledge Base* — LSKB) (OLIVEIRA; GOMES, 2011; OLIVEIRA; SANTOS; GOMES, 2014). As LSKBs são LR caracterizados por ter uma cobertura significativa de palavras de uma língua, inter-relacionadas através de relações semânticas estabelecidas através do significado das palavras (OLIVEIRA, 2018).

B.5.1 Utilização de recursos de domínio público

O LR Onto.PT foi criado usando o ECO, visando superar as limitações proporcionada por Bases de Conhecimento Léxico (LKBs — do inglês *Lexical Knowledge Bases*) em português (OLIVEIRA; GOMES, 2014). Essas limitações incluem: ser estruturado como uma *wordnet*; abranger uma ampla gama de relações semânticas; ser criado automaticamente por meio da exploração de recursos textuais disponíveis e LKBs em português e ser de domínio público.

O ECO foi implementado usando LR disponíveis para o português. Uma das principais fontes de conhecimento utilizadas na criação do Onto.PT foram dicionários, em razão da sua organização em palavras e significados, bem como as suas definições sistemáticas e vocabulário. Foram utilizados os seguintes dicionários: a rede léxico-semântica PAPEL 2.0 (Palavras Associadas Porto Editora - Linguateca) (OLIVEIRA; SANTOS; GOMES, 2010), extraída automaticamente de um dicionário proprietário; o dicionário eletrônico Dicionário Aberto (SIMÕES; SANROMÁN; ALMEIDA, 2012; SIMÕES; IRIARTE SANROMÁN; ALMEIDA, 2013) e o dicionário eletrônico Wiktionary.

Além disso, foram utilizados dois *thesaurus* baseados em *synset*: TeP 2.0 (MAZIERO et al., 2008), elaborado por especialistas e OpenThesaurus.PT, criado de forma colaborativa. OpenThesaurus é o *thesaurus* oficial usado pelo OpenOffice (NABER, 2004). O projeto OpenThesaurusPT[16] começou como um esforço colaborativo para portar o projeto alemão OpenThesaurus para outras línguas, como o português (TEIXEIRA; SARMENTO; OLIVEIRA, 2010). O projeto OpenThesaurus fundamenta-se no Wordnet e utiliza *synsets* para armazenar informações sobre sinônimos e antônimos. Ele também conserva algumas informações sobre relações de superordenação e su-

[15] o recurso Onto.PT pode ser baixado através do link `http://ontopt.dei.uc.pt/index.php?sec=download_ontopt`.

[16] O LR OpenThesaurusPT pode ser encontrado no link `http://openthesaurus.caixamagica.pt/`. A lista de palavras da versão do ano de 2006 pode ser encontrada pelo link `https://paginas.fe.up.pt/~arocha/AED1/0607/trabalhos/thesaurus.txt`.

bordenação (isto é, hipernímia e hiponímia) entre os conjuntos de sinônimos.

B.5.2 Processos de desenvolvimento do Onto.PT

A técnica proposta por Oliveira; Gomes (2014) consiste em três etapas: Extração, Clusterização e Ontologização. A Tabela 13 ilustra o resultado para cada uma das etapas utilizadas na desenvolvimento do ECO.

Extração		
gado	s.m.	conjunto de animais criados para diversos fins; rebanho
tb_triple_1 = rebanho SINONIMO_DE gado		
tb_triple_2 = animal MEMBRO_DE gado		
Clusterização		
$synset_1$ = {manada, rebanho, mancheia, boiada}		
$synset_1 + tb\_triple_1$ = {manada, rebanho, mancheia, boiada, gado}		
Ontologização		
$synset_2$ = {bicho, animal, alimal, béstia, minante}		
sb_triple_1 = $synset_2$ MEMBRO_DE $synset_1$		

Tabela 13 – Exemplo das três etapas de criação. Fonte: Adaptado de (OLIVEIRA; GOMES, 2014)

Para integrar o conhecimento léxico-semântico, o ECO exige que esta informação seja representada como triplos relacionais, denotando instâncias de relações semânticas. Esses triplos consistem em dois argumentos conectados por um tipo de relação, como em:

$$animal \quad \text{HIPERONIMO\_DE} \quad cachorro$$

Na etapa de extração é realizada a aquisição automática de triplas baseados em termos (do inglês, *term-based triples* — tb-triples), triplas relacionais que conectam LI. As regularidades nas definições dos dicionários são exploradas nesse processo para extrair instâncias de relações semânticas, as quais conectam palavras identificadas pelo seu lema. Ademais, instâncias que envolvem a extração de relações semânticas são adquiridas e mantidas entre os itens lexicais.

Observe as definições:

- candeia s.f. utensílio doméstico rústico usado para iluminação¸ com pavio abastecido a óleo
- espiga s.f. parte das gramíneas que contém os grãos
- inquietar v.t. causar ansiedade
- severo adj. grave , crítico

A partir dessas definições é possível extrair diversas relações, tais como:

- *iluminação* HIPERONIMO_DE *candeia*
- *iluminação* FINALIDADE_DE *candeia*
- *grão* PARTE_DE *espiga*
- *inquietar* CAUSADOR_DE *ansiedade*
- *grave* SINONIMO_DE *severo*
- *crítico* SINONIMO_DE *severo*

Na etapa de clusterização, os sentidos das palavras são descobertos explorando as informações extraídas. Observando exclusivamente as relações de sinonímia, verifica-se que estas tendem a formar *clusters* (agrupamentos) de palavras. Esses *clusters* são descobertos automaticamente e devem, além de agrupar palavras semelhantes, incluir somente palavras que estejam conectadas, direta ou indiretamente, por sinonímia, o que significa que podem ser vistas como *synsets*. Em outras palavras, esta etapa resulta no estabelecimento de conceitos e sentidos de palavras, podendo, assim, ser entendida como um tipo de indução do sentido das palavras. O resultado é um *thesaurus*, onde cada conceito é representado por um conjunto de palavras sinônimas, assemelhando-se a uma *wordnet*.

Na etapa de ontologização, originalmente batizada como *ontologising* por Pantel (2005), consiste em associar os *synsets* descobertos a uma representação do seu significado. A ontologização é realizada sobre os tb-triples não sinonímicos, com o intuito de transformar o conhecimento estruturado em termos para uma estrutura ontológica, organizada em conceitos. Cada LI de um tb-triple é atribuído ao *synset* mais adequado. Caso não exista um *synset* adequado, um novo *synset* é criado com esse LI.

- S: (adj) **hospital**, bem-querente, bem-intencionado, caridoso, benevolente, benévolo
- S: (adj) **hospital**, caridoso, esmoler, benfazejo, bem-fazejo, bem-fazente
- S: (s) **hospital**, esprital, nosocômio
  - temParte
    - S: (s) enfermaria
      - parteDe
      - hiperonimoDe
        - S: (s) ambulatório
      - hiponimoDe
      - meioPara
        - S: (s) paciente, enfermo, doente
  - referidoPorAlgoComPropriedade
    - S: (adj) hospitalar, hospitalário
    - S: (adj) nosocómico, nosocomial
  - hiperonimoDe
    - S: (s) gafaria, leprosaria
    - S: (s) leprocómio
    - S: (s) lazareto
    - S: (s) hospital psiquiátrico, manicômio, hospício, rilhafoles, casa-de-orates, manicómio
  - hiponimoDe
    - S: (s) edifício, edificamento, edificação
    - S: (s) instituto, instituição, fundação, instauração, infra-estrutura, estabeleza, implantação, implante, estabelecimento, inauguração
    - S: (s) telhado, mobiliário, teito, tecto, habitação, dinastia, meisom, casa, cosque, mesão

Figura 33 – Relações e *synsets* a partir da entrada 'hospital'. Fonte: Adaptado de (OLIVEIRA; GOMES, 2011)

A Figura 33 mostra o resultado das Relações e *synsets* a partir da entrada 'hospital' no onto.pt on-line. A atribuição de sentimento ao ONTO.PT foi realizada de forma automática por Oliveira; Santos; Gomes (2014), utilizando os *synsets* encontrados na criação do ONTO.PT, bem como o SL SentiLex-PT (SILVA; CARVALHO; SARMENTO, 2012)).

B.6 OpLexicon

Souza et al. (2011); Souza; Vieira (2012) desenvolveram o SL OpLexicon[17] (*Opinion Lexicon*). A técnica proposta por Souza et al. (2011) consiste na utilização de três métodos existentes na literatura, que formam três léxicos de opinião diferentes que se unem para criar um grande léxico.

B.6.1 O método de Turney

O primeiro método utilizado é o de Turney (2002), que é baseado em *corpus*. O *corpus* utilizado nos experimentos contém 1.316 documentos e cerca de um milhão de palavras, resultante das junção de dois *corpus*. O primeiro corpus é composto por 346 resenhas de filmes escritas em português brasileiro e extraídas dos sites CinePlayers e Cinema com Rapadura. O segundo corpus é composto por 970 textos jornalísticos

[17] O recurso OpLexicon pode ser baixado pelo *link* `https://github.com/sillasgonzaga/lexiconPT/blob/master/data-raw/oplexicon_v3.0.zip`.

sobre diferentes temas extraídos do corpus PLN-BR CATEG[18] (BRUCKSCHEN et al., 2008).

Todas as expressões foram extraídas do corpus e anotadas utilizando Informações Mútuas Pontuais (verificar a seção 4.1.2). Para garantir uma maior precisão, apenas as expressões cuja polaridade estava acima da polaridade média da classe foram selecionadas.

B.6.2 O método de Kamps

O segundo método utilizado é o de Kamps et al. (2004), que é baseado em *thesaurus*. Este método baseia-se em uma função de distância baseada em sinonímia e antonímia, que é definida como o comprimento do menor caminho entre o caminho mínimo de sinônimos de uma palavra para outra, ou entre o caminho mínimo de seus antônimos.

A Figura 34 mostra o Conjuntos de Sementes utilizados para realizar o cálculo da polaridade.

bom, ótimo, excelente, feliz, brilhante, fenomenal, fantástico, espetacular, melhor, satisfatório : **Positiva**

ruim, péssimo, horrível, infeliz, estúpido, odioso, pior, feio, insatisfatório : **Negativa**

Figura 34 – Conjunto de Sementes

O cálculo da polaridade de uma palavra, $word$, é definida pela Equação (10) (OSGOOD; SUCI; TANNENBAUM, 1957):

$$EVA(word) = \frac{min[d(word, p)] - min[d(word, n)]}{min[d(p, n)]} \tag{10}$$

para cada p e n.

[18] O *dataset* PLN-BR CATEG representa uma amostra aleatória estratificada e proporcional em relação à distribuição do *dataset* PLN-BR FULL,com relação aos textos dos cadernos do jornal. Este conjunto é composto por 30% dos textos do corpus PLN-BR FULL, abrangendo aproximadamente 30 mil textos e 9.780.220 tokens. Ademais, o PLN-BR CATEG inclui exclusivamente notícias e reportagens que estão sob os direitos de republicação da Folha de São Paulo (BRUCKSCHEN et al., 2008).

O *dataset* PLN-BR pode ser obtido entrando em contato com sandra@icmc.usp.br. Mais informações podem ser obtidas pelo link `https://sites.google.com/icmc.usp.br/corpus-pln-br/`.

Onde:

- EVA representa a *Evaluation* (Avaliação);
- $word$ é a palavra analisada;
- d é a distância entre as palavras;
- p é uma semente positiva;
- n é uma semente negativa;
- $Freq_{t,d}$ é a quantidade de vezes que o termo ocorre em uma publicação;
- Max é o termo que possui maior número de ocorrências na publicação;
- N é o número total de publicações;
- n_t é o número de publicações que possui o termo t.

Como LR, utilizou-se o *thesaurus* TeP 2.0 (*Electronic Thesaurus for Brazilian Portuguese*) (MAZIERO et al., 2008), que pode ser visto como uma extensão do TeP. O TeP foi desenvolvido utilizando as diretrizes da WNP (WordNet de Princeton) (FELLBAUM, 1998), consequentemente, o TeP 2.0 é pautado sobre as mesmas diretrizes. O TeP 2.0 pauta-se especificamente: na divisão das unidades lexicais nas categorias nome, verbo, adjetivo e advérbio; no construto synset e na definição de antonímia especificada pelos desenvolvedores da WNP.

O conjunto de sinônimos (do inglês, *synonym set* — synset) constitui o elemento fundamental na estrutura de uma rede wordnet, representando um grupo de unidades lexicais sinônimas ou quase sinônimas. Essas unidades permitem que o falante deduza o conceito que elas evocam. Em outras palavras, o synset é uma coleção de unidades lexicais da mesma categoria sintática que são intercambiáveis em determinados contextos, por exemplo: *bicycle*, *bike*, *wheel*, *cycle*, que em determinados contextos podem ser utilizados para se referir a *bicycle* (bicicleta). Por definição, um synset é estruturado para representar um único conceito, o qual é lexicalizado por suas unidades constituintes.

Já a antonímia abrange diversos tipos de oposição de significado. Existem três categorias: **antonímia complementar:** associa pares de itens lexicais em uma relação de contradição, onde a validade de um implica a falsidade do outro, tal como vivo e morto; **antonímia gradual:** liga itens lexicais indicando extremos opostos de uma escala, tal como pequeno e grande e **antonímia recíproca:** une pares de itens lexicais que se implicam mutuamente, tal como comprar e vender.

O TeP 2.0 possui 44.077 LI e anotação de *synsets* e antonímia. De maneira similar ao que ocorre na WNP para a língua inglesa, ao calcularmos a distância semântica

entre os adjetivos “bom” e “ruim”, descobrimos que eles estão intimamente ligados (KAMPS et al., 2004). Embora “bom” e “ruim” sejam antônimos, portanto com significados opostos, ainda assim estão relacionados devido à relação de antonímia. Parte dessa relação pode ser atribuída à ampla aplicabilidade desses adjetivos.[19]

O cálculo da distância entre os adjetivos“bom” e “ruim” é 4, sendo que “bom” apresenta 10 sentidos (em comparação com 25 no WNP para a língua inglesa) (KAMPS et al., 2004). Dessa forma, foi selecionado apenas palavras cuja distância para pelo menos um dos conjuntos de sementes é igual ou inferior a 3.

B.6.3 O método de Mihalcea

O terceiro método utilizado é o de Mihalcea; Banea; Wiebe (2007), que é baseado na utilização um sistema de tradução automática online em vez de um dicionário bilíngue. O LR utilizado no terceiro método é o *Liu's English Opinion Lexicon*[20] (HU; LIU, 2004). Esse recurso é composto por uma lista contendo cerca de 6.800 palavras de sentimento em inglês, divida em dois arquivos: um contendo a lista de palavras com polaridade positiva e outra contendo a lista de palavras com polaridade negativa.

Para contornar as desvantagens do método de Mihalcea; Banea; Wiebe (2007), como o uso do dicionário bilíngue e o trabalho manual envolvido, foi utilizado o serviço de tradução online Google Tradutor. Todas palavras e expressões foram traduzidos. Os LI que o sistema de tradução não conseguiu traduzir - devido à grande presença de variação linguística, como erros ortográficos comuns, no léxico original - foram descartados por revisão manual.

B.6.4 A aplicação dos três métodos

A aplicação dos três métodos descritos gerou três SL distintos, compostos por 359 expressões oriundos do SL baseado em corpus, 2.400 palavras oriundos do SL baseado em *thesaurus* e 4.909 expressões oriundos do SL baseado em tradução. O SL final, resultante da combinação desses três conjuntos, é composto por 7.077 LI com polaridade definida, excluindo-se os LI com polaridade neutra. Nos casos em que a polaridade de um LI divergia entre duas fontes, a decisão foi tomada com base em heurísticas simples, considerando a confiabilidade das fontes. Observa-se que, para o léxico baseado em tradução, criado a partir de recursos originalmente destinados a outras línguas e considerando a imperfeição do processo de tradução, a preferência foi dada à outra fonte em situações de conflito.

[19] Considere o “Problema do Mundo Pequeno”, no qual Milgram (1967) desenvolveu o conceito dos seis graus de separação. Esta teoria sugere que apenas seis elos ou laços de amizade são necessários para conectar duas pessoas quaisquer no mundo.

[20] O recurso *Liu's English Opinion Lexicon* pode ser baixado através do *link* `https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html`.

B.7 ReLi-Lex

Freitas et al. (2012) desenvolveram o ReLi[21] (**Re**senhas de **Li**vros), um *corpus* de resenhas de livros anotado manualmente quanto à expressão de opinião. O ReLi foi criado no âmbito do projeto Anotadores Semânticos baseados em Aprendizado Ativo, do Laboratório de Engenharia de Algoritmos e Redes Neurais (do inglês, Laboratory of Engineering of Algorithms and Neural Networks — LEARN), coordenado por Ruy Milidiú (Departamento de Informática - PUC-Rio).

O ReLi contém 1600 resenhas de 13 livros, de um total de 7 autores, totalizando 12.470 frases e 259.978 palavras. Para cada livro foram coletadas cerca de 200 resenhas e, quando esse número não pôde ser atingido, as resenhas foram completadas com outras obras do mesmo autor até atingir um número próximo de duzentos. As resenhas possuem formato livre, e não há separação formal entre pós e contras.

O ReLi foi automaticamente anotado com POS *tags* e manualmente anotado quanto à expressão de opinião. No processo de anotação da opinião, foram marcadas (i) a polaridade da frase que expressa opinião; (ii) segmento(s) que expressa(m) a opinião; (iii) polaridade desses segmentos. O corpus possui 4.210 opiniões positivas e 1.024 opiniões negativas, sendo 2.883 sentenças positivas, 596 sentenças negativas e 212 sentenças contendo ao mesmo tempo opiniões positivas e negativas.

B.7.1 Processos de desenvolvimento do ReLi-Lex

Freitas (2013) desenvolveram o SL ReLi-Lex[22] (*ReLi Lexicon*). O ReLi-Lex foi derivado do *corpus* ReLi. A partir do corpus anotado, foram extraídas as sequências marcadas como núcleos de opinião e agrupadas segundo a POS. Dessa lista de sequências marcadas como núcleos de opinião, foram selecionadas manualmente as candidatas a entradas. As classes consideradas foram adjetivo, verbo, substantivo e expressões multi-vocabulares[23] (do inglês, *Multi-Word Expressions* — MWE). Para cada classe, há uma lista de entradas com os lemas, obtidos automaticamente com o analisador morfossintático PALAVRAS (BICK, 2000), e as polaridades.

O processo de seleção das entradas não servia apenas para eliminar as palavras que não indicavam a presença de opinião e polaridade, como também auxiliava a identificar LI que apenas apresentavam polaridade/opinião no contexto específico de resenhas de livros. Para aproveitar as particularidades da qualificação encontradas nesse contexto, foi criada entradas separadas com LI específicos do domínio cultural. Por exemplo, os LI *imprevisível* e *perturbador* são características/ações considera-

[21] O *dataset* ReLi pode ser obtido através do *link* `https://www.linguateca.pt/Repositorio/ReLi/`.

[22] O recurso ReLi-Lex pode ser baixado pelo *link* `https://www.linguateca.pt/Repositorio/ReLi/`.

[23] Expressões multi-vocabulares são LI formados por duas ou mais palavras que atuam como uma única 'unidade', exibindo características idiossincráticas formais e/ou funcionais em comparação às combinações livres de palavras (MASINI, 2019). Em outras palavras, são expressões que adquirem um significado que difere do significado das palavras quando consideradas isoladamente.

das positivas no corpus, mas dificilmente apresentem essa característica em outros contextos, como pode-se observar nos exemplos a seguir:

O mais interessante é que o final foi *imprevisível* e surpreendente.

Sim, o final é ainda mais genial e *perturbador*!.

Para assegurar maior precisão na análise dos LI seguiu-se as seguintes diretrizes: confirmar se o LI comumente expressa afetividade e opinião e averiguar se a polaridade expressa pelo LI é de caráter geral ou específica de algum domínio.

Para realizar essas diretrizes, os LI foram comparados com as ocorrências nos variados corpora do Projeto AC/DC[24] (Projeto de Acesso a corpos/Disponibilização de corpos) (COSTA; SANTOS; ROCHA, 2009; FREITAS; SANTOS; GONÇALVES, 2011; SANTOS, 2012; SANTOS; BICK, 2000), especificamente no corpus Floresta, que inclui textos jornalísticos e de blogs (FREITAS; SANTOS, 2015; BICK et al., 2002).

Dois critérios nortearam a incorporação de um LI no ReLi-Lex: o LI deveria apresentar polaridade e opinião e o LI deveria apresentar (relativa) estabilidade quanto ao tipo de polaridade. Se o uso de um LI se restringia a um tipo de texto (especificamente, textos opinativos), o LI seria incluído. Se, por outro lado, a polaridade/opinião flutuava sem permitir a identificação de qualquer regularidade, o LI seria excluído.

Após o término do processo de incorporação dos LI ao ReLi-Lex, o SL foi separado em oito arquivos. Cada arquivo contendo um tipo de classe e um tipo de polaridade (positivo ou negativo), totalizando 568 LI.

B.8 SentiLex-PT

Existe duas versões do SentiLex-PT (*Sentiment Lexicon for Portuguese*). A primeira versão (SentiLex-PT01[25]) foi desenvolvida por Silva et al. (2010), enquanto que a última versão (SentiLex-PT02[26]) foi desenvolvida Silva; Carvalho; Sarmento (2012) (CARVALHO; SILVA, 2015).

A técnica proposta por Silva; Carvalho; Sarmento (2012) consiste na expansão automática de um léxico de sentimentos em português para identificação de opiniões referentes a entidades humanas. Esta metodologia baseia na identificação de candidatos a adjetivos humanos através de um conjunto limitado de LR: um léxico de adjetivos e dicionários de nomes de pessoas, profissões e cargos oficiais. Estes LR são

[24] O projeto AC/DC, iniciado em 1999, surgiu da necessidade de juntar os poucos recursos disponíveis num único ponto na rede e dessa forma facilitar a comparação e a reutilização do material. O Projeto AC/DC pode ser acessado através do *link*: `https://www.linguateca.pt/ACDC/`.

[25] O recurso SentiLex-PT01 pode ser baixado pelo *link* `https://github.com/caiomsouza/u-tad-eds-proyecto-final/tree/master/lexicon/SentiLex-PT01`.

[26] O recurso SentiLex-PT02 pode ser baixado pelo *link* `https://b2share.eudat.eu/records/93ab120efdaa4662baec6adee8e7585f`.

empregados para selecionar adjetivos potenciais de uma vasta coleção de n-gramas. O léxico de adjetivos é também utilizado, em conjunto com um dicionário de sinônimos, na fase de determinação da polaridade.

Para extrair opiniões direcionadas a entidades humanas, Silva; Carvalho; Sarmento (2012) direcionaram a sua atenção para os predicados adjetivais. Segundo Overall; Vallejos; Gildea (2018), os predicados adjetivais são conceituados como predicados nos quais se verifica a presença de atribuições ao sujeito, isto é, adjetivos que modificam substantivos. Por exemplo, nas frases "*I am hungry*" (eu estou faminto) ou "*my father is tall*" (meu pai é alto), os adjetivos "*hungry*" (faminto) e "*tall*" (alto) modificam o sujeito, dando-lhes características. Tais predicados adjetivais são, portanto, reconhecidos como estruturas que expressam tanto propriedades permanentes quanto temporárias (COSTA ARRAIS; GALUCIO, 2020).

B.8.1 O Léxico de Adjetivos

O Léxico de Adjetivos utilizado contém 24.792 lemas, parcialmente anotados com a sua categoria semântica (humano, não-humano) e polaridade (positivo, negativo, neutro). 4.546 registros incluem informação sobre a sua possível categoria semântica: 4.034 adjetivos foram atribuídos manualmente ao atributo humano e os 511 lemas restante foram atribuídos ao atributo não-humano. Os adjetivos humanos são caracterizados por coocorrerem com um sujeito humano (por exemplo, o primeiro-ministro é popular). Em contrapartida, caso isso não ocorra, são caracterizados com não-humanos (por exemplo, o primeiro-ministro é esporádico).

Para expandir o Léxico de Adjetivos, foi explorado o uso da sinonímia entre adjetivos em diversos tesouros públicos disponíveis para o português. Foi utilizado especificamente os LR PAPEL 2.0[27] (OLIVEIRA; SANTOS; GOMES, 2010), TeP (MAZIERO et al., 2008) e DicSin[28]. Esses RL anteriormente mencionados abrangem 87.327 lemas diferentes; distribuídos em 136.913 pares de sinônimos, dos quais 36.326 envolvem adjetivos.

A polaridade dos adjetivos pode ser caracterizada com um espectro de adjetivos polares (KENNEDY, 2001a). Os adjetivos polares são classificados conforme a sua extensão de significado: adjetivos com polaridade positiva expressam atributos de pessoas de forma positiva, enquanto que adjetivos com polaridade negativa expressam atributos de pessoas de forma negativa (KENNEDY, 1997, 2001b, 2013).

Os adjetivos polares foram recolhidos, na sua maioria, de uma base de dados léxico-sintáctica de adjetivos intransitivos humanos disponíveis no português europeu contemporâneo (CARVALHO, 2007). Este recurso descreve sistematicamente as propriedades sintácticas e semânticas de 4.250 lemas e substantivos morfo-

[27] O recurso PAPEL 2.0 pode ser baixado através do link `https://www.linguateca.pt/PAPEL/`.
[28] O recurso DicSin pode ser consultado pelo link `https://www.dicsin.com`.

sintaticamente associados (por exemplo, estúpido/estupidez).

B.8.2 Dicionários de Nomes, Profissões e Cargos Oficiais

O Dicionário de Nomes foi compilado a partir de listas públicas de nomes de professores do ensino secundário colocados no recrutamento de 2009, disponíveis no site do Ministério da Educação de Portugal. Foi obtido uma lista de 562 nomes próprios e 1388 sobrenomes. Os nomes próprios correspondem ao primeiro elemento da combinação de cada nome, enquanto os sobrenomes são identificados como os elementos remanescentes, após a remoção de todas as preposições e conjunções possíveis, bem como dos dois primeiros tokens de cada nome. Isso se deve ao fato de, em países lusófonos (países de língua portuguesa), ser comum que as pessoas tenham dois nomes próprios.

O Dicionário de Profissões e Cargos Oficiais é composto por 383 lemas (aproximadamente 1200 formas flexionadas) que denotam uma profissão ou cargo oficial. As entradas do dicionário foram compiladas de maneira semi-automática a partir de corpora de notícias, explorando estruturas sintáticas nas quais tais tipos de substantivos ocorrem tipicamente (por exemplo, em aposição a uma entidade nomeada humana).

B.8.3 Processos de desenvolvimento do SentiLex-PT

O *corpus* utilizado nos experimentos é o o WPT05[29], uma coleção de mais de 10 milhões de documentos da web portuguesa (BATISTA; SILVA, 2010). Foi utilizado os n-gramas (e as suas frequências) gerados a partir dos documentos da coleção com língua automaticamente identificada como português, aproximadamente 7 milhões de documentos com 26 GB de texto (aproximadamente 7 milhões de documentos, totalizando 26 Gb de texto).

Filtraram-se os tokens com comprimento superior a 32 caracteres, mas não foi excluído n-gramas de baixa frequência do conjunto. Buscava-se encontrar a ocorrência de múltiplos padrões com um determinado lema no processo de classificação, n-gramas de baixa frequência podem se combinar para produzir padrões de alta frequência. Foi explorado 8 milhões de unigramas, 501 milhões de trigramas, 984 milhões de tetragramas e 1.321 milhões de pentagramas. Este corpus contém uma amostra ampla e representativa de documentos em português disponíveis na Web, incluindo uma gama abrangente de tipos e gêneros de textos.

Durante a identificação de múltiplos padrões vinculados a um lema específico no processo classificatório, n-gramas de baixa frequência podem se combinar para produzir padrões de alta frequência. Foram examinados 8 milhões de unigramas, 501 milhões de trigramas, 984 milhões de tetragramas e 1.321 milhões de pentagramas, formando um corpus que oferece uma amostra ampla e fiel dos documentos em por-

[29] O *dataset* WPT05 pode ser obtido através do *link* `https://www.linguateca.pt/WPT/WPT05.html`.

tuguês disponíveis online, abrangendo uma vasta gama de tipos e gêneros de textos.

Para identificar candidatos a adjetivos humanos, foi criado uma biblioteca de padrões léxico-sintáticos feitos à mão, representando construções copulares e adnominais elementares onde tais predicados podem ser encontrados. Essa biblioteca de padrões é então aplicada ao WPT05, para reunir evidências sobre o comportamento dos adjetivos. As frequências de adjetivos e padrões no corpus de n-gramas são então usadas como recursos de entrada para um classificador binário que foi treinado e testado usando os adjetivos rotulados manualmente. Para refinar os resultados fornecidos pelos padrões léxicos e filtrar potenciais casos errados, observou-se: o número de correspondências no corpus e o número e tipo de padrões instanciados.

Por exemplo, o adjetivo ventilado corresponde apenas uma vez, enquanto impotente tem um total de 97 correspondências, instanciando 9 padrões diferentes. Pode-se inferir que o adjetivo impotente tem maior probabilidade de ser considerado um adjetivo humano válido do que ventilado. Também pode-se supor que os adjetivos que não correspondem a nenhum padrão em toda a coleção são raros na linguagem ou não têm comportamento humano.

Foi treinado um classificador estatístico para distinguir automaticamente adjetivos com alta evidência humana (do inglês, *adjetives with high-human evidence* — AWHHE) de adjetivos com baixa evidência humana (do inglês, *adjetives with low-human evidence* — AWLHE). A classificação automática é realizada com base nas seguintes características identificadas: frequência de cada padrão; número total de padrões instanciados; frequência de correspondências e frequência do lema no n- coleção de gramas. Esses atributos estão associados a cada adjetivo reconhecido em nosso léxico original, inclusive aqueles cuja categoria semântica já é conhecida.

Uma vez que os adjetivos humanos no léxico foram identificados, concentrou-se em atribuir polaridades aos AWHHE. O procedimento começa derivando um grafo de sinônimos, denominado *syngraph* (*synonym graph*), onde os nós são os lemas humanos previamente identificados e as arestas representam relações de sinonímia entre lemas.

Cada nó do *syngraph* é nomeado como a concatenação de um lema, sua POS tag e classe semântica. Esta combinação, nomeada como *qualiflemma* (*qualified lemma*), foi anteriormente aplicada à normalização de verbetes em dicionários. Ao ter uma rede de *qualiflemma*, em vez de apenas lemas, podemos evitar a propagação de relações de sinonímia entre lemas de categorias sintático-semânticas distintas (homógrafos), e possibilitar a atribuição de diferentes polaridades a tais lemas.

Para atribuir polaridades automaticamente aos *qualiflemma* não rotulados, foi treinado um classificador estatístico, que explora um vetor de características extraído do *syngraph*. Foi utilizado 80% dos *qualiflemma* polares para gerar o *syngraph* e os 20% restantes foram utilizados para aprender um modelo para atribuir polaridades a lemas.

No *syngraph* encontramos nós que apresentam as seguintes polaridades: -1, 0, 1 e nulo, onde nulo designa polaridade não atribuída. O objetivo é aprender um modelo que preveja a polaridade de um nó com polaridade nula dada a informação de polaridade de sua vizinhança.

Um *qualiflemma* no *syngraph* com polaridade não atribuída pode possivelmente ter nós adjacentes exibindo as quatro polaridades distintas, tornando a decisão complexa. Uma situação em que todos os nós adjacentes possuem polaridade nula é bastante comum. No entanto, podemos tentar observar através dos nós adjacentes e atribuir polaridades com base nas polaridades da nuvem dos *qualiflemma* conectados no *syngraph*. Capturamos essas informações calculando os caminhos mais curtos e as distâncias até os nós mais próximos com polaridade atribuída.

A distancia foi calculada usando o *Dijkstra's shortest-path algorithm* (Algoritmo do Caminho Mais Curto, proposto por Dijkstra (1959)). O cálculo foi realizado sobre um *syngraph* modificado, ao qual foi adicionado três nós iniciais, rotulados como "1", "-1" e "0", cada um representando um valor de polaridade. Esses nós são então diretamente conectados a todos os nós que representam *qualiflemma* com a mesma polaridade atribuída.

As distâncias de cada *qualiflemma* q, para os nós positivo, zero e negativo correspondem, respectivamente, às características d_{pos_q}, d_{zer_q} e d_{neg_q} usadas na classificação estatística subsequente. Além dessas características, também foi calculado três pesos de polaridade (w_{pos_q}, w_{zer_q} e w_{neg_q}) como a soma dos inversos das distâncias de cada nó ao nó inicial correspondente.

Para w_{pos_q}, temos:

$$w_{pos_q} = \sum_i \frac{1}{1 + d_{pos_i}},$$

onde i representa os nós adjacentes a q.

A maior parte das informações referente à polaridade foram atribuídas manualmente. No entanto, a polaridade de alguns adjetivos foram automaticamente classificados utilizando a ferramenta Classificador Léxico de Análise de Julgamento (do inglês, *Judgment Analysis Lexicon Classifier* — JALC), desenvolvida por Silva; Carvalho; Sarmento (2012) (CARVALHO; SILVA, 2015).

As formas flexionadas dos verbos e das expressões idiomáticas, bem como os respetivos atributos morfológicos, foram extraídos semi-automaticamente do LABEL-LEX-sw[30] (RANCHHOD; MOTA; BAPTISTA, 1999; ELEUTÉRIO et al., 2003). O recurso LABEL-LEX é constituído de um sistema que integra diversos recursos linguísticos de ampla cobertura, como dicionários e gramáticas, desenvolvidos pela equipe

[30]O recurso LABEL-LEX-sw pode ser obtido através do site `https://label.ist.utl.pt/pt/downloads_pt.php`.

do Laboratório de Engenharia da Linguagem (LabEL) (CARVALHO; SILVA, 2015).

Após todas as etapas descritas, o resultado foi um LR contendo 7.014 lemas e 82.347 formas flexionadas. Os adjetivos e substantivos foram flexionados em gênero (masculino e feminino), enquanto os atributos morfológicos que caracterizam os verbos ou expressões idiomáticas foram flexionados em tempo, pessoa e número.

B.9 SentiWordNet-PT-BR

Bastos (2023) desenvolveu o SentiWordNet-PT-BR[31] baseando-se na pontuação das palavras do SentiWordNet e padronização da identificação lexical das palavras do WordNet. O SentiWordNet-PT-BR foi desenvolvido com base em alguns projetos existentes, são eles: WordNet (FELLBAUM, 1998), Open Multilingual Wordnet (BOND; PAIK, 2012; BOND et al., 2016), OpenWordnet-PT (PAIVA; RADEMAKER; MELO, 2012) e SentiWordNet (ESULI; SEBASTIANI, 2006; BACCIANELLA; ESULI; SEBASTIANI, 2010).

B.9.1 WordNet

A WordNet[32], também chamada de WordNet de Princeton (do inglês, Wordnet of English — PWN), é uma rede de palavras interligadas através de um conjunto de sinônimos (do inglês, *synonym set — synset*) (FELLBAUM, 1998). O *synset* é o elemento fundamental na estrutura de uma rede *wordnet* e é representado por um grupo de unidades lexicais. Essas unidades compreendem itens linguísticos que são sinônimos ou quase sinônimos, permitindo ao usuário inferir o conceito que elas evocam.

Um *synset* pode ser compreendido como um conjunto de unidades lexicais que pertencem à mesma categoria sintática (como Substantivos, verbos, adjetivos e advérbios) e são estruturados para representar conceitos distintos, expresso pelas suas unidades constituintes. Synsets são interligados por meio de relações conceituais-semânticas e lexicais. Os itens linguísticos presentes em um *synset* são intercambiáveis, permitindo que sejam permutados entre si em diferentes contextos sem alterar o significado fundamental da frase.

[31] O LR SentiWordNet-PT-BR pode ser baixado pelo *link* `https://github.com/Pedro-Thales/SentiWordNet-PT-BR`.

[32] O LR WordNet pode ser baixado pelo *link* `https://wordnet.princeton.edu/node/5`.

Figura 35 – Exemplo de resultado da busca de palavras no WordNet Search

A WordNet possui uma versão on-line chamada WordNet Search[33]. A Figura 35 apresenta o resultado obtido ao pesquisar a palavra *bicycle* no WordNet Search. A busca retorna dois *synsets*: um para substantivos e outro para verbos. Além de retornar os sinônimos ou quase sinônimos, também apresenta a definição que representa cada *synset*.

No *synset* dos substantivos, temos os itens linguísticos: *bicycle*, *bike*, *wheel* e *cycle*, que é definido como *"a wheeled vehicle that has two wheels and is moved by foot pedals"* (um veículo com rodas que tem duas rodas e é movido por pedais). Já no *synset* dos verbos temos os itens linguísticos: *bicycle*, *cycle*, *bike*, *pedal* e *wheel*, que é definido como *"ride a bicycle"* (andar de bicicleta). Em ambos os *synsets*, os itens linguísticos presentes em cada *synset* podem ser substituídos por termos que estão no mesmo *synset*, sem que o contexto da frase seja alterado.

B.9.2 Open Multilingual Wordnet

O Open Multilingual Wordnet (OMW) é um projeto que visa fornecer o acesso de maneira facilitada a *wordnets* em vários idiomas, todos vinculados ao PWN. O OMW está disponível em duas versões:

- **OMW Versão 1[34]:** a primeira versão vincula *wordnets* criados manualmente e criados automaticamente para mais de 150 idiomas por meio do WNP (BOND; PAIK, 2012);

[33] O WordNet Search pode ser acessado através do navegador, pelo site `http://wordnetweb.princeton.edu/perl/webwn`.

[34] O primeira versão do Open Multilingual Wordnet pode ser acessada através do navegador, pelo site `https://omwn.org/omw1.html`.

- **OMW Versão 2**[35]**:** Já a segunda versão, em vez de vincular todos os idiomas através do PWN, utiliza o *Collaborative Interlingual Index* (Índice Interlingual Colaborativo) para vincular as *wordnets*. Trata-se de uma ferramenta que permite a ligação de várias *wordnets* de diferentes idiomas, de forma semelhante às ontologias, facilitando a tradução e a comparação entre as palavras de diferentes línguas (BOND et al., 2016).

B.9.3 OpenWordnet-PT

O OpenWordnet-PT[36] é uma *wordnet* para o português disponibilizada de maneira gratuita. Assim como o PWN, possui tanto uma versão para download quanto uma versão online[37] (PAIVA; RADEMAKER; MELO, 2012).

B.9.4 SentiWordNet

O SentiWordNet[38] é um recurso que atribui três pontuações de sentimento (positividade, negatividade e objetividade) a cada *synset* do WordNet, com valores entre 0 e 1, cuja soma é sempre igual a 1 (ESULI; SEBASTIANI, 2006; BACCIANELLA; ESULI; SEBASTIANI, 2010).

B.10 UNILEX

O método proposto por Souza; Pereira; Dalip (2017) utilizou um *dataset*[39] contendo 14.083 publicações coletados da rede social X, tendo como principal assunto a política, que foram rotulados e comparados para a criação da base do dicionário. O dicionário foi nomeado de Léxico Unificado (do inglês, *Unified Lexicon* — UniLex[40]), referindo-se à unificação de diversas técnicas da literatura para a composição de um método de dicionário léxico.

O *dataset* primeiramente passou por uma etapa de pré-processamento, onde foi realizada a eliminação de acentuações, abreviações e *stopwords* (como preposições, artigos e conectivos das palavras conhecidas da língua portuguesa).

A remoção dos acentos e abreviações foi realizada com intuito de igualar as palavras. Isso foi realizado em razão da ocorrência de alterações tanto por causa da cultura que os usuários possuem na escrita ao se expressar, como também por erros de digitação ou erros ortográficos. As abreviações serão transformadas para a

[35] O segunda versão do Open Multilingual Wordnet pode ser acessada através do navegador, pelo site `https://github.com/omwn/omwn.github.io`.

[36] O RL OpenWordnet-PT pode ser baixado pelo *link* `https://github.com/own-pt/openWordnet-PT`.

[37] A versão on-line do OpenWordnet-PT pode ser acessado através do navegador, pelo site `https://www.openwordnet-pt.org/`.

[38] O RL OpenWordnet-PT pode ser baixado pelo *link* `https://github.com/aesuli/SentiWordNet`.

[39] O *dataset* utilizado pode ser baixado através do link `https://dicionariounilex.wixsite.com/unilex`.

[40] O RL UniLex pode ser baixado pelo link `https://dicionariounilex.wixsite.com/unilex`.

forma correta da linguagem, evitando que a análise interprete as mesmas palavras com significados diferentes.

A remoção de *stopwords* foi realizada com intuito de subtrair as palavras que não agregam um sentido específico na frase, ou seja, como a sua função é realizar ligação de uma palavra a outra, não devem ser consideradas na análise. As *stopwords* são removidas por não serem consideradas relevantes, isto é, por não refletirem a essência das palavras quando isoladas, diferentemente de outras que podem ser classificadas como positivas, negativas ou neutras.

A rotulação do *dataset* foi realizada através da classificação manual das publicações, indicando se o sentimento expresso na frase é positiva, negativa ou neutra. Na fase de classificação das polaridades, levaram-se em conta as intenções do usuário ao escrever a publicação. A opinião do rotulador não foi considerada, analisando-se apenas o conteúdo transmitido pelo *post*. As publicações que continham datas, menções a outros usuários ou *hashtags* foram classificados como neutro. A rotulação foi executada por quatro rotuladores, cada um classificando aproximadamente 3.500 publicações do *dataset*.

Foram criadas três listas de LI: a lista TB_SIN_01, armazena todas as palavras identificadas como *stopwords* para métodos de recuperação de informação na língua portuguesa; a lista TB_SIN_02 armazena as palavras cujas polaridades foram classificadas como positivas e a lista TB_SIN_03 armazena as palavras cujas polaridades foram classificadas como negativas.

A contabilização para a ocorrência das palavras segue o seguinte conjunto de regras: se a palavra estiver na lista TB_SIN_01 é descartada; se a palavra estiver na lista TB_SIN_02 é contabilizada como ocorrência positiva e se a palavra estiver na lista TB_SIN_03 é contabilizada como ocorrência negativa.

Realiza-se a contabilização das ocorrências para determinar o quão positiva ou negativa é uma palavra. O número de vezes que uma palavra aparece é utilizado como peso, e este peso é empregado para estimar a polaridade de um *post*.

Realizou-se a comparação entre as listas de LI com o objetivo de identificar aquelas que apresentam um número de ocorrências equivalente ou muito próximo. Caso a porcentagem de ocorrências de uma determinada palavra nas listas TB_SIN_02 ou TB_SIN_03 exceda 60% do total de ocorrências, ela é mantida em uma lista e excluída da outra. Além disso, quando as ocorrências foram iguais em ambas as listas, a palavra é removida das duas listas.

Dessa forma, é possível estabelecer uma filtragem mais pontual nas palavras contidas nas listas de LI, estabelecendo uma classificação mais eficiente e equilibrada entre o positivo e o negativo. Isso é feito indicando qual das listas exerce maior influência sobre a palavra, para que assim a palavra mantenha o sentimento correspondente nos resultados da análise.

Após a finalização da etapa de comparação entre a listas de LI, as listas TB_SIN_02 e TB_SIN_03 foram unificadas em uma única tabela. Nesta tabela, cada palavra foi incluída juntamente com a polaridade correspondente à sua lista de origem, resultando na criação do LR UNILEX, totalizando em um léxico de 3.845 LI.

Para poder realizar a comparação com outros SL foi necessário realizar a tradução das palavras para o inglês. As palavras foram traduzidas usando o Goslate[41], uma biblioteca de python que possibilita a tradução de forma gratuita através da API do Google Tradutor.

A comparação com outros SL foi realizada através do iFeel 2.0[42] (ARAÚJO et al., 2014; ARAUJO et al., 2016), uma aplicação Web gratuita, que permite detectar sentimentos em qualquer forma de texto, incluindo dados de mídias sociais não estruturados. O iFeel 2.0 pode realizar a SA utilizando diversos SL e em diversas línguas (ARAUJO et al., 2016; RIBEIRO et al., 2016; ARAUJO et al., 2016).

A comparação com outros SL também foi realizada com o uso da Frequência do Termo – Inverso da Frequência nos Documento (do inglês, *Term Frequency – Inverse Document Frequency* — TF-IDF), onde o peso de cada palavra seria substituído pelo valor da sua frequência (AIZAWA, 2003; WU et al., 2008; GROOTENDORST, 2022).

Segundo Baeza-yates; Ribeiro-neto (1999, 2011), a técnica TF-IDF é composta pelas técnicas: Frequência do Termo (TF — *Term Frequency*) e Inverso da Frequência nos Documento (IDF — *Inverse Document Frequency*). O TF refere-se à frequência com que palavras ou termos aparecem em um documento específico. Este cálculo é enriquecido pelo inverso da frequência do termo em toda a coleção de documentos, com o objetivo de reduzir a relevância de termos que surgem frequentemente nesses documentos. Já o IDF constitui uma abordagem estatística que fundamenta a ponderação de termos, baseando-se na função inversa do número de documentos em que o termo aparece (HARMAN et al., 2019).

O valor TF-IDF de uma palavra, independente de qual seja a polaridade, aumenta proporcionalmente com o número de suas ocorrências no documento. Esse valor é balanceado pelo inverso da frequência de cada termo. Assim, torna-se possível diferenciar o fato de que alguns termos são, em geral, mais comuns que outros em contextos positivos ou negativos, conforme demonstrado na Equação (11) (BAEZA-YATES; RIBEIRO-NETO, 1999, 2011):

$$W_{t,d} = \left(\frac{Freq_{t,d}}{Max}\right) \times \log_2\left(\frac{N}{n_t}\right) \tag{11}$$

[41] Informações sobre o Goslate pode ser encontradas no seguinte link `https://pythonhosted.org/goslate/#module-goslate`.

[42] Infelizmente, em julho de 2020 a ferramenta online do iFeel foi descontinuada. No entanto, ainda é possível testar a ferramenta através da sua versão em Docker, iFeel Docker. As instruções de como utilizar o iFeel Docker podem ser encontradas através do *link* `https://docs.google.com/document/d/1DrqMYz6z5v_xACzKgYIXfnS748SLqbpSdJSGt-emzUg/edit`.

Onde:

- $W_{t,d}$ é o peso do termo t no documento d;
- $Freq_{t,d}$ é a quantidade de vezes que o termo ocorre em uma publicação;
- Max é o termo que possui maior número de ocorrências na publicação;
- N é o número total de publicações;
- n_t é o número de publicações que possui o termo t.

B.11 WordNet Affect BR

Pasqualotti; Vieira (2008); Pasqualotti (2015) desenvolveram o WordNetAffectBR[43]. O WordNetAffectBR foi desenvolvido utilizando os seguintes RL: Base Affect, WordNet e WordNet Affect.

B.11.1 Base Affect

A Base Affect constitui-se como um LR elaborado manualmente, o qual engloba LI associados a estados afetivos. Parte das informações foram coletadas de dicionários e documentos científicos que tratam sobre a psicologia das emoções, o resto dos dados foram inseridas baseadas de forma intuitiva e arbitrária pelos autores, com revisões e validação por psicólogos e lexicógrafos.

A Base Affect é primordialmente estruturada segundo as classes gramaticais presentes na WordNet, abrangendo substantivos, verbos, adjetivos e advérbios. Cada registro na Base Affect inclui dados que estabelecem a correlação entre o inglês e o italiano, POS tags, assim como relações de sinonímia e antonímia. Além disso, cada entrada contém uma definição análoga à 'glossa' encontrada no LR WordNet, além de LI pertencentes a classes gramaticais distintas, os quais convergem para uma mesma categoria psicológica, e informações afetivas. As informações afetivas referem-se às teorias de emoção baseadas no conceito da avaliação cognitiva, teorias das emoções básicas e teorias dimensionais.

A teoria dimensional representa as emoções como um ponto em um espaço emocional, tais como *valence* (valência), *arousal* (excitação) e *dominance* (dominância) (MÄNTYLÄ et al., 2016). No entanto, somente *valence* e *arousal* foram utilizados na Base Affect.

A *valence* indica o quão positiva ou negativa é a emoção. Enquanto que *arousal* e *dominance* representam, respectivamente, o nível de ativação ou excitação associado a uma emoção e o grau de controle ou poder que uma pessoa sente quando

[43]O LR WordNetAffectBR pode ser baixado pelo *link* `https://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/wordnetaffectbr/`.

experimenta a emoção. Por exemplo, temos que a raiva é uma emoção com alto grau de *arousal*, enquanto a tristeza é apresenta *arousal* baixo. Por outro lado, a *dominance* pode ser alta em emoções como raiva, mas baixa em estados como medo ou submissão.

O conceito de avaliação cognitiva refere-se ao modelo de emoções OCC[44], proposto por Ortony; Clore; Collins (2022). O OCC é um modelo definido por psicólogos, que abrange um total de 22 emoções, originadas a partir de respostas a estímulos positivos ou negativos. A Figura 36 mostra a estrutura do modelo OCC.

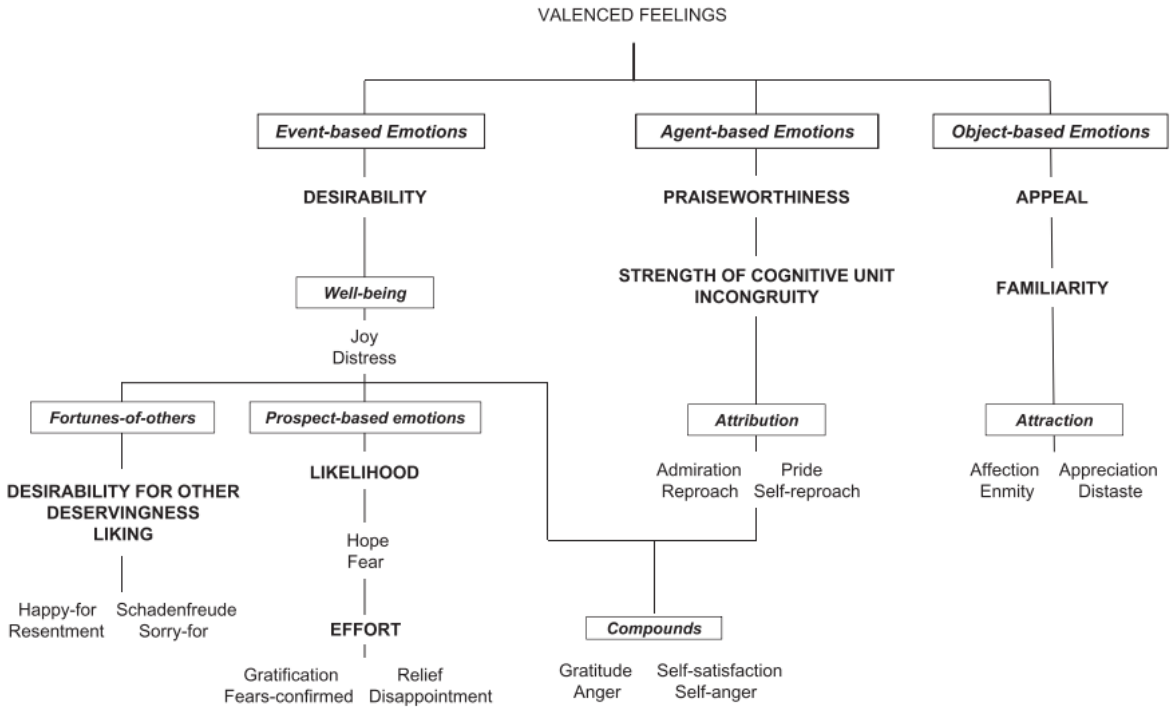

Figura 36 – Estrutura do Modelo OCC. Fonte: (ORTONY; CLORE; COLLINS, 2022)

No modelo OCC, as emoções são categorizadas levando em consideração o tipo de estímulo recebido:

- ***Event-based Emotions:*** Na categoria *Event-based Emotions* (Emoções baseada em Eventos) são categorizadas emoções relativas a *Well-being* (bem-estar), *Fortunes-of-others* (sorte dos outros) e *Prospect-based emotions* (emoções baseadas em perspectiva) que decorrem de eventos com outros indivíduos ou com o próprio indivíduo.

  – *Well-being:* decorrem de sentimentos relacionados ao bem-estar do vivenciador do evento. Na categoria *Well-being* enquadra-se os sentimentos de *Joy* (alegria) e *Distress* (sofrimento).

[44]OCC é um acrônimo derivado das iniciais dos sobrenomes de seus criadores, Ortony, Clore e Collins.

  - *Fortunes-of-others:* decorrem de sentimentos relacionados a sorte de uma pessoa ao vivenciar um evento. Na categoria *Fortunes-of-others* enquadra-se os sentimentos de *Happy-for* (feliz por), *Resentment* (ressentimento), *Schadenfreude*[45] e *Sorry-for* (desculpa por).
  - *Prospect-based emotions:* decorrem da perspectiva das crenças do vivenciador do evento sobre o *status* de um evento previsto. Os *status* podem ser categorizados como *Unconfirmed* (não confirmado), *Confirmed* (confirmado) ou *Disconfirmed* (desconfirmado). O *status* é categorizado como *Unconfirmed* quando o *status* é desconhecido pelo vivenciador, *Confirmed* quando o vivenciador acredita que o evento previsto aconteceu e *Disconfirmed* quando o vivenciador acredita que o evento previsto não aconteceu. Na categoria *Prospect-based emotions* enquadra-se os sentimentos de *Hope* (esperança), *Fear* (medo), *Gratification* (gratificação), *Fears-confirmed* (medos confirmados), *Relief* (alívio) e *Disappointment* (desapontamento).

- ***Agent-based Emotions:*** Na categoria *Agent-based Emotions* (Emoções baseada em Agentes) são categorizadas emoções de *Attribution* (Atribuição) que decorrem de ações com outros indivíduos ou com o próprio indivíduo. Nessa categoria enquadra-se os sentimentos de *Admiration* (admiração), *Reproach* (reprovação), *Pride* (orgulho) e *Self-reproach* (auto-reprovação).

  Além disso, também existem emoções *Compounds* (compostas), que levam em consideração os sentimentos de *Well-being* e *Attribution*. Da combinação de sentimentos dessas categorias surgiram os sentimentos *Gratitude*, *Anger*, *Self-satisfaction* e *Self-anger*.

- ***Object-based Emotions:*** Na categoria *Object-based Emotions* (Emoções baseada em Objetos) são categorizadas emoções de *Attraction* (atração) que são direcionadas a objetos. Nessa categoria enquadra-se os sentimentos de *Affection* (afeição), *Enmity* (inimizade), *Apprecition* (apreciação) e *Distaste* (desgosto).

A teoria das emoções básicas refere-se ao *framework* denominado Racionador Afetivo (do inglês, *Affective Reasoner* — AR), desenvolvido por Elliott (1992). O *framework* AR, fundamentado no modelo de emoções OCC, foi desenvolvido para para criar agentes[46] capazes de responder emocionalmente. A interface e a comunicação

[45] *Schadenfreude* é uma palavra composta das palavras alemãs *Schaden*, que significa dano, e *Freude*, que significa alegria, e é usada hoje em dia como um empréstimo na língua inglesa (VAN DIJK; OUWERKERK, 2014). A emoção *Schadenfreude* foi definida como uma emoção sentida ao obter prazer com o infortúnio de outra pessoa (SMITH, 2018).

[46] Agentes são sistemas computacionais capazes de realizar tarefas de forma autônoma ou semi-autônoma (WOOLDRIDGE, 1997; WEISS, 2001).

dos agentes com o mundo real podem dar-se através de expressões faciais ou em respostas por dialogo, estabelecendo, assim, os métodos pelos quais os agentes utilizam para se comunicar e expressar suas emoções durante a interação com o usuário. Na Base Affect, cada LI corresponde a uma das 24 categorias de emoção identificadas por Elliott (1992).

B.11.2 WordNet Domains e WordNet Affect

O WordNet Affect (VALITUTTI; STRAPPARAVA; STOCK, 2004; STRAPPARAVA; VALITUTTI et al., 2004) foi desenvolvido pelo grupo de pesquisa em Tecnologias Cognitivas e de Comunicação (do inglês, *Cognitive and Communication Technologies* — TCC), pertencente ao ITC-IRST, que corresponde ao Instituto Trentino de Cultura (do italiano, *Istituto Trentino di Cultura* — ITC), vinculado ao Instituto de Investigação Científica e Tecnológica (do italiano, *Istituto per la Ricerca Scientifica e Tecnologica* — IRST). O ITC-IRST, uma instituição europeia de pesquisa multidisciplinar especializada nas áreas de tecnologia, inovação, humanidades e ciências sociais, com sede em Trento, foi posteriormente renomeado como Fundação Bruno Kessler (do italiano, *Fondazione Bruno Kessler* — FBK).

A WordNet Affect é parte integrante da base lexical multi-lingual WordNet Domains[47] (MAGNINI; CAVAGLIA, 2000; BENTIVOGLI et al., 2004). A WordNet Domains é um LR criado de maneira semi-automática, aumentando o WordNet com rótulos de domínio.

Um domínio pode englobar *synsets* de distintas categorias sintáticas e de diversas sub-hierarquias no WordNet. Os domínios têm a capacidade de agrupar palavras que são utilizadas com o mesmo sentido em conjuntos homogêneos, o que contribui para a diminuição da polissemia nas entradas do WordNet. Os rótulos de domínio servem como uma ferramenta para atenuar a ambiguidade, sendo aplicáveis em algoritmos de desambiguação semântica das palavras, mitigando, assim, os desafios impostos pela polissemia[48] (MAGNINI et al., 2002; GLIOZZO; STRAPPARAVA; DAGAN, 2004; MAGNINI; STRAPPARAVA, 2004).

Como exemplo de palavra polissêmica, podemos citar, a palavra *bank* (banco), que na *wordnet* apresenta dez sentidos diferentes para a classe gramatical substantivo e oito para a classe gramatical verbo. Com a inclusão de um rótulo de domínio a polissemia seria solucionada, pois, estaria referenciando o uso da palavra a um domínio adequado. Por exemplo, usaria o rótulo economia para referenciar a entidade ou ins-

[47] O LR WordNet Affect pode ser obtido através do formulário presente no *link* `https://wndomains.fbk.eu/download.html`.

[48] Um problema de polissemia ocorre quando uma palavra possui à capacidade de apresentar diversos significados, o que possibilita a elaboração de jogos de linguagem e trocadilhos, bem como amplia as dimensões de ambiguidade e conotação, dada a existência de palavras com múltiplas interpretações (JURAFSKY; MARTIN, 2023).

tituição financeira; o rótulo geografia para referenciar um banco de área presente na maré baixa; o rótulo arquitetura para referenciar o objeto que se dispõe ao redor de uma mesa, utilizado para as pessoas sentarem (PASQUALOTTI, 2008).

Os Synsets do WordNet foram anotados com pelo menos um rótulo de domínio semântico, selecionado dentre um conjunto de cerca de duzentos rótulos estruturados hierarquicamente (ALFIO GLIOZZO, 2009). Cada hierarquia é estruturada através de níveis de especificidade, onde, por exemplo, o domínio denominado de *religion* (religião), pertencente a hierarquia *DOCTRINES* (doutrinas), contém os subníveis *mythology* (mitologia), *occultism* (ocultismo) e *theology* (teologia) (BENTIVOGLI et al., 2004).

WordNet Affect acrescenta ao WordNet Domains uma hierarquia adicional de rótulos de domínios afetivos, com a qual os *synsets* que representam conceitos afetivos são anotados posteriormente. Esses rótulos podem ser: *emotion* (emoção), como raiva (substantivo) e temer (verbo); *mood* (humor), como animosidade (substantivo) e amável (adjetivo); *trait* (traço de personalidade), como agressividade (substantivo) e competitivo (adjetivo); *cognitive state* (estado cognitivo), como confusão (substantivo) e atordoado (adjetivo); *physical state* (estado físico), como doença (substantivo) e com tudo incluído (adjetivo); *hedonic signal* (sinal hedônico), como ferido (substantivo) e sofrimento (substantivo); *emotion-eliciting situation* (situação que provoca emoção), como estranheza (substantivo) e fora de perigo (adjetivo); *emotional response* (resposta emocional), como suar frio (substantivo) e tremer (verbo); *behaviour* (comportamento), como ofensa (substantivo) e inibido (adjetivo); *attitude* (atitude), como intolerância (substantivo) e defensivo (substantivo) e *sensation* (sensação), como frieza (substantivo) e sentir (verbo).

B.11.3 Processos de desenvolvimento do WordNetAffectBR

A construção da base de dados WordnetAffectBR baseou-se na diretriz de que os *synsets* identificados na Wordnet deveriam estar correlacionados com o código que determina a localização do *synset* na base Wordnet Affect. Além disso, determinou-se que a classificação dos rótulos afetivos, que foram baseados no modelo OCC, deveria ser “emoção”.

A criação da base de dados do WordnetAffectBR foi dividida em três etapas, visando ampliar e aprofundar a análise semântica e afetiva das palavras relacionadas a emoções. A criação da base de dados do WordnetAffectBR segue as seguintes etapas:

1. **Formação da Base Referencial:** Inicialmente, a WordnetAffectBR foi constituída por palavras relacionadas a emoções, especificamente classificadas como adjetivos e substantivos,encontradas no Modelo OCC. Esta etapa inicial criou

uma base referencial contendo 21 palavras, pois as emoções *fear* e *fears-confirmed* foram agrupadas em uma única emoção definida como *fear*;

2. **Expansão com Palavras de Synsets Relacionados:** A base de dados foi então expandida para incluir palavras do mesmo synset que estivessem no primeiro nível, dobrando o léxico referencial para 42 palavras. Esta expansão visou enriquecer a base com termos semanticamente e afetivamente relacionados às emoções iniciais.

3. **Ampliação Mediante Relações Semânticas e Lexicais:** A terceira etapa envolveu uma expansão mais complexa, utilizando as relações semânticas e lexicais definidas pelas bases Wordnet e os rótulos afetivos da Wordnet Affect. Para os adjetivos, foram aplicadas as relações de sinonímia, similaridade, e a indicação "ver também", enquanto que para os substantivos, as relações de sinonímia, hipernímia, e hiponímia foram exploradas. Esta etapa permitiu uma ampliação significativa do conjunto de dados, culminando em uma lista compreensiva de 403 palavras relacionadas a emoções.

Para criar o WordnetAffectBR também foi necessário realizar a tradução das palavras encontradas na criação da base de dados. A partir da tradução, foram identificadas palavras que as diferenças entre as traduções geravam relações de sinonímia, permitindo, dessa forma, incrementar a base de palavras. Em algumas situações, quando ocorria divergência entre as traduções, era escolhido uma das palavras e descartava-se as outras.

Algumas palavras apresentavam problemas de duplicidade e divergência quanto à sua aplicação. Para arrumas esses problemas as palavras duplicadas foram excluídas da lista final e algumas palavras tiveram que ser traduzidas através de um expressão para manter o sentido de palavra de emoção.

Após a etapa de validação, a lista original de palavras, que, após o processo de tradução, havia alcançado um total de 457 termos, foi refinada para um conjunto de 289 palavras. Para a criação da base afetiva, desenvolveu-se uma ferramenta de chat capaz de identificar palavras relacionadas a emoções no diálogo, representando-as na tela através de *emoticons*.

A figura 37 mostra a interface da ferramenta de chat. Essa representação visual foi viabilizada pela análise e categorização das palavras da base de emoções em 15 grupos distintos, cada um associado a uma ou mais imagens. Estas imagens foram cuidadosamente selecionadas de modo a refletir de forma fidedigna a emoção correspondente ao grupo de palavras ao qual estavam atreladas.

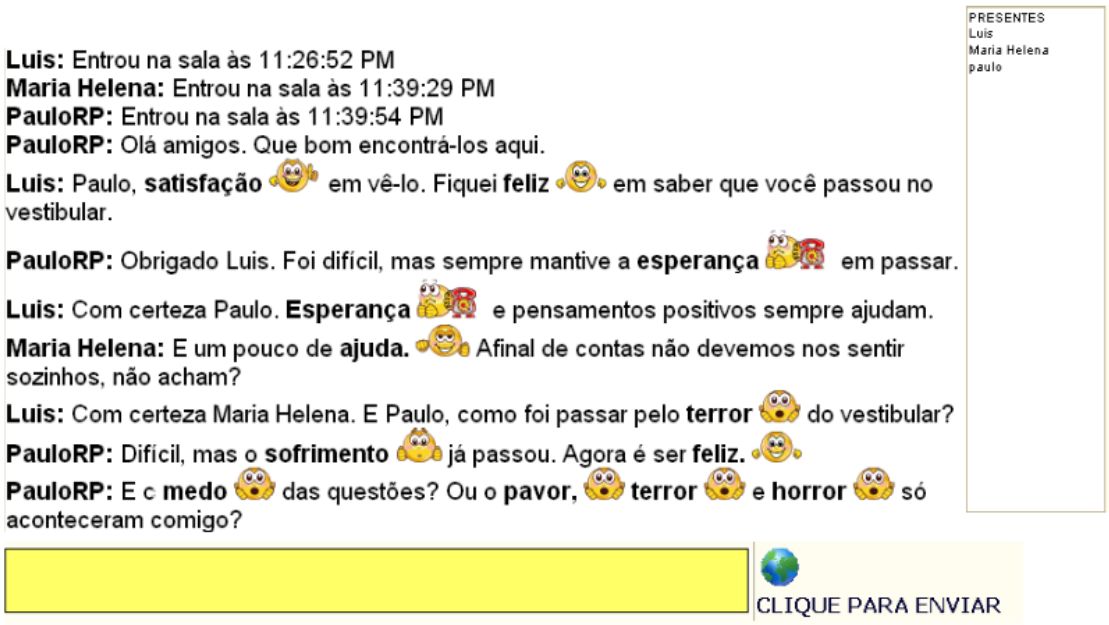

Figura 37 – Interface da tela do *Chat* - *Emoticon*. Fonte: (PASQUALOTTI, 2008)

A fim de validar essa metodologia de agrupamento e a eficácia das imagens na representação das emoções expressas pelas palavras, disponibilizou-se um formulário online. Este instrumento permitiu que um conjunto de pessoas avaliasse e emitisse sua opinião sobre a adequação e precisão da associação entre os grupos de palavras e as imagens selecionadas.

APÊNDICE C – Acesso aos Recursos Externos

Nesta seção estará disposto os *links* os quais os recursos utilizados no desenvolvimento deste trabalho podem ser obtidos.

C.1 Analisador de Dependência Sintática

spaCy: Informações sobre o spaCy podem ser encontradas através do *link* `https://spacy.io/`.

C.2 Dataset

***Dataset* da competição ABSAPT-2022:** O *dataset* usado na competição ABSAPT-2022 (SILVA et al., 2022) pode ser baixado entrando em contato com absapt2022@inf.ufpel.edu.br. Após o contato você receberá por e-mail com o **corpus** compactado e a senha necessária para descompactá-lo. Mais informações podem ser obtidas pelo link `https://sites.google.com/inf.ufpel.edu.br/absapt2022/`.

C.3 Ontologia de Domínio

Hontology: O LR Hontology está disponível para *download* através do *link* `https://portulanclarin.net/repository/browse/hontology/a83c9d04cb7a11e1a404080027e73ea2359e10ea62b940109aabe03684aa5ea4/`.

Protégé: O *software* Protégé está disponível para *download* através do *link* `https://protege.stanford.edu/`. Enquanto que a sua versão *on-line* pode ser acessada através do link `https://webprotege.stanford.edu/`.

C.4 Léxicos de Sentimentos

AffectPT-br: O LR AffectPT-br está disponível para *download* através do *link* `https://github.com/LaCAfe/AffectPT-br`

EmoLex: O LR EmoLex está disponível para *download* através do *link* `http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`

LeIA: O LR LeIA está disponível para *download* através do *link* `https://github.com/rafjaa/LeIA`.

LIWC2007pt: O LR LIWC2007pt está disponível para *download* através do *link* http://143.107.183.175:21380/portlex/index.php/pt/projetos/liwc.

Onto.PT: O LR Onto.PT está disponível para *download* através do *link* http://ontopt.dei.uc.pt/index.php?sec=download_ontopt.

OpLexicon: O LR OpLexicon está disponível para *download* através do *link* https://github.com/sillasgonzaga/lexiconPT/blob/master/data-raw/oplexicon_v3.0.zip.

Reli-Lex-PT: O LR ReLi-Lex está disponível para *download* através do *link* https://www.linguateca.pt/Repositorio/ReLi/.

SentiLex-PT02: O LR SentiLex-PT02 está disponível para *download* através do *link* https://b2share.eudat.eu/records/93ab120efdaa4662baec6adee8e7585f.

SentiWordNet-PT-BR: O LR SentiWordNet-PT-BR está disponível para *download* através do *link* https://github.com/Pedro-Thales/SentiWordNet-PT-BR.

UNILEX: O RL UniLex está disponível para *download* através do *link* https://dicionariounilex.wixsite.com/unilex.

WordNetAffectBR: O LR WordNetAffectBR está disponível para *download* através do *link* https://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/wordnetaffectbr/.

Anexos

ANEXO A – Estrutura da Ontologia Hontology

- **Acomodação**
  - Albergue
  - Alojamento de Férias
  - Apartamento
  - Bangalô e Campismo
  - Cama e café da manhã
  - Casa de Hóspedes
  - Hostal
  - Hotel
    - Hotel adega
    - Hotel aquático
    - Hotel casa da árvore
    - Hotel Casamata
    - Hotel cápsula
    - Hotel de gelo
  - Hotel Flutuante
  - Motel
  - Outra hospitalidade
  - Pensão
  - Pousada
- **Aparência**
- **Avaliação da Acomodação**
- **Categorias de Hotéis**
  - Conforto
  - Conforto classe superior

  - Conforto Superior
  - Luxo
  - Luxo superior
  - Padrão
  - Padrão superior
  - Primeira Classe
  - Turista
  - Turista superior
- **Endereço $\equiv$ Localização**
  - Cidade
  - Código Postal
  - País
  - Rua
- **Facilidade**
  - Acesso para Deficientes
  - Facilidades do banheiro
    * Balança
    * Banheira
    * Chuveiro
    * Pia
    * Produtos de Higiene
    * Roupão
    * Banheiro
    * Secador
    * Toalha
    * Torneira
  - Facilidades do Quarto
    * Ar-condicionado
    * Cabide
    * Cama
    * Candeeiro
    * Carpete

    - Chaleira
        - Chaleira Elétrica
        - Chaleira Manual
    - Colchão
    - Cortina
    - Cozinha
    - Espelho
    - Ferro de Engomar
        - Ferro a Vapor
    - Frigobar
    - Máquina de Café
    - Rede Sem Fio
    - Roupeiro
    - Sacada
    - Salgados
    - Telefone
    - Tomada
    - Travesseiro
    - TV
    - Mesa de Passar Roupa
- Instalação externa
    - Acesso à Internet
    - Bar da piscina
    - Golfe
    - Jacuzzi Exterior
    - Jardim
    - Parque de Estacionamento
    - Piscina exterior
    - Quadra
        - Quadra de futebol
        - Quadra de tênis
    - Rede Sem Fio
- Instalação interna
    - Acesso à Internet

- Berço
- Cabeleireiro
- Banheiro
- Cassino
- Centro de spa
  - Banho Turco
  - Ginásio $\equiv$ Sala de ginástica
  - Hidromassagem
  - Massagem
  - Sala de ginástica $\equiv$ Ginásio
  - Sauna
- Clube para crianças
- Corredor
- Discoteca
- Elevador
- Engraxamento de Sapatos
- Escadaria
- Lavanderia
  - Sala de lavanderia
  - Serviço de lavanderia
- Estacionamento
- Piscina coberta
- Praça de Brinquedos
- Quadra
  - Quadra de futebol
  - Quadra de tênis
- Rede Sem Fio
- Restaurante
- Reunião
  - Sala de reunião
- Sala de Conferências
- Sala de Jogos
- Sala de leitura
- Sala de Malas
- Salão de Baile

        - Salão de beleza
        - Sistema de Aquecimento
        - Tabacaria
        - Área de Fumantes
    - Motorista
- **Horário**
    - Horário da Piscina Coberta
    - Horário da Piscina Externa
    - Horário do Restaurante
    - Horário do Spa
- **Hospitalidade**
- **Ponto de interesse**
    - Aeroporto
    - Arena
    - Centro comercial
    - Centro da cidade
    - Centro histórico
    - Construção Histórica
    - Estação de Ônibus
    - Estádio
    - Igreja
    - Jardins
    - Monumento
    - Parada de Ônibus
    - Parque
    - Praia
    - Teatro
    - Universidade
- **Preço**
    - Preço da Internet

  - Preço do Bar
  - Preço do Berço
  - Preço do Café
  - Preço do Café da Manhã
  - Preço do Estacionamento
  - Preço do Quarto
  - Preço do Restaurante

- **Quadro de Funcionários**
  - Animador
  - Funcionários da Cozinha
  - Funcionários da Limpeza
  - Funcionários da Piscina
  - Funcionários da Recepção $\equiv$ Serviço de Check Out $\equiv$ Serviço de Check In
  - Funcionários da Área de Lazer
  - Funcionários do Bar
  - Funcionários do Restaurante
  - Gerência
  - Porteiro
  - Serviço de Check In $\equiv$ Funcionários da Recepção
  - Serviço de Check Out $\equiv$ Funcionários da Recepção

- **Quarto**
  - Quarto de albergue
    - Quarto Feminino com 10 Camas
    - Quarto Feminino com 12 Camas
    - Quarto Feminino com 4 Camas
    - Quarto Feminino com 6 Camas
    - Quarto Feminino com 8 Camas
    - Quarto Misto com 10 Camas
    - Quarto Misto com 12 Camas
    - Quarto Misto com 4 Camas
    - Quarto misto com 6 Camas

    - Quarto Misto com 8 Camas
    - Twin Privado com Casa de Banho Compartilhada
- Quarto de hotel
    - Duplo
        - SuperiorTwinRoomWithBalconyAndSeaView
        - TwinRoomWithLandView
        - TwinRoomWithPoolView
        - TwinRoomWithSeaView
    - Quarto de casal
        - ConfortDoubleRoomWithSeaView
        - DoubleRoomWithLandView
        - DoubleRoomWithPoolView
        - DoubleRoomWithSeaView
        - SuperiorDoubleRoomWithSideSeaView
    - Quarto de luxo
        - Quarto Duplo de Luxo com Vista para o Mar
        - Quarto Familiar de Luxo com Vista para o Mar
    - Quarto de solteiro
    - Quarto Familiar
        - Quarto Familiar Com Sacada
        - Suíte Familiar com Vista para o Mar
        - Suíte Familiar Júnior com Vista para o Mar
    - Quarto standard
    - Quarto Triplo
    - RomanticRoom
    - Suíte
        - Suíte Júnior
            - Suíte Júnior com Vista Frontal para o Mar
            - Suíte Júnior com Vista para o Mar
        - Suíte com Sacada e Vista para o Mar
- Tipos de Quartos em Apartamentos
    - Apartamento com 2 Camas
    - Apartamento com 3 Camas
    - Apartamento com Cama Individual

        - Apartamento de Luxo
        - Apartamento Queen com Duas Camas
        - Apartamento Studio
        - Duas Camas com Terraço
        - Quarto Duplo de Luxo
        - Quarto Familiar
            - Quarto Familiar Com Sacada
            - Suíte Familiar com Vista para o Mar
            - Suíte Familiar Júnior com Vista para o Mar
        - Quarto Twin de Luxo

- **Rede de hotéis**
    - AccorHotels
        - Adagio
        - AllSeasons
        - EtapHotel
        - Formula1
        - HotelF1
        - Ibis
        - Mercure
        - MGallery
        - Novotel
        - Orbis
        - Pullman
        - Sofitel
        - SuiteNovotel
        - ThalassaSea&Spa
    - BestWesternHotels
    - ChoiceHotelsInternational
    - DominaHotels
    - HiltonHotels
    - HyattHotels
    - IbirostarHotels
    - MarriottHotels

  - PestanaGroupHotels
  - PortoBayHotels
  - RadissonHotels
  - TivoliHotels
  - VilaGaléHotels

- **Refeição**
  - Almoço
  - Café da Manhã
  - Jantar

- **Serviço**
  - Aluguel de Bicicletas
  - Aluguel de Veículos
  - Babá
  - Café da Manhã no Quarto
  - Funcionários da Recepção $\equiv$ Serviço de *Check Out* $\equiv$ Serviço de *Check In*
  - Convidado - entrega diária do jornal escolhido
  - Porteiro
  - Posto de Turismo
  - Porteiro
  - Posto de Turismo
  - Serviço de *Check In* $\equiv$ Funcionários da Recepção
  - Serviço de *Check Out* $\equiv$ Funcionários da Recepção
  - Serviço de Câmbio
  - Serviço de Limpeza
  - Serviço de Limpeza a Seco
  - Serviço de Quarto
  - Serviço de Lua-de-mel
  - Transporte para o Aeroporto

- **Tipos de Hóspede**
  - Casal

    * Casal Jovem
    * Casal Maduro
- Colega de Trabalho
- Família
    * Família com Filhos Mais Jovens
    * Família com Filhos mais Velhos
    * Família Grande
- Grupo de Amigos
- Viajante de Negócios
- Viajante Sozinho