# Processamento de Linguagem Natural do Português Brasileiro para detecção de *cyberbullying*

José Luiz Villela M. Mioni
Programa de Pós-Graduação em Ciência da Computação
Universidade Estadual de Londrina
Londrina Paraná Brasil
jose.luiz.villela@uel.br

Cinthyan Renata S. C. de Barbosa
Programa de Pós-Graduação em Ciência da Computação
Universidade Estadual de Londrina
Londrina Paraná Brasil
cinthyan@uel.br

Bruno Fantineli C. de Oliveira
Programa de Pós-Graduação em Ciência da Computação
Universidade Estadual de Londrina
Londrina Paraná Brasil
bruno.fantineli@uel.br

## RESUMO

O objetivo geral desta pesquisa é a detecção automática de comportamentos violentos em redes sociais usando técnicas de Processamento de Linguagem Natural (PLN) e estratégias de Aprendizado de Máquina. A metodologia usada nesta pesquisa combina linguística computacional, mais especificamente o PLN, mecanismos básicos de Inteligência Artificial e *webscraping* com Python em uma ferramenta de PLN que sinaliza possível *cyberbullying* e discursos violentos em redes sociais. Esta ferramenta representa uma importante contribuição no combate ao *cyberbullying*, de modo a proteger os usuários dos efeitos dessa prática, como depressão, ansiedade e até mesmo o suicídio.

## PALAVRAS-CHAVE

Processamento de Linguagem Natural, Discurso de Ódio.

## ABSTRACT

The overall objective of this research is the automatic identification of violent behavior on social networks using Natural Language Processing (NLP) techniques and Machine Learning strategies. The methodology used in this research is a combination of computational linguistics, more specifically NLP, basic Artificial Intelligence mechanisms and *webscraping* with Python in an NLP tool that detects possible *cyberbullying* and violent speech on social networks. This tool represents an important contribution to the prevention of *cyberbullying*, in order to protect users from the effects of this kind of practice, including depression, anxiety and even suicide.

## KEYWORDS

Natural Language Processing, Cyberbullying.

## CCS CONCEPTS

• Computing methodologies • Artificial intelligence • Natural language processing • Discourse, dialogue and pragmatics.

## 1 Introdução

O mundo passa por diversas mudanças tecnológicas à medida que avanços ocorrem em diferentes frentes de desenvolvimento. Desde demandas de mercado a novos horizontes, isso tudo é obtido por meio de novas perspectivas e ferramentas.

Sendo parte fundamental da experiência humana, a capacidade de se comunicar e se conectar faz presente em diversas iniciativas tecnológicas e não tecnológicas ao passo que evoluímos enquanto sociedade. Seja na comunicação entre indivíduos que usam a mesma linguagem ou mesmo idioma ou na interação entre esses utilizando de linguagens artificiais e não só naturais para interagir com diferentes tipos de hardware e software, a habilidade de se comunicar e ser compreendido entre emissores e receptores permanece como um alicerce das evoluções tecnológicas e de qualidade de vida.

Um dos motivos da existência do Processamento de Linguagem Natural (PLN) é o intuito de permitir a comunicação entre indivíduo utilizando-se de língua natural e no auxílio de ferramentas tecnológicas na compreensão do discurso da língua humana por software e outros dispositivos. PLN vem sido usado por mais de cinquenta anos, crescendo além do campo da linguística à medida que computadores são cada vez mais utilizados na humanidade.

Segundo Jurafsky e Martin [1], o PLN tem como objetivo extrair textos em Linguagem Natural (LN) e executar tarefas relevantes, permitindo o diálogo entre homem e máquina, melhorando a comunicação humano-computador ou fazer processamento de texto ou fala (discurso).

O PLN abrange diversos usos da habilidade de detectar e processar a palavra escrita e suas informações [2], tais como *chatbots*, geração de

conteúdo [3] e diversos outros softwares utilizados no reconhecimento de palavras e discurso.

Os desafios que envolvem a linguística computacional são, em alguns casos, motivados pela perspectiva científica, onde alguém procura prover uma explicação para um fenômeno linguístico ou psicolinguístico [4], [5]. Em outros casos, a motivação pode ser puramente tecnológica, buscando montar um componente funcional de um sistema de linguagem natural [6].

À medida que a tecnologia avança e nos conectamos, problemas e soluções do mundo real são amplificados por meio de maior alcance e conexão proporcionados pelo mundo digital. Um dos problemas amplificados pelas maiores capacidades de comunicação permitidas pelo mundo digital é, infelizmente, a potencialização do impacto de um problema antigo já conhecido no mundo: o *bullying*. Sua versão digital, conhecida como *cyberbullying*, infelizmente já foi apontada como um fator agravante ou até mesmo como o principal fator motivador em casos de suicídio de jovens de diferentes idades [7].

*Cyberbullying*, discurso de ódio e outras atividades maléficas realizadas *online* podem criar ambientes agressivos e, às vezes, perigosos [7], sendo esses relacionados a distúrbios psicológicos e, em casos mais severos, resultar em ferimentos autoinduzidos ou até mesmo o suicídio, também apontado por Bauman, Russel and Walker [8].

Porém, diversos fatores relacionados ao dia a dia da comunicação digital dificultam o emprego de diferentes tecnologias no intuito da diminuição do impacto de relações tóxicas na internet. A grande diversidade de dados e demográficos de usuário, assim como a variedade tecnológica entre plataformas, formatos de agressão e escassez de produções, especificamente voltados à Língua Portuguesa faz da descoberta digital, fundamental na prevenção das ações trágicas no mundo real, complexas e desafiadoras.

Assim, este trabalho busca atuar na direção de mitigar problemas do mundo real gerados pelo *cyberbullying*. Justifica-se sua disseminação pelo devido aumento de número de conteúdos ofensivos, comportamentos tóxicos e assédio de indivíduos em ambientes da internet em diversas comunidades na web, dentre essas as de jogos e plataformas. Casos de depressão, ansiedade, outros traumas psicológicos e até mesmo suicídio ainda são relatados em várias plataformas.

O trabalho tem como objetivo a detecção automática de comportamentos violentos em redes sociais, usando técnicas de PLN e estratégias de Aprendizado de Máquina para poder assim, no futuro, possibilitar a prevenção de eventos trágicos motivados ou potencializados pela violência *online*, como tiroteios em escolas, casos clínicos de doenças psiquiátricas e perda de vida humana por meio do suicídio. Será fornecido suporte tecnológico, proteção individual e filtragem de conteúdo tóxico. Posto isso, os objetivos específicos do trabalho almejam lidar com os aspectos pragmáticos de textos da web a fim de oferecer uma acurácia maior à detecção de tais comportamentos, de modo a solucionar os problemas previamente identificados.

Os resultados estimados são: a atribuição e remoção de indivíduos perigosos do ecossistema *online*; contribuir para o estudo de PLN em conteúdos inseridos por usuários em plataformas como redes sociais; aumentar a percepção na comunidade sobre os perigos e riscos inerentes ao comportamento tóxico *online*; potencial prevenção de traumas psicológicos, violência e perda de vida humana; permitir o crescimento de software disponível sobre o tema na Língua Portuguesa.

Este trabalho se organiza como segue. Na Seção 2 trata-se do assunto *cyberbullying*. Na Seção 3 é apresentado o sistema de PLN partindo do Analisador Léxico-morfológico para podermos desenvolver a fase de Análise Pragmática. Nessa seção são apontados alguns dados resultantes da implementação. E, finalmente, na Seção 4, temos as conclusões e trabalhos futuros.

## 2 *Cyberbullying*

O suicídio é um problema de saúde pública que afeta, diariamente, populações do mundo inteiro. No Brasil, os dados são preocupantes e segundo um estudo realizado em 2017 pelo Sistema de Informações de Mortalidade (SIM) [9], um órgão que faz parte do Ministério da Saúde, a taxa de suicídio entre os jovens vem aumentando desde 2002. Esses dados, que apresentam um crescimento de 10% no número de mortes, mostra que o problema atinge, principalmente, indivíduos na faixa etária de 15 a 29 anos [10].

Infelizmente, casos de *cyberbullying* responsáveis por traumas psicológicos severos [11] [12] inclusive a perda de vidas devido ao suicídio, podem ser encontrados no mundo todo [13] [14].

Também se faz importante analisar a correlação de profanidade em texto [15], assim como mapear e incrementar o corpus de acordo com dados linguísticos [16] e verificar alguma técnica para separação de dialetos dentro da linguagem [17].

## 3 Processamento de Linguagem Natural

Processamento de Linguagem Natural, conhecido como PLN, é o campo da Inteligência Artificial (IA) que pesquisa como os computadores podem ser utilizados para entender e manipular texto ou fala em linguagem natural para fazer coisas úteis. Ou seja, desenvolve modelos computacionais para analisar e

gerar interações entre humanos e computadores, as quais ocorrem por meio de Linguagem Natural (LN) [18].

Pode-se definir o objetivo do PLN como a execução de tarefas envolvendo a linguagem humana, permitindo a comunicação humano-máquina, melhorando a comunicação entre pessoas ou simplesmente obtendo processamentos úteis a partir de um texto ou discurso oral [1].

Já as convenções que regem o uso da linguagem nas interações sociais é um vasto leque de competências linguísticas sociais, imprescindíveis no ato de comunicação, dentro da pragmática principalmente.

As habilidades de comunicação são as que se constituem pré-requisito para a atribuição de função comunicacional à linguagem, isto é, que possibilitam o uso da linguagem em um processo de interação [1].

A título de exemplo do ponto mencionado, diversos atos aumentam a complexidade da compreensão do discurso, enquanto o mesmo se conecta à profanidade ou violência. O discurso de profanidade ou o comportamento agressivo pode servir como incentivo social, ao contrário do *bullying*, com o objetivo de reforçar laços ou até mesmo amplificar esforços competitivos em grupo em esportes ou outras atividades que demandam coordenação de time.

Miura [19] salienta que a análise de uma sentença em linguagem natural pode resultar em mais de uma possível interpretação. Essa ambiguidade pode se manifestar de diversas formas, sendo elas: morfológica (léxica, flexiva ou léxico-flexiva), sintática (identificação do constituinte ou coesão) ou semântica (léxica, unidade polilexicais, escopo ou papel temático).

A *tokenização* aborda um processo de separação das palavras. Dessa forma, letras maiúsculas e minúsculas são reconhecidas, assim como palavras compostas, verificação ortográfica e quebra de caracteres.

O processo de *análise léxica* (ou *léxico-morfológica*) é composto da classificação em palavras que armazenam o seu significado.

A tarefa de *parsing* (*análise sintática)*, etapa atual do projeto, consiste da extração de informações sintáticas de uma frase representada por meio de regras gramáticas e/ou árvores sintáticas.

A *análise semântica* aborda o processo de análise do significado das palavras, ou seja, interpretar expressões fixadas, sentenças inteiras e enunciados no contexto [20].

Já em relação à *análise pragmática*, o estudo fundamenta-se em reconhecimento de palavras dentro de um contexto [21]. Não apenas a frase única, mas o contexto do texto é então analisado.

Para clarear o que significa cada fase de análise, pequenos exemplos serão elucidados.

O processo de tokenização consiste da separação de uma sentença em tokens, partes de uma sentença que futuramente serão processadas pelo analisador léxico.

A Análise Léxico-morfológica é responsável por: fazer a verificação ortográfica e classificação léxico-morfológica, podendo ser classificadas, por exemplo, como substantivo, verbo, advérbio, pronome, numeral, preposição, conjunção, interjeição, artigo e adjetivo; identificar as partes, segundo sua estrutura e formação em: radical, tema, vogal temática, dentre outras [22].

A Análise Sintática é composta de um conjunto de tarefas que definem a função sintática de cada token em uma frase. A tarefa da Análise Sintática ou *parsing* trata-se de extrair as informações contidas em uma frase, sendo estas representadas por meio de uma gramática [23].

A finalidade do *parsing* é analisar e gerar sentenças corretas de acordo com a estrutura de cada palavra [22]. O sistema de perguntas e respostas objetiva analisar uma pergunta formulada em linguagem humana e determinar sua resposta. Comumente atua em um domínio restrito, de maneira que os sistemas de PLN podem explorar o conteúdo do domínio específico na construção de suas bases de conhecimento [23].

Para reconhecer uma frase é necessário verificar a relação lógica entre as palavras. A análise sintática é aquela onde uma sequência de unidades lexicais, tipicamente uma oração, será decomposta para determinar sua descrição estrutural de acordo com uma gramática formal [24].

Essa fase é responsável por organizar o conjunto das palavras e então aplicam-se regras gramaticais à sentença para reconhecer a estrutura e extrair seus significados [23]. Thanaki [25] diz que a análise sintática é uma área de grande importância para lidar com a sintaxe da LN, visto que algumas sentenças podem possuir erros.

Jurafsky e Martin [1] destacam os dois formalismos que desempenham papéis importantes à análise sintática: o primeiro é a *gramática de estrutura frasal*, ou seja, as palavras agrupam-se e formam uma unidade e essa forma constituintes. Já o segundo formalismo é a *gramática de dependência*, utilizada para processamento de fala e linguagem. Assim, é possível gerar uma estrutura de árvore pode meio da derivação da frase de entrada que descreve a formação sintática da sentença analisada. Outras gramáticas podem ser aplicadas, como as descritas por Barbosa [26] [27].

A análise semântica trata de analisar os significados das palavras, ou seja, interpretar as

expressões fixadas, sentenças inteiras e enunciados no contexto [20], pois as frases podem ser ambíguas.

Oliveira e Navaux [28] afirmam que a semântica pode ser dividida em léxica e gramatical. A *semântica léxica* busca uma representação conceitual para descrever o sentido, sendo que, para construir essa representação pode ser feita a decomposição semântica das unidades léxicas (em primitivas ou em traços semânticos) ou ser utilizadas redes semânticas. A *semântica gramatical* (também chamada composicional) procura identificar o sentido por meio de uma fórmula lógico-semântica. Porém, pode ocorrer ambiguidade.

Por fim, a Pragmática refere-se às convenções que regem o uso da linguagem nas interações sociais, ou seja, a um vasto leque de competências linguísticas sociais, imprescindíveis no ato de comunicação [29]. As habilidades de comunicação são as que se constituem pré-requisito para a atribuição de função comunicacional à linguagem, isto é, que possibilitam o uso dessa em um processo de interação [30]. A pragmática estuda os atos da fala e os contextos nos quais eles se realizam [31].

## 4 Ferramenta e recursos

O presente resultado da ferramenta aqui proposta é obtido por meio do uso da ferramenta Spark NLP [32], software estado da arte, largamente usado na indústria em diferentes atividades de Processamento de Linguagem Natural. Ferramentas como Spacy ou NTLK, apesar de excelentes em seus campos como, por exemplo, a análise sintática e catalogação de termos, não se aplicam de maneira específica como os componentes encontrados dentro do Spark NLP.

Um exemplo dos motivos do trabalho em se beneficiar da versatilidade e resultados que podem ser objetivos com a ferramenta Spark NLP são os resultados já atingidos na Língua Inglesa pela ferramenta. Tecnologias baseadas nessa podem ler sentimentos e *cyberbullying*, como demonstrados na Figura 1.

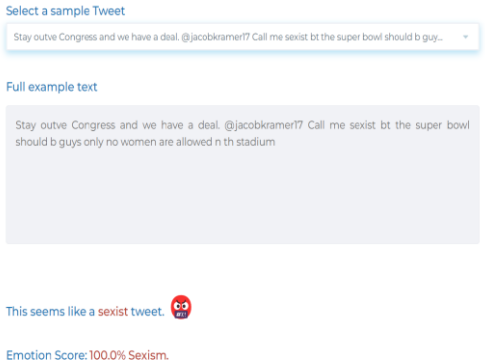


**Figura 1: Análise de *tweet* de exemplo pela ferramenta Spark NLP [32]**

Utilizando a *pipeline* (um *pipeline* de aprendizado de máquina é a construção de ponta a ponta que orquestra o fluxo de dados e a saída de um modelo de *machine learning*. Inclui entrada de dados brutos, recursos, saídas, o modelo em si, seus parâmetros e saídas de previsão) de código aberto *explain document* [32], disponível no Spark NLP, podemos inserir uma frase alvo contendo *cyberbulying* para ser analisada pela aplicação, como observado na Figura 2.

```
from sparknlp.pretrained import PretrainedPipeline
from sparknlp.base import *
```

```
pipeline = PretrainedPipeline('explain_document_md', lang = 'pt')
```

```
text = 'João é feio por ter vindo de Portugal'
result = pipeline.annotate(text)
```

**Figura 2: Importando uma *pipeline* pré-treinada no Spark NLP e inserindo uma frase alvo**

Uma vez que os dados estejam prontos para análise na aplicação, é possível realizar tarefas como tokenização, separação em diferentes partes de discurso (ou *Parts of Speech Tagging)* e indicação de Entidades Nomeadas (*Named Entity Recognition* - NERs). O processo de tokenização disponível é ilustrado na Figura 3.

```
result['token']
```

```
['João', 'é', 'feio', 'por', 'ter', 'vindo', 'de', 'Portugal']
```

**Figura 3: Separação em tokens da frase alvo**

A aplicação desconstrói as sentenças em verbetes, separando "*João é feio por ter vindo de Portugal*" nos diferentes tokens 'João', 'é', 'feio', 'por', 'ter', 'vindo', 'de' e 'Portugal'.

Ela permite os primeiros passos da realização do processo de POS - *Part of Speech Tagging* (classificar todas as sentenças do texto por categorias gramaticais e a relação de dependência entre palavras), conforme ilustrado na Figura 4.

```
result['pos']
```

```
['PROPN', 'AUX', 'ADJ', 'ADP', 'AUX', 'VERB', 'ADP', 'PROPN']
```

**Figura 4: Separação da frase alvo em POS tags**

A aplicação separa a frase alvo entre substantivos (do inglês *noun*, representado na ferramenta como *NOUN)*, substantivos próprios (do inglês, *proper noun*, representado pela sigla PROPN), entre outros pontos diferentes componentes estabelecidos ao se usar POS [33].

É destacado por Ferreira [34] que a nível de palavras, as principais *parts of speech tags* fornecem informações significativas sobre uma palavra em seu contexto. Além disso, elementos como a subjetividade, polaridade ou até mesmo a presença de

sentimentos na escrita de um determinado autor podem ser classificados pela quantidade de adjetivos presentes no texto [35]. Neste trabalho foi utilizada a Tabela 1 de Tripathi [36] para classificações de NERs.

**Tabela 1– Entidades NER e suas descrições [36]**

| Tipo de NER | Descrição | Exemplo |
|---|---|---|
| PERSON | Pessoas, reais ou fictícias | João |
| NORP | Grupos de Nacionalidades, Religiões ou entidades políticas | Partido Democrático Trabalhista |
| FAC | Construções, estradas, e outros pontos geográficos | Aeroporto de Guarulhos |
| ORG | Empresas, instituições, etc | Microsoft |
| GPE | Países, cidades, estados | França |
| LOC | Localizações não pertencentes ao GPE | Europa |
| PRODUCT | Objetos, veículos, etc. | Formula 1 |
| EVENT | Conflitos, eventos esportivos ou históricos | Olimpíadas |
| WORK_OF_ART | Títulos de livros, canções e filmes | O Senhor dos Anéis |
| LAW | Itens que constituem leis | Lei Maria da Penha |
| LANGUAGE | Linguagens nomeadas | Português Brasileiro |
| DATE | Datas ou períodos históricos | 20 de Julho de 2020 |
| TIME | Unidades de tempo menores que 24h | Dez minutos |
| PERCENT | Percentagem | Cinquenta por cento, 50% |
| MONEY | Valores monetários | Quinze centavos |
| QUANTITY | Medidas diversas | 10 quilômetros |
| ORDINAL | Posicionamento numérico | Primeiro |
| CARDINAL | Numerais não cobertos por outra NER | 2, quarenta, 3 |

Ainda dentro da aplicação, é possível assim determinar NERs existentes dentro da frase alvo anterior. Esse processo é ilustrado na Figura 5.

```
result['ner']
['B-PER', 'O', 'O', 'O', 'O', 'O', 'O', 'B-LOC']
result['entities']
['João', 'Portugal']
```

**Figura 5: Separação da frase alvo em NERs**

As duas NERs encontradas na frase alvo são "João", uma entidade de pessoa (do inglés *person)* refletida em B-PER e a entidade de local "Portugal", apontada como B-LOC.

## 5 Conclusões

Entre próximos passos e trabalhos futuros, estão o desenvolvimento e inserção do módulo de Análise Pragmática capaz de categorizar diferentes índices de agressão em sentenças da Língua Portuguesa, assim como a criação de um componente instalável capaz de monitorar o conteúdo gerado e curado por usuários em plataformas *online*.

Testes iniciais foram realizados em um corpus que contém conteúdo real gerado por usuários na internet durante um momento de emoção de um programa de TV composto de 450 *tweets* na Língua Portuguesa [37].

Munido de tal conhecimento, este trabalho buscou aplicar as técnicas descritas nesse documento para trabalhar em um software inicial na construção de um PMV (Produto Mínimo Viável) de um analisador de agressão em discurso da Língua Portuguesa. Esse resultado foi obtido por meio da combinação de mecanismos já existentes de lematização, tokenização, separação e afins combinados ao conteúdo da Análise Pragmática. Uma *pipeline* pré-treinada foi utilizada no reconhecimento e atribuição de palavras da Língua Portuguesa em diferenças sentenças.

Dessa forma, a ferramenta aborda fases de análises léxico-morfológica, sintática e semântica do Português Brasileiro. É importante disseminar esses trabalhos iniciais para que possamos dar continuidade na fase de análise pragmática.

Diante dos problemas descritos, a elaboração deste trabalho permite um passo na direção de contribuir com a sociedade e com o relacionamento entre diferentes pessoas no mundo *online*, utilizando técnicas de PLN na identificação de discurso violento e perseguição na internet.

## REFERÊNCIAS

[1] Daniel Jurafsky S. and James H. Martin. 2020. *Speech and Language Processing: an introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. (3ª. ed.). New Jersey, USA: Stanford. 615p. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

[2] Barr Avron, Paul R. Cohen and Edward A. Feigenbaum. 1989. *The Handbook of Artificial Intelligence*. Vol. IV. Massachusetts: Addison Wesley Publishing Company. 728p. https://ia800500.us.archive.org/31/items/handbookofartific04 barr/handbookofartific04barr.pdf

[3] Hugo G. Oliveira, Diogo Costa and Alexandre M. Pinto. Automatic Generation of Internet Memes from Portuguese News Headlines. 2016. *Proceedings of the 12th International Conference on the Computational Processing of the Portuguese Language (PROPOR' 16)*, LNCS, Springer, Tomar, Portugal, 340–347. DOI: https://doi.org/10.1007/978-3-319-41552-9_34

[4] Bárbara B. C. da Silva and Ivandré Paraboni. 2018. Personality Recognition from Facebook Text. *Proceedings of International Conference on the Computational Processing of the 13th Portuguese Language(PROPOR'18)*, LNCS, Springer. Canela, 107-114. DOI: https://doi.org/10.1007/ 978-3-319-99722-3_11

[5] Allisfrank dos Santos, Jorge D. Barros Junior and Heloisa A. Camargo. 2018. Annotation of a Corpus of Tweets for Sentiment Analysis. *Proceedings of International Conference on the Computational Processing of the 13th Portuguese Language (PROPOR'18),* LNCS, Springer. Canela, 294-302. DOI: https://doi.org/10.1007/978-3-319-99722-3_30

[6] ACL - *Association for Computational Linguistics*. 2022. What is computational linguistics? (Sept. 2022). https://www.aclweb.org/portal/

[7] Fiocruz. 2014. *Cyberbullying e casos de suicídio aumentam entre jovens* (fev. 2014) https://agencia.fiocruz.br/cyberbullying-e-casos-de-suic%C3%ADdio-aumentam-entre-jovens.

[8] Sheri Bauman, Russel Toomey and Jenny L. Walker. 2013. Associations among bullying, cyberbullying, and suicide in high school students. *Journal of Adolescence*, 36, 2 (Jan., 2013), 341-350. DOI: https://doi.org/10.1016/ j.adolescence. 2012.12.001

[9] Secretaria de Vigilância em Saúde – Ministério da Saúde. 2017. Suicídio: Saber, agir e prevenir. *Periódicos Eletrônicos em Psicologia*. (set., 2017), https://bvsms.saude.gov.br/bvs/publicacoes/suicidio_saber_agir_prevenir .pdf

[10] Hospital Santa Monica. 2018. *Cyberbullying e suicídio: como influenciam crianças e adolescentes?* (jul., 2018), https://hospitalsantamonica.com.br/ cyberbullying-e-suicidio-como-influenciam-criancas-e-adolescentes/

[11] Sameer Hinduja and Justin W. Patchin. 2010. Bullying, Cyberbullying, and Suicide. *Archives of Suicide Research* 14, 3 (July, 2010), 206-221. DOI: https://doi.org/10/1080/13811118.2010.494133

[12] Wendy Craig, Meyran Boniel-Nissim, Nathan King, Sophie D. Walsh, Maartje Boer, Peter D. Donelly, Yossi Harel-Fisch, Marta Malinowska-Ciéslik, Margarida G. de Matos, Alina Cosma, Regina Van den Eijnden, Alessio Vieno, Frank J. Elgar, Michal Molcho, Ylva Bjereld and William Pickett. 2010. Social Media Use and Cyber-Bullying: A Cross-National Analysis of Young People in 42 Countries. *Journal of Adolescent Health.* 66, 6 (June, 2020), Elsevier, S100-S108. DOI: https://doi.org/10/1016/j.jadohealth.2020.03.006

[13] Sofia Berne, Ann Frisén and Jesper Berne. 2019. Cyberbullying in Childhood and Adolescence: Assessment, Negative Consequences and Prevention Strategies. *Policing Schools: Scholl Violence and the Jurdification of Youth* (Lunneblad, J. (Ed.). Springer International Publishing (Sept., 2019), 141-152. DOI: https://doi.org/10.1007/978-3-030-18605-0_10

[14] Marisa Pinto. 2012. *Menina de 15 anos suicida-se por sofrer de Cyberbullying* (out., 2012), https://pplware.sapo.pt/informacao/menina-de-15-anos-suicida-se-por-sofrer-de-cyberbullying/

[15] Gustavo Laboreiro and Eugênio Oliveira. 2014. What we can learn from looking at profanity. *Prooceedings of International Conference on the 11th Computational Processing of the Portuguese Language (PROPOR'14).* In: Baptista, J., Mamede, N., Candeias, S., Paraboni, I., Pardo, T. A. S. and Nunes, M. G. V. (Org.), Springer, Springer, São Carlos. 108–113. DOI: https://doi.org/10.1007/978-3-319-09761-9_11

[16] José J. Almeida. 2019. *Dicionário de Calão e Expressões Idiomáticas* (1ª ed.). Lisboa: Guerra e Paz.

[17] Soren Wichmann. 2019. How to distinguish languages and dialects. *Computational Linguistics*, 45, 4 (Dec., 2019), 823–831. DOI: https://doi.org/10.1162/ coli_a_00366.

[18] Silvio L. Pereira. 2020. *Processamento de Linguagem Natural*. https://www.ime.usp.br/~slago/IA-pln.pdf

[19] Newton K. Miura. 2019. *Geração incremental de parsers dependentes de contexto para o português brasileiro.* 132f. Departamento de Engenharia de Computação e Sistemas Digitais. Escola Politécnica da Universidade de São Paulo. São Paulo. Tese de Doutorado.

[20] Cliff Goddard. 2012. *Semantic analysis:* a Pratical Introduction (2ª. ed.). Oxford: Oxford University Press, 512p.

[21] Daniel N. Muller. 2003. *Processamento de Linguagem Natural*. https: //www.inf.ufrgs.br/~danielnm/docs/pln.pdf

[22] Miriam L. C. S. Domingues. 2011. *Abordagem para o Desenvolvimento de um Etiquetador de Alta Acurácia para o Português do Brasil.* 154f. Departamento de Engenharia Elétrica. Universidade Estadual do Pará. Belém. Tese de Doutorado.

[23] Carolinne R. Faria. 2021. *Ferramenta Carolina para Identificação de Pragas e Doenças na Cultura da Soja utilizando Processamento de Linguagem Natural*. 87f. Departamento de Computação. Universidade Estadual de Londrina. Londrina. Dissertação de Mestrado.

[24] Peter Ljunglöf and Mats Wirén. 2010. *Handbook of Natural Language Processing*. (2ª ed.). Nitin Indurkhya and Fred J. Damerau (Ed.). Boca Raton: Chapman & Hall/CRC, p.59-91.

[25] Jalaj Thanaki. 2017. *Python Natural Language Processing: Advanced Machine Learning and Deep Learning Techniques for Natural Language Processing.* (1ª ed.). Birmigham: Packt Publishing. 798p.

[26] Cinthyan R. S. C. de Barbosa. 2004. *Técnicas de parsing para Gramática Livre de Contexto Lexicalizada da Língua Portuguesa.* 171f. Departamento de Engenharia Eletrônica e Computação. Instituto Tecnológico da Aeronáutica, São José dos Campos. Tese de Doutorado.

[27] Cinthyan R. S. C. de Barbosa. 1998. *Gramática para consultas radiológicas em Língua Portuguesa.* 143f. Instituto de Informática. Universidade Federal do Rio Grande do Sul. Porto Alegre. Dissertação de Mestrado.

[28] Fábio A. D. de Oliveira e Philippe O. A. Navaux. 2002. Processamento de Linguagem Natural: princípios básicos e a implementação de um analisador sintático de sentenças da Língua Portuguesa. *Revista de Ciência da Informação*, 1,5 (maio, 2002), 6-14.

[29] Krista M. Wilkison. 1998. Profiles of Language and Communication Skills in Autism. *Mental Retardation and Developmental Disabilities Research Review*, 4 (Jan., 1998), 73–79. DOI: https://doi.org/10.1002/(SICI)1098-2779 (1998)4:2<73::AID-MRDD3>3.0.CO;2-Y

[30] Simone A. Lopes-Herrera. 2009. O uso da linguagem no autismo de alto funcionamento e na síndrome de Asperger: uma perspectiva pragmática na intervenção fonoaudiológica. *Cadernos de Comunicação e Linguagem*, 1, 2 (2009), 87-106.

[31] Françoise Armengaud. 2006. *A Pragmática.* Tradução Marcos Marcionilo. São Paulo: Parábola Editorial.159p.

[32] John Snow Labs Inc. 2022. *Spark NLP: State of the Art Natural Language Processing*. https://nlp.johnsnowlabs.com

[33] Rami Al-Rfou. 2015. *Part of Speech Tagging.* https://polyglot.readthedocs.io/ en/latest/POS.html

[34] Renato C. B. Ferreira. 2017. *Uma Abordagem Semiautomática para Identificação de Elementos de Processo de Negócio em Texto em Linguagem Natural.* 103f. Instituto de Informática. Universidade Federal do Rio Grande do Sul. Porto Alegre. Dissertação de Mestrado.

[35] Guilherme Y. Sakurai. 2019. *Processamento de Linguagem Natural - Detecção de Fake News.* 37f. Departamento de Computação. Universidade Estadual de Londrina. Londrina. Trabalho de Conclusão de Curso.

[36] AshutoshTripathi. 2020. Named Entity Recognition NER using spaCy | NLP. *Towards Data Science*. https://towardsdatascience.com/named-entity-recognition-ner-using-spacy-nlp-part-4-28da2ece57c6

[37] José L. V. M. Mioni. 2023. *Processamento da Língua Portuguesa na Detecção de Toxicidade na rede social Twitter*. 82f. Departamento de Computação. Universidade Estadual de Londrina. Londrina. Dissertação de Mestrado.