The Construction of a Corpus for Detecting Irony and Sarcasm in Portuguese

Gabriel Schubert M.[1], **Larissa A. de Freitas** [1]

[1]CDTec - Ciência da Computação
Universidade Federal de Pelotas (UFPel)
Rua Gomes Carneiro, 1 - 96010-610 - Pelotas, RS - Brasil

gsmarten@inf.ufpel.edu.br, larissa@inf.ufpel.edu.br

Abstract. *Portuguese is a low-resource language, where a few works developed corpora for specific Natural Language Processing tasks, such as sarcasm and irony detection, sentiment analysis and others. In this work, we developed a corpus in the Portuguese language to sarcasm and irony detection task. In the future, we intend to develop a tool to recognize sarcasm and irony and we intent to use the corpus presented in this article.*

Resumo. *O português é uma língua com poucos recursos, onde poucos trabalhos desenvolvem e disponibilizam corpora para tarefas específicas de Processamento de Língua Natural, como detecção de sarcasmo e ironia, análise de sentimento e outras. Neste trabalho, nós desenvolvemos um corpus de sarcasmo e ironia para o Português. No futuro, pretendemos desenvolver uma ferramenta que reconheça sarcasmo e ironia e utilizar o corpus apresentado neste artigo.*

1. Introdução

Apesar dos avanços na área de Análise de Sentimento [Pang and Lee 2008], essa área ainda se depara com vários obstáculos. Dentre eles, destaca-se a linguagem figurada.

O primeiro passo para o processamento de textos irônicos é a detecção automática desse fenômeno linguístico. Porém, para detectar estes fenômenos precisamos de recursos, ou seja, de corpus anotado.

Tendo em vista este contexto, o objetivo geral deste trabalho é construir um corpus sobre sarcasmo e ironia anotado automaticamente.

É importante ressaltar que não é encontrado nenhum corpus proveniente de *sites* na literatura de língua portuguesa, apenas corpora provenientes da rede social Twitter como [Freitas et al. 2014] e [Silva and Bonfante 2018]

O próximo passo deste trabalho é utilizar o corpus criado em uma competição específica de detecção de sarcasmo e ironia em língua portuguesa. Além disso, pretendemos criar um sistema de detecção de sarcasmo e ironia.

Este artigo está estruturado da seguinte forma: na seção 2 é apresentado o referencial teórico; na seção 3 é apresentado os trabalhos relacionados; na seção 4 é apresentado a coleta e o processamento dos textos; na seção 5 é apresentado os resultados obtidos e, por fim, na seção 6 é apresentado as considerações finais e trabalhos futuros.

2. Referencial Teórico

Normalmente, o conceito de sarcasmo e ironia são considerados muito similares, entretanto, em uma frase irônica o sarcasmo não está presente e em uma frase sarcástica a ironia pode ser presente de uma diferente forma. [Lee and Katz 1998] menciona sarcasmo como uma forma de ofender ou ridicularizar uma pessoa em específico, já a ironia representa a expressão de algo contrário ao que realmente é verdade.

Para a automatização da coleta dos dados presentes em *sites* noticiários, será utilizada uma ferramente denominada *web crawler* (nome em português "rastreador web"). Como definido em [Cho and Garcia-Molina 1999], um *crawler* é um programa que coleta automaticamente páginas da web a partir de um URL inicial.

Em relação a seleção dos *sites* para a coleta, *sites* com notícias não sarcásticas (HuffPost Brasil[1] e Nexo Jornal[2]) e notícias sarcásticas (Sensacionalista[3] e The piaui Herald[4]) foram escolhidos, sendo eles:

- **Sensacionalista:** *site* noticiário satírico criado em 2009, como o próprio nome já diz, um *site* cujo conteúdo é relacionado ao sensacionalismo [Angrimani 1994]. Nele, utilizamos as notícias que estão na categoria de país, as quais abordam mais de cinco mil notícias sensacionalistas.
- **The piauí Herald:** *site* noticiário com notícias arquivadas datadas desde 2009. Neste caso, não há uma divisão em categorias especificada, logo são coletados todos os dados provenientes dele, que envolve em torno de 2,100 notícias sensacionalistas.
- **HuffPost Brasil:** *site* noticiário com notícias reais datadas desde o final de 2017. Mesmo sendo um *site* com conteúdo mais recente, nele se é encontrado mais de 5,000 notícias reais abordando diversos assuntos sem uma categoria específica.
- **Nexo Jornal:** *site* noticiário com notícias reais fundado em novembro de 2015. Nele, utilizamos as notícias que estão na categoria de sociedade, em torno de 5,000 notícias reais.

3. Trabalhos Relacionados

A busca pelo desenvolvimento de material relacionado ao Processamento de Linguagem Natural [Collobert et al. 2011] vem crescendo ao decorrer dos últimos anos pela comunidade acadêmica.

No caso deste artigo, o foco é na construção de um corpus anotado que permita a detecção de sarcasmo ou ironia. Sendo assim, este corpus será proveniente de textos de fontes como rede sociais e *sites* informativos.

Na academia, existem algumas competições, nas quais corpora sobre sarcasmo e ironia anotados automaticamente são de grande importância. Exemplos dessas competições são o SemEval[5] (sigla do inglês *Semantic Evaluation*), o FIRE[6] (sigla do

[1]https://www.huffpostbrasil.com/
[2]https://www.nexojornal.com.br/
[3]https://www.sensacionalista.com.br/
[4]https://piaui.folha.uol.com.br/herald/
[5]http://alt.qcri.org/semeval2020/
[6]http://fire.irsi.res.in/fire

inglês *Forum for Information Retrieval Evaluation*) e o IroSvA[7] (sigla do inglês *Irony Detection in Spanish Variants*).

3.1. Das Competições

Durante os anos de 2018 e 2019, algumas competições focadas em detecção de ironia foram realizadas, almejando tal detecção em idiomas como o Inglês, Espanhol e Árabe, tendo seus corpora descritos nas subseções abaixo.

3.1.1. Inglês - SemEval 2018

O conjunto de dados usado no SemEval 2018 [Van Hee et al. 2018] consiste de 4,792 *tweets* divididos em 2,396 irônicos e 2,396 não irônicos que foram coletados entre dezembro de 2014 e janeiro de 2015. A coleta foi baseada em *tweets* contendo as hashtags *#irony*, *#sarcasm* e *#not*. Além disso, também foi feita a adição de 1,792 *tweets* não irônicos de um segundo corpus, durante o mesmo período de tempo e com as mesmas hashtags mencionadas previamente.

3.1.2. Espanhol - IroSvA 2019

O conjunto de dados usado no IroSvA 2019 [Ortega-Bueno et al. 2019] consiste de 3,000 notícias dividindo-as entre 2,000 irônicas e 1,000 não irônicas. Além disso, estes dados tem uma variedade de nove tópicos distintos e são provenientes de três *sites* cubanos populares de notícias (Cubadabate[8], OnCuba10[9] e CubaSí[10])

3.1.3. Árabe - FIRE 2019

O conjunto de dados usado no FIRE 2019 [Ghanem et al. 2019] consiste de 5,030 *tweets* divididos entre 2,614 irônicos e 2,416 não irônicos e foram coletados entre os anos de 2011 e 2018. Assim, eles tratam de diferentes assuntos políticos e eventos relacionados à região de Magrebe, na África, e do Oriente Médio. Por conseguinte, para está coleta foram consideradas um conjunto de palavras predefinidas e referentes a algumas figuras políticas específicas, as quais estavam presentes nos assuntos a Primavera Árabe e as eleições presidenciais do Egito e dos Estados Unidos.

3.2. Corpus Proveniente de Sites

Como é mencionado em [Misra and Arora 2019], os conjuntos de dados provenientes de *tweets* são, em sua maioria, escritos com erros ortográficos e de maneira informal, principalmente por serem provenientes de uma rede social [Amir et al. 2016]. Sendo assim, essa plataforma está limitada a uma comunidade leiga, não profissional, onde os *tweets* são totalmente dependentes do contexto disponível para poder ser definido.

[7]https://www.autoritas.net/IroSvA2019/
[8]http://www.cubadebate.cu/
[9]https://oncubanews.com/
[10]http://cubasi.cu/

Logo, uma abordagem como a de [Misra and Arora 2019] acaba proporcionando vantagens por serem *sites* populares relacionados a notícias escritas por profissionais e, assim, possibilitando definir os textos mais facilmente como sarcásticos, irônicos ou textos de noticiários reais (não irônicos e não sarcásticos) que são meticulosamente independentes e disponibilizam todo o contexto necessário para compreensão em seus parágrafos.

O corpus proposto por [Misra and Arora 2019] consiste de 26,709 manchetes divididas em 11,725 irônicas e 14,984 não irônicas. Os autores coletaram os textos de dois *websites*, The Onion[11] (notícias sarcásticas) e HuffPost[12] (notícias não sarcásticas).

3.3. Nossa Abordagem

Este trabalho tem como principal a relação com o [Misra and Arora 2019], descrito acima, principalmente por abordarmos o uso de *sites* de noticiários com o foco em notícias reais e sensacionalistas [Angrimani 1994] para a construção do nosso corpus anotado.

Não houve a tentativa de utilizar corporas traduzidos do inglês para o português, porque ironia ou sarcasmo pode variar de um idioma para o outro. Também não há nenhuma rotulação manual dos corpus coletados, sendo utilizado como critério, exclusivamente, o fato de um *site* ser categorizado como de notícias reais ou não e assim ser coletado pelo *web crawler*. Considerando também que até uma coleta manual possa haver erros dependendo de como cada coletor interpreta algo como irônico ou não.

Assim, almejamos usar este corpus anotado para a criação de uma competição voltada a tarefa de detecção de sarcasmo e ironia no idioma Português, semelhante as competições SemEval 2018, IroSvA 2019 e FIRE 2019.

4. Coleta e Processamento dos Textos

Toda a coleta e processamento realizado nos dados deste trabalho será com base nos quatro *sites* apresentados anteriormente. Portanto, isso será feito a partir de um programa desenvolvido na linguagem de programação *Python 3.8* e usufruindo de módulos nativos e não nativos.

4.1. Coleta

Tendo em vista a etapa de coleta, criamos uma abordagem que possa flexibilizar a forma que os dados são coletados, podendo assim não fazer somente o uso dos *sites* base, mas tendo a possibilidade de adicionar outros *sites* ao realizar o *crawling*. Para que isso ocorra, é utilizado um arquivo de formato *JSON* (Figuras 1(a), 1(b), 2(a) e 2(b)) como o arquivo de configuração, descrito na seção 4.1.1, assim evitando a necessidade de interferir diretamente em alterações no código fonte do programa utilizado para a coleta.

4.1.1. Arquivo de Configuração

O arquivo de configuração em formato *JSON* possibilita a troca de atributos e variáveis utilizados no programa de forma simples e rápida.

[11]https://www.theonion.com/
[12]https://www.huffpost.com/

```
"sensacionalista": [
    {
        "url": "https://www.sensacionalista.com.br/pais/page/",
        "sarcasm": true
    },
    {
        "html_class": "last",
        "regex": "title=\"[^\"]*",
        "remove": [7,0]
    },
    {
        "raw_args":
        {
            "html_class": "td_module_8 td_module_wrap",
            "regex": "<a[^<]*</a>"
        },
        "regex": ["href=\\\"[^\\\"]*", "title=\\\"[^\\\"]*"],
        "remove": [[6,0], [8,0]],
        "html_options": ["div", "p402_premium"]
    }
]
```

(a) Site Sensacionalista

```
"piauiherald": [
    {
        "url": "https://piaui.folha.uol.com.br/herald/",
        "sarcasm": true,
        "as_archived": true
    },
    {
        "html_class": "tab-btn",
        "regex": "data-tab=\"arquivo_\\d{4}\">",
        "remove": [18,2],
        "element": "li",
        "shorter": 7
    },
    {
        "raw_args":
        {
            "html_class": "bloco size-2",
            "regex": "<a href=[^>]*>\\s<h2[^<]*"
        },
        "regex": ["<a href=\\\"[^\\\"]*\\\">", "<h2 class=\\\"bloco-title\\\">\\s*.*"],
        "remove": [[9,2], [26,0]],
        "html_options": ["div", "post-inner"]
    }
]
```

(b) The piuaí Herald

Figura 1. Configuração dos sites sarcásticos para a coleta do corpus

```
"huffpostbrasil": [
    {
        "url": "https://www.huffpostbrasil.com/noticias/",
        "sarcasm": false
    },
    {
        "html_class": "pagination__link",
        "regex": "href=\\\"/noticias/\\d*/\\\"",
        "remove": [16,2]
    },
    {
        "raw_args":
        {
            "html_class": "apage-rail-cards",
            "regex": "<a class=\\\"[^<]*</a>"
        },
        "regex": ["href=\\\"[^\\\"]*\\\"", "target=\\\"_self\\\">[^<]*<"],
        "remove": [[7,1],[15,0]],
        "html_options": ["div", "post-contents yr-entry-text"],
        "url_prefix": 31
    }
]
```

(a) Site HuffPost Brasil

```
"nexojornal": [
    {
        "url": "https://www.nexojornal.com.br/tema/Sociedade?pagina=",
        "sarcasm": false
    },
    {
        "html_class": "Pagination__link___1VkYg",
        "regex": ">\\d{3}</a>",
        "remove": [1,4]
    },
    {
        "raw_args":
        {
            "html_class": "Teaser__title-dark___1HEzZ",
            "regex": "<a alt=\\\"[^>]*>",
            "element": "h4"
        },
        "regex": ["href=\\\"[^\\\"]*\\\"", "title=\\\"[^\\\"]*\\\">"],
        "remove": [[6,1], [7,2]],
        "html_options": ["div", "Default__text-area___38Dm5"],
        "url_prefix": 30
    }
]
```

(b) Nexo Jornal

Figura 2. Configuração dos sites não sarcásticos para a coleta do corpus

Neste arquivo, há diferentes valores no formato atributo-valor referente a um ou mais *sites*. Desta forma, tal arquivo deverá começar com uma lista contendo um ou mais atributos referente ao nome do *site*.

Dentro de cada atributo referente ao nome do *site* há três outras lista, sendo a primeira referente a dados base do *site* (URL, se é sarcástico ou não), a segunda há variáveis necessárias para a coleta da quantidade total de páginas a serem requisitadas e a terceira contendo expressões regulares e nomes de elementos do HTML requisitado anteriormente.

4.1.2. Execução da Coleta

Utilizando o pacote para *Python "Requests: HTTP for Humans*" e tendo os dados do arquivo de configuração, podem ser feitas as requisições HTTP para cada URL a ser obtido. Assim, primeiro se é feita a requisição no URL base e encontrado o número máximo de páginas ou o número de anos contendo notícias arquivadas.

Consequentemente, com a obtenção deste dado, é acrescentado ao final do URL base tais números, criando, assim, uma lista de URL para que ocorra a requisição em cada um, além do armazenamento em uma lista. Então, com a requisição de cada página armazenada, podemos ter acesso aos dados brutos do HTML de cada página e realizar o seu refinamento.

4.2. Processamento dos Dados Coletados

Para processar estes dados e adquirir somente o necessário, que neste primeiro caso seria o título da notícia e o URL referente a mesma, utilizaremos outro pacote *Python* chamado "*Beautiful Soup*". Com isto, as classes referentes aos elementos e o HTML obtido com as requisições, podemos escolher especificamente quais partes queremos extrair, excluindo assim divisões contendo abas de menu, anúncios, textos de outras seções, entre outros.

Com os dados brutos filtrados, será utilizado o módulo nativo do *Python* para operações com expressões regulares, mesmo que haja uma grande quantidade de conteúdo desnecessário eliminado, visto que ainda haverá a necessidade de selecionar somente o texto do título e o URL fazendo referência a página do título, excluindo, assim, elementos do HTML que não necessitam serem arquivados.

Desta parte adiante, já terá sido adquirido o necessário para a construção de um corpus anotado contendo informações referentes se os textos são sarcásticos ou não para cada conjunto de título e link.

5. Resultados Obtidos

Levando em consideração os quatro *sites* predefinidos para este levantamento de dados, obteve-se um total de 18,003 notícias, divididas em 7229 classificadas como irônicas e 10774 não irônicas, coletadas no dia 3 de Julho de 2020. Em síntese, é apresentado abaixo uma Tabela 1 mostrando a diferença entre a nossa abordagem em relação as demais descritas anteriormente.

De acordo com [Sardinha 2000] o corpus coletado é classificado como médio-grande, contendo um total de 1,000,010 palavras, sendo dividido em 663,775 palavras

Corpus/Estatística	Irônicas	Não Irônicas
SemEval 2018	2396	2396
IroSvA 2019	2000	1000
FIRE 2019	2614	2416
[Misra and Arora 2019]	11724	14985
Nossa Abordagem	7229	10774

Tabela 1. Estatística Geral dos Corpora.

coletadas das notícias irônicas e 336,235 palavras coletadas das notícias não irônicas. Sendo assim, nas Figuras 3(a) e 3(b) são apresentadas as nuvens de palavras dos títulos sarcásticos e não sarcásticos, respectivamente.

(a) Títulos sarcásticos (b) Títulos não sarcásticos

Figura 3. Nuvem de palavras

6. Considerações Finais e Trabalhos Futuros

Pela falta de corpus anotado provenientes de *sites* focados na tarefa de detecção de sarcasmo ou ironia em língua portuguesa até o momento, há de se enfatizar que a criação de mais recurso voltado para este idioma é de extrema importância.

Com a divulgação desse tipo de recurso, novos estudos relacionados a área no âmbito acadêmico Brasileiro podem surgir e consequentemente uma maior produção nacional.

O código fonte, juntamente com o dataset gerado, utilizado para a criação automatizada deste corpus, encontra-se disponível no repositório `https://github.com/schuberty/PLNCrawler`.

Como mencionado anteriormente, almejamos usar este corpus anotado para a criação de uma competição voltada a tarefa de detecção de sarcasmo e ironia no idioma Português, já que no momento não há nenhum outro do tipo em processo de desenvolvimento para o público nacional.

Por conseguinte, pretendemos criar um detector de sarcasmo e ironia, assim como as abordagens: [Misra and Arora 2019], [Freitas et al. 2014], e

[Silva and Bonfante 2018], e aplicá-lo no corpus anotado provenientes dos quatro *sites*: Sensacionalista, The piaui Herald, HuffPost Brasil e Nexo Jornal.

Referências

Amir, S., Wallace, B. C., Lyu, H., and Silva, P. C. M. J. (2016). Modelling context with user embeddings for sarcasm detection in social media. *arXiv preprint arXiv:1607.00976*.

Angrimani, D. (1994). *Espreme que sai sangue: um estudo do sensacionalismo na imprensa*, volume 47. Summus Editorial.

Cho, J. and Garcia-Molina, H. (1999). The evolution of the web and implications for an incremental crawler. Technical report, Stanford.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.

Freitas, L. A., Vanin, A., Hogetop, D., Bochernitsan, M., and Vieira, R. (2014). Pathways for irony detection in tweets. In *29th Symposium on Applied Computing*, pages 628–633.

Ghanem, B., Karoui, J., Benamara, F., Moriceau, V., and Rosso, P. (2019). Idat at fire2019: Overview of the track on irony detection in arabic tweets. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, pages 10–13.

Lee, C. J. and Katz, A. N. (1998). The differential role of ridicule in sarcasm and irony. *Metaphor and symbol*, 13(1):1–15.

Misra, R. and Arora, P. (2019). Sarcasm detection using hybrid neural network. *arXiv preprint arXiv:1908.07414*.

Ortega-Bueno, R., Rangel, F., Hernández Farıas, D., Rosso, P., Montes-y Gómez, M., and Medina Pagola, J. E. (2019). Overview of the task on irony detection in spanish variants. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR-WS. org*.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135.

Sardinha, T. B. (2000). Lingüística de corpus: histórico e problemática. *Delta: documentação de estudos em lingüística teórica e aplicada*, 16(2):323–367.

Silva, F. R. A. and Bonfante, A. G. (2018). Detecção de ironia e sarcasmo em língua portuguesa: uma abordagem utilizando deep learning. Monografia (Bacharel em Ciência da Computação), UFMG (Universidade Federal do Mato Grosso), Brasil.

Van Hee, C., Lefever, E., and Hoste, V. (2018). Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.