Complexidade textual em notícias satíricas: uma análise para o português do Brasil

Gabriela Wick-Pedro[1], **Roney L. S. Santos**[2]

[1]Universidade Federal de São Carlos (UFSCar), São Carlos, Brasil

[2]Universidade de São Paulo (USP), São Carlos, Brasil

gwpedro@estudante.ufscar.br, roneysantos@usp.br

Abstract. *This article presents an analysis of the textual complexity of satirical and true news for Brazilian Portuguese. The Fake has been a big problem nowadays. Satirical content is an important point in the automatic detection of false news as its use can cause underlying confusion in the analysis. To carry out this research, the NILC-Metrix tool was applied, and 16 measures were evaluated, including descriptive, syntactic and semantic aspects, noting a greater complexity for real texts.*

Resumo. *Neste artigo é apresentada uma análise da complexidade textual de notícias satíricas e verdadeiras para o português do Brasil. As chamadas Fake News – ou notícias falsas – têm sido um grande problema na atualidade. O conteúdo satírico é um ponto importante na detecção automática de notícias falsas, pois seu uso pode causar confusão subjacente na análise. Para realização desta pesquisa, foi aplicada a ferramenta NILC-Metrix e avaliadas 16 medidas, entre aspectos descritivos, sintáticos e semânticos, notando-se uma maior complexidade para os textos verdadeiros.*

1. Introdução

Historicamente, a capacidade de divulgação das notícias falsas está intrinsecamente ligada aos suportes de cada época, como os papiros e pergaminhos na Antiguidade, a criação da imprensa na década de 1430 e surgimento do rádio e da TV nos séculos XIX e XX, respectivamente. Na atualidade, com a Internet, a velocidade da comunicação cresceu e acelerou o tempo de disseminação das notícias, atingindo o maior número de pessoas em todo lugar do planeta em um curto espaço de tempo. As redes sociais como novos meios de comunicação permitiram um maior espaço para o indivíduo expor seus pensamentos, opiniões, emoções e posições de suas ideias, o que fortaleceu um comportamento individualista. Assim, o ambiente digital se tornou, com o passar dos anos, o habitat ideal para a disseminação da desinformação, pois é a partir das atuais tecnologias que uma notícia falsa é desenvolvida e compartilhada de uma forma sistêmica. O excesso de informação veiculada nas redes influenciou a sociedade atual, impactando diretamente em sua forma de pensar e agir no mundo contemporâneo.

Esse dinamismo das mídias sociais possibilita uma maior rapidez de leitura, o que resulta em baixo aprofundamento de grande parte dos textos veiculados na rede e desqualifica o leitor para uma compreensão de textos mais complexos. Logo, para realizar o processo de leitura, o leitor precisa compreender totalmente o texto e ter o conhecimento

prévio para obter integralmente o seu sentido [Leffa 1996]. Tal cenário colabora para minimizar a capacidade de entender argumentos mais difíceis da linguagem, de fazer uma análise crítica do que está sendo lido, além de favorecer na ascensão da disseminação de notícias falsas.

A heterogeneidade do termo e as diversas definições para o conceito de notícias falsas, ou popularmente, *Fake News*, transpassa as limitações conceituais, pois uma notícia pode ser projetada intencionalmente para enganar o leitor, ser criada para atrair cliques e obter lucro ou ser notícias satíricas com o objetivo de entreter [Rubin et al. 2015, Wardle and Derakhshan 2017, Tandoc Jr et al. 2018]. Grosso modo, Fake News pode ser definido como notícias imprecisas e, muitas vezes, fabricadas intencionalmente [Quandt et al. 2019].

Notam-se esforços acadêmicos recentes que procuram estudar conteúdo enganoso (do inglês, "*deception*"), averiguar o comportamento e o perfil dos usuários que compartilham e produzem esse tipo de notícia e como elas se espalham pela rede. Em particular, muitas frentes vêm sendo exploradas pelo Processamento de Língua Natural (PLN). Vale citar, por exemplo, o esforço para a língua portuguesa de [Monteiro et al. 2018] e [Silva et al. 2020], que fazem uso de características linguísticas para identificar automaticamente as *Fake News*.

Dentro desse contexto, as notícias satíricas podem criar dificuldades de entendimento e falsas crenças em leitores mais desatentos [Rubin et al. 2016]. Para a Literatura, a sátira é a representação literária de um estilo escrito em verso ou prosa que tem como objetivo a crítica às instituições, à sociedade e aos hábitos culturais de um povo. A sátira é considerada um gênero literário focado na crítica de um determinado tema, utilizando-se da ironia, do sarcasmo e da paródia para apontar falhas morais, políticas e sociais [Kreuz and Roberts 1993, Simpson 2003, Attardo 2014]. Para além da censura, a obra satírica acarreta entretenimento e leva o público ao riso por meio do absurdo, do exagero ou do ridículo.

Portanto, mostra-se relevante a tarefa de descrição e detecção automática de notícias satíricas. O presente trabalho propõe-se a investigar, com base em dados levantados por ferramentas de análise textual, especificamente o NILC-Metrix, como se dá a relação de inteligibilidade em notícias satíricas e verdadeiras para o português do Brasil, a fim de verificar as principais divergências linguísticas entre o conteúdo analisado, como aspectos lexicais, sintáticos e semânticos.

2. Métodos e Materiais

2.1. NILC-Metrix

O NILC-Metrix [Leal 2021] é um sistema computacional que contém por volta de 200 métricas propostas em estudos de discurso, psicolinguística, linguística cognitiva e computacional, que tem o objetivo de analisar a complexidade textual para o português.

O NILC-Metrix[1] está dividido em 14 categorias, que vão desde informações morfossintáticas e frequência de palavras até medidas mais robustas, como medidas psicolinguísticas e de legibilidade e facilidade de leitura do texto. A documentação do sistema,

[1]Disponível em http://fw.nilc.icmc.usp.br:23380/nilcmetrix

bem como as explicações de todas as métricas estão disponíveis online[2].

Segundo o autor, o NILC-Metrix pode ajudar os pesquisadores a investigar: (i) como as características do texto se correlacionam com a compreensão da leitura; (ii) quais são as características mais desafiadoras de um determinado texto, ou seja, quais características tornam um texto ou corpus mais complexo; (iii) quais textos têm as características mais adequadas para desenvolver as habilidades dos alunos-alvo; e (iv) quais partes de um texto são desproporcionalmente complexas e devem ser simplificadas para atender a um determinado público. Todos os pontos citados acima são úteis para o trabalho neste artigo, justificando a escolha do sistema para a análise dos textos satíricos e verdadeiros, os quais são explicados na seção a seguir.

2.2. O Corpus

O corpus desta pesquisa é composto por 300 notícias do domínio político, sendo 150 notícias satíricas e 150 notícias verdadeiras. As notícias satíricas foram extraídas automaticamente do site *Sensacionalista*[3], noticiário eletrônico que brinca com vários tópicos da política ou do entretenimento brasileiro. Já para as notícias verdadeiras, a coleta foi feita manualmente: primeiro palavras-chave foram identificadas e depois uma busca manual por cada notícia verdadeira equivalente à satírica. As características detalhadas podem ser observadas na Tabela 1.

NOTÍCIAS	QTD. NOTÍCIAS	TOKENS	TYPES	SENTENÇAS
Verdadeiras	150	107.133	11.304	5.721
Satíricas	150	22.963	4.843	1.212
TOTAL	**300**	**130.096**	**16.147**	**6.933**

Tabela 1. Características do corpus

Optou-se por não balancear o corpus para evitar a perda de informações na análise, uma vez que o número de palavras, sentenças ou diversidade lexical pode ser uma característica para a descrição desse tipo de conteúdo.

3. Resultados Obtidos

3.1. Índice Flesch

Considera-se o texto como um resultado parcial da comunicação do leitor com processos cognitivos, contextuais e linguísticos [Koch 1995]. Do inglês "*readability*", a leiturabilidade de um texto tem a finalidade de calcular o nível de facilidade de leitura do leitor. Nesse sentido, entende-se que o tamanho das sentenças e o vocabulário do leitor aumentam (ou diminuem) a capacidade de leitura de um texto [DuBay 2004].

Apesar de muitos estudos considerarem leiturabilidade e legibilidade como sinônimos, aqui, entende-se os dois conceitos de forma distinta. De acordo com [Resende and de Souza 2011], o termo leiturabilidade refere-se ao que está inserido no ato de ler, considerando o papel, a habilidade, as características, os conhecimentos e a experiência do leitor na atividade de leitura. Já legibilidade corresponde aos "elementos

[2]Disponível em http://fw.nilc.icmc.usp.br:23380/metrixdoc

[3]Disponível em https://blogs.oglobo.globo.com/sensacionalista

e recursos que o próprio texto, em sua materialidade, oferece ao leitor" [Resende e Souza 2011], i.e., relaciona-se com a facilidade se reconhecimento da forma das letras.

O Índice Flesch [Flesch 1979] procura uma correlação entre tamanho da sentença e o tamanho da palavra. A fórmula de Flesch, adaptada para o português por [Martins et al. 1996], é mostrada na Equação a seguir:

$$248,835 - 1.015\left(\frac{palavras}{sentencas}\right) - 84,6\left(\frac{silabas}{palavras}\right) \quad (1)$$

A Figura 1 apresenta a estatística da leiturabilidade das notícias satíricas e verdadeiras de acordo com o Índice Flesch. De acordo com a descrição da Tabela 2 e as informações apresentadas na terceira coluna do gráfico, as notícias verdadeiras são mais difíceis (87) em relação às satíricas (55). Ainda, em comparação, a primeira coluna do gráfico mostra que as notícias satíricas (4) são muito mais fáceis do que as notícias verdadeiras (1). A segunda coluna indica uma facilidade de leiturabilidade muito maior nos textos satíricos (90) em contraste com os textos verídicos (61).

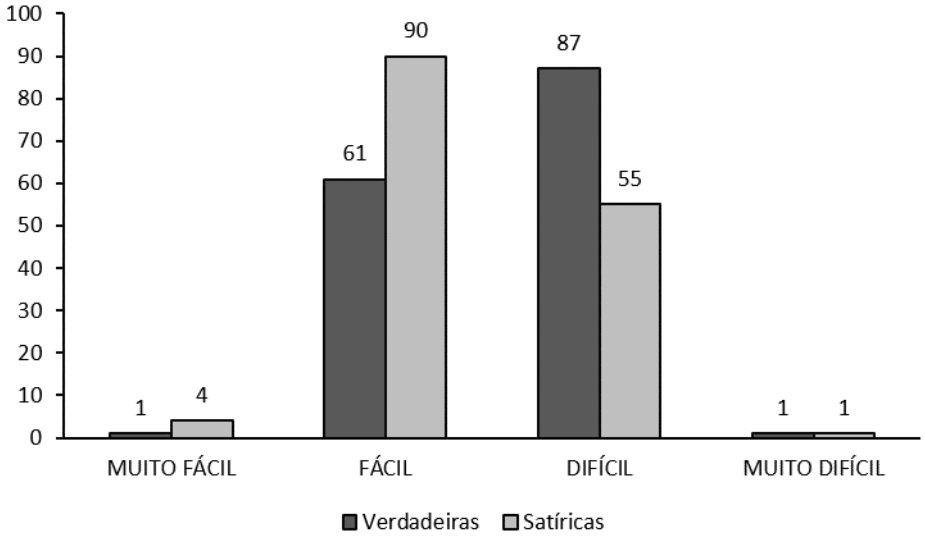

Figura 1. Estatística de leiturabilidade do corpus de acordo com o Índice Flesch Brasileiro

ESCORE	NÍVEL DE COMPLEXIDADE	GRAU ESCOLAR
100-75	Muito Fácil	1º a 5º ano
75-50	Fácil	6º a 9º ano
50-25	Difícil	Ensino Médio
25-00	Muito Difícil	Ensino Superior

Tabela 2. Escore do Índice Flesch Brasileiro: Nível de leiturabilidade por grau escolar

3.2. Avaliação da Complexidade Textual em Português do Brasil

Em geral, foram investigadas 200 medidas do NILC-Metrix, agrupadas em 14 categorias: *medidas descritivas, simplicidade textual, coesão referencial, coesão semântica, medidas psicolinguísticas, diversidade textual, conectivos, conectivos, léxico temporal, complexidade sintática, densidade de padrões sintáticos, informações morfossintáticas de palavras, informações semânticas de palavras, frequência de palavras e índices de leiturabilidade*. Entretanto, para a análise, foram selecionadas apenas as medidas com resultados mais relevantes, tais quais as medidas que tiveram valores distantes, bem como medidas importantes que são utilizadas em comparações de complexidade de texto, como o TTR [Templin 1957]. A Tabela 3 apresenta as médias de cada tipo de notícia quando aplicadas as medidas do NILC-Metrix.

CATEGORIA	*MÉTRICAS*	*VERDADEIRAS*	*SATÍRICAS*
MEDIDAS DESCRITIVAS	**Frequência de palavras de conteúdo**	611.825,41	490.304,19
	Número de palavras	594,08	156,44
	Número de sentenças	32,02	8,36
	Palavras por sentença	19,35	19,49
COESÃO REFERÊNCIAL	**Pronome anafórico do caso reto**	0,27	0,19
	Pronome demonstrativo anafórico	0,30	0,21
	Referência anafórica	1,55	1,05
	Referência anafórica (adjacente)	0,33	0,24
	Sobreposição de argumentos	0,70	0,97
	Sobreposição de argumentos (adjacentes)	0,84	1,06
	Sobreposição de radical de palavras	0,93	1,38
	Sobreposição de radical de palavras (adjacentes)	1,14	1,53
DIVERSIDADE LEXICAL	**TTR**	0,73	0,73
COMPLEXIDADE SINTÁTICA	**Incidência de orações subordinadas**	0,18	0,30
	Preposições por sentença	1,53	2,16
INFORMAÇÕES MORFOSSINTÁTICAS DE PALAVRAS	**Pronomes em 1ª Pessoa**	0,12	0,11
	Pronomes em 3ª Pessoa	0,71	0,49
	Verbos flexionados	0,25	0,38
	Verbos não flexionados	0,15	0,25
INFORMAÇÕES SEMÂNTICAS DE PALAVRAS	**Ambiguidade adjetival**	2,57	3,18
	Ambiguidade preposicional	0,28	0,49
	Ambiguidade verbal	9,99	10,22

Tabela 3. Características do corpus

Como esperado, os índices de medidas descritivas (*frequência de palavras de conteúdo, números de palavras, números de sentenças*) têm valores mais elevados em notícias verdadeiras, pois geralmente são textos mais longos e, consequentemente, mais complexos. Além disso, a ocorrência de pronomes em 3ª pessoa também é mais expressiva em notícias verdadeiras, o que pode indicar uma impessoalidade maior em textos reais em comparação às notícias satíricas.

Em relação aos mecanismos coesivos de referenciação textual, as medidas de referência anafórica – *pronome anafórico do caso reto, pronome demonstrativo anafórico, referência anafórica e referência anafórica (adjacente)* – mostram-se mais presentes em textos verdadeiros. Por se tratar de um recurso coesivo que busca a manutenção de sentidos apresentados anteriormente, quanto maior a métrica, maior a complexidade textual.

As medidas de *sobreposição de argumentos, sobreposição de argumentos (adjacentes), sobreposição de radical de palavras e sobreposição de radical de palavras (adjacentes)* possuem maior valor nas notícias satíricas. Entretanto, a repetição referencial

é uma característica de simplificação textual. Assim, quanto maior for o índice, menos complexo será o texto.

Por fim, nota-se que a ambiguidade lexical (*ambiguidade adjetival, ambiguidade preposicional e ambiguidade verbal*) é mais evidente em notícias satíricas. a ambiguidade é uma característica da construção de sentido da sátira e da paródia, uma vez que seu uso intencional causa no leitor uma confusão de sentido, podendo gerar um efeito de humor no texto. Contudo, para a compreensão do sentido ambíguo, é necessário o conhecimento extralinguístico daquilo que se lê e, dessa forma, quanto mais sentidos um texto tiver, maior será o esforço requerido do leitor para a desambiguação.

4. Conclusões

Em resumo, este artigo apresentou uma breve análise sobre a complexidade textual de notícias verdadeiras e satíricas para o português do Brasil por meio de métricas da ferramenta NILC-Metrix. Como já foi abordado neste artigo, sabe-se que compreensão de um conteúdo satírico está intrinsecamente ligada à dispositivos extralinguísticos e ao conhecimento de mundo do leitor. No entanto, aspectos linguísticos presentes na estrutura das notícias podem ser um indicativo da sátira na notícia, como a forte presença de ambiguidade lexical ser um elemento para a construção do humor.

Finalmente, como há poucas evidências de trabalhos que abordam a descrição de notícias satíricas, sobretudo para o português brasileiro, os índices aqui apresentados, mostram-se úteis não só para a descrição de notícias satíricas, como para a detecção automática de conteúdo enganoso. Além disso, os dados aqui discutidos podem ser usados futuramente como features linguísticas na implementação de classificadores automáticos de complexidade de textos jornalísticos utilizando algoritmos de aprendizado de máquina.

Agradecimentos

Os autores agradecem à CAPES (Código Financeiro 001), ao Escritório de Pesquisas da USP (PRP 668) e ao Centro de Inteligência Artificial (C4AI) da Universidade de São Paulo, apoiado pela IBM e FAPESP (nº 2019/07665-4)

Referências

Attardo, S. (2014). *Encyclopedia of humor studies*. Sage Publications.

DuBay, W. H. (2004). The principles of readability. *Online Submission*.

Flesch, R. (1979). How to write plain english: A book for consumers and lawyers.

Koch, I. (1995). O texto: construção de sentidos. *Organon*, 9(23).

Kreuz, R. J. and Roberts, R. M. (1993). On satire and parody: The importance of being ironic. *Metaphor and Symbol*, 8(2):97–109.

Leal, S. E. (2021). *Predição da complexidade sentencial do português brasileiro escrito, usando métricas linguísticas, psicolinguísticas e de rastreamento ocular*. PhD thesis, Universidade de São Paulo.

Leffa, V. J. (1996). *Aspectos da leitura*, volume 7. Sagra Porto Alegre.

Martins, T. B., Ghiraldelo, C. M., Nunes, M. d. G. V., and de Oliveira Junior, O. N. (1996). *Readability formulas applied to textbooks in brazilian portuguese*. ICMC-USP.

Monteiro, R. A., Santos, R. L. S., Pardo, T. A. S., de Almeida, T. A., Ruiz, E. E. S., and Vale, O. A. (2018). Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *Computational Processing of the Portuguese Language*, pages 324–334.

Quandt, T., Frischlich, L., Boberg, S., and Schatto-Eckrodt, T. (2019). *Fake News*, pages 1–6. American Cancer Society.

Resende, N. R. and de Souza, A. C. (2011). A atividade tradutória e a relevância da leitura: legibilidade e leiturabilidade de textos humorísticos traduzidos. *Revista Gatilho*, 13.

Rubin, V. L., Chen, Y., and Conroy, N. K. (2015). Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Rubin, V. L., Conroy, N., Chen, Y., and Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17.

Silva, R. M., Santos, R. L., Almeida, T. A., and Pardo, T. A. (2020). Towards automatically filtering fake news in portuguese. *Expert Systems with Applications*, 146:113–199.

Simpson, P. (2003). *On the discourse of satire: Towards a stylistic model of satirical humour*, volume 2. John Benjamins Publishing.

Tandoc Jr, E. C., Lim, Z. W., and Ling, R. (2018). Defining "fake news" a typology of scholarly definitions. *Digital journalism*, 6(2):137–153.

Templin, M. C. (1957). Certain language skills in children; their development and interrelationships.

Wardle, C. and Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe*, 27.