Utilizando Pistas Linguística para Detectar Conteúdo Enganoso em Português

Rodrigo F. Rodrigues [1], **Larissa A. de Freitas** [2]

[1]Centro de Desenvolvimento Tecnológico (CDTec)
Universidade Federal de Pelotas (UFPel) – Pelotas, RS – Brasil

{rfrodrigues,larissa}@inf.ufpel.edu.br

Abstract. *Greater access to internet-connected cell phones and the popularization of social networks have led to a significant increase in the generation and sharing of false news. Studies available in the literature, based on linguistic clues, retrieving authors of misleading content exhibits different verbal and nonverbal behavior than authors of true content. Thus, this article presents the LC-Tool tool, which extracts 29 linguistic clues from texts. Still, we tested the tool in three corpus about deceptive content available on the Internet. Finally, we realized that some linguistic clues could be extensive for the Portuguese language (e.g.: avg number of verbs and avg pausality). In other linguistic clues, they need to be validated, as they are affected by the context and domain of the messages.*

Resumo. *O maior acesso a celulares conectados à internet e a popularização das redes sociais levaram a um aumento significativo na geração e no compartilhamento de notícias falsas. Estudos disponíveis na literatura, baseados em pistas linguísticas, sugerem que os autores de conteúdo enganoso exibem comportamento verbal e não verbal diferente dos autores de conteúdo verdadeiro. Desta forma, neste artigo apresentamos a ferramenta LC-Tool, a qual extrai 29 pistas linguísticas de textos. Ainda, testamos a ferramenta em três corpus sobre conteúdo enganoso disponíveis na Internet. Por fim, percebemos que algumas pistas linguísticas podem ser extensíveis para o idioma português (por exemplo: média do número de verbos e média de pausalidade) e que em outras pistas linguísticas precisam ser validadas, pois são afetadas pelo contexto e domínio das mensagens.*

1. Introdução

O volume de informações geradas a cada minuto é enorme, nas redes sociais online (RSO), como Facebook, Twitter e Whatsapp. Uma vez que, facilitam o compartilhamento rápido de informações (Zhou and Zhang 2008). Desta maneira, surge um grande problema, a verificação da veracidade dos conteúdos compartilhados. Assim sendo, destaca-se uma área do Processamento de Língua Natural (PLN), chamada detecção de conteúdo enganoso, que pode ser realizada através do uso de pistas linguísticas.

As pistas linguísticas podem ser verbais ou não verbais (Zhou and Zhang 2008). Em que o primeiro enfoca como o engano é transmitido em uma linguagem natural e o segundo sobre o que é transmitido.

Muitos dos trabalhos de detecção de conteúdo enganoso disponíveis na literatura estão restritos ao idioma em Inglês (Zhou et al. 2003; Zhou et al. 2004; Zhou and Zhang 2008) e outros idiomas, como o Chinês (Zhou and Sung 2008) e o Russo (Litvinova et al. 2017). Para o idioma Português, existe uma escassez de conjuntos de dados rotulados sobre conteúdo enganoso. Posto isto, o presente artigo apresenta a implementação de uma ferramenta que extrai pistas linguísticas de três corpus rotulado sobre conteúdo enganoso para o idioma Português do Brasil.

O restante deste artigo está organizado da seguinte forma. Na seção 2, revisamos brevemente os trabalhos relacionados. Na seção 3, introduzimos a ferramenta proposta. Na seção 4, apresentamos a metodologia. Ne seção 5, apresentamos os resultados. Por último, na seção 6, concluímos este artigo.

2. Trabalhos Relacionados

O trabalho de (Zhou and Zhang 2008), resume os principais comportamentos de autores de conteúdo enganosos de acordo com construtores linguísticos.

Em (Fuller et al. 2006), as pistas linguísticas disponíveis nas ferramentas A99A e LIWC são avaliadas. Ainda, neste trabalho, os autores verificam que algumas pistas línguísticas, tais como: #p13, #p8, #p9 e #p14 têm diferenças significativas em mensagens sobre conteúdo enganoso em comparação com mensagens sobre conteúdo verdadeiro.

Em (Zhou et al. 2003), as pistas linguísticas são classificadas em 8 construtores linguísticos. São eles: quantidade, complexidade, não imediatismo, expressividade, afeto, especificidade, diversidade e informalidade.

Tabela 1. Resumo das pistas linguísticas.

Construtor	Pista linguística	Comportamento esperado em conteúdo enganoso	Trabalhos relacionados
Quantidade	Avg number of sentences per text - #p1 Avg number of verbs - #p2	+ +	(Zhou and Zhang 2008) (Fuller et al. 2006) (Zhou et al. 2003)
Complexidade	Avg size of words - #p3 Avg pausality - #p4 Avg number of sentences (in words) - #p5	- - -	(Zhou and Zhang 2008) (Fuller et al. 2006) (Zhou et al. 2003)
Não imediatismo	Avg number of modal verbs - #p6 Avg self reference (1st person pronoun) - #p7 Avg group reference (2nd person pronoun) - #p8 Avg another reference per text (3rd person pronoun) - #p9	+ - + +	(Zhou and Zhang 2008) (Fuller et al. 2006) (Zhou et al. 2003)
Expressividade	Avg emotiveness - #p10	+	(Zhou and Zhang 2008) (Fuller et al. 2006) (Zhou et al. 2003)
Afeto	Positive affect - #p11 Negative affect - #p12	+ +	(Zhou and Zhang 2008) (Fuller et al. 2006) (Zhou et al. 2003)
Especificidade	Avg spatiotemporal words - #p13	-	(Zhou and Zhang 2008) (Fuller et al. 2006) (Zhou et al. 2003)
Diversidade	Avg lexical diversity - #p14 Avg number of types per text - #p15	- -	(Zhou and Zhang 2008) (Fuller et al. 2006) (Zhou et al. 2003)
Informalidade	Avg misspelled words - #p16	+	(Zhou and Zhang 2008) (Fuller et al. 2006) (Zhou et al. 2003)
Punctuation Cue	Average number of exclamation marks #p17	+	(Fernandez and Devaraj 2019)

A Tabela 1 contextualiza os construtores linguísticos nos trabalhos relaciona-

dos. O símbolo (+) indica que o comportamento da pista linguística é mais atenuado em conteúdo enganoso, enquanto o símbolo (-) indica que o comportamento da pista linguística é menos atenuado em conteúdo enganoso.

3. Ferramenta Proposta

A ferramenta LC-Tool[1] implementa um total de 29 pistas linguísticas, que podem ser aplicadas a corpus sobre conteúdo enganoso.

Na primeira etapa, é obtido o conjunto de dados a serem processados. Em seguida, na segunda etapa, é realizado o cálculo das metainformações (tokens, caracteres, sinais de pontuação e outros). Por fim, na terceira etapa, são calculadas as pistas linguísticas.

Para o desenvolvimento da ferramenta, usamos: (i) as tags *Universal POS tags*[2], com o propósito de identificar a marcação de partes da fala; (ii) o LeIA (Léxico para Inferência Adaptada)[3], com a intenção de realizar analise de sentimento no domínio de sentença; (iii) freeoffice pt-BR[4] em conjunto com o spaCy[5], com o objetivo de encontrar erros ortográficos.

4. Metodologia

Para realização deste trabalho, o primeiro passo foi buscar conjuntos de dados rotulados com conteúdo enganoso para a língua portuguesa do Brasil. Em segundo lugar, era necessário um estudo sobre detecção de conteúdo enganoso com base em abordagens que fazem uso de pistas linguísticas. Em terceiro lugar, implementamos uma ferramenta que extrai pistas linguísticas de corpus rotulados com conteúdo enganoso, usando a linguagem de programação Python. Por fim, cada conjunto de dados foi utilizado na ferramenta para que as pistas linguísticas fossem calculadas. Assim, é possível avaliar os resultados indicados pelas pistas linguísticas em cada conjunto de dados.

Em nossos experimentos, utilizamos três conjuntos de dados: anônimo-1[6], FakeTweetBr[7] e Fake.br-Corpus[8]. O anônimo-1 contém notícias sobre a cura do COVID-19. O FakeTweetBr é um corpus de notícias falsas do Twitter. O Fake.br-Corpus contém notícias classificadas em seis grandes categorias (política, TV e celebridades, sociedade e notícias diárias, ciência e tecnologia, economia, religião).

5. Análise dos Resultados

As pistas linguísticas sugerem a probabilidade de que o transmissor esteja tentando enganar. Sendo assim, as pistas linguísticas estão inseridas em várias unidades de texto, incluindo palavras, frases, sentenças ou mensagens. A Tabela 2 apresenta os resultados obtidos nos três conjuntos de dados.

[1]https://github.com/pseudorfrodrigues/LinguisticCluesTool
[2]https://universaldependencies.org/docs/u/pos/
[3]https://github.com/Deceptive-Content-Utilities/LeIA
[4]https://www.freeoffice.com/pt/baixar/dicionarios
[5]https://spacy.io/
[6]https://wp.ufpel.edu.br/midiars/datasets/
[7]https://github.com/prc992/FakeTweet.Br
[8]https://github.com/roneysco/Fake.br-Corpus

Tabela 2. Resultados obtidos nos três datasets.

Pista linguística	anônimo-1: outro	anônimo-1: desmentido	anôniomo-1: desinformação	Fake.br: True	Fake.br: False	FakeTweetBr: True	FakeTweetBr: False
#p1	2.09	2.32	**2.75**	58.30	12.00	3.73	**4.07**
#p2	10.47	12.05	**16.58**	11.88	**12.97**	10.98	**13.00**
#p3	4.80	5.05	**4.52**	4.88	4.90	6.20	**5.54**
#p4	0.96	0.93	**0.77**	2.81	**2.64**	1.39	**1.26**
#p5	8.79	8.50	**8.27**	18.92	**15.05**	7.72	**6.99**
#p6	2.78	2.17	**3.02**	2.64	**2.95**	3.40	**3.42**
#p7	0.00∼	0.00∼	0.50	0.21	**0.19**	0.12	**0.09**
#p8	0.00∼	0.00∼	0.00∼	0.16	0.13	0.00∼	**0.09**
#p9	5.98	9.88	**12.06**	8.18	8.11	5.67	**7.25**
#p10	0.50	0.44	**0.50**	0.45	0.43	0.36	0.35
#p11	0.22	0.26	0.12	0.31	0.30	0.26	**0.28**
#p12	0.22	0.58	**0.75**	0.69	0.69	0.55	**0.60**
#p13	9.19	11.81	**11.56**	10.69	11.87	16.68	**15.41**
#p14	0.68	0.70	**0.69**	0.34	0.52	0.70	0.72
#p15	12.43	13.79	15.62	379.79	**93.26**	20.15	20.60
#p16	0.71	2.94	2.20	1.39	**1.83**	3.82	3.16
#p17	0.21	0.48	**1.01**	0.03	**0.31**	0.68	**2.20**

O presente trabalho buscou descobrir as possibilidades do uso de pistas linguísticas para distinguir conteúdos verdadeiros de enganosos no idioma português brasileiro. Encontramos 6 pistas linguísticas (#p2, #p4, #p5, #p6, #p12 e 17) que obtiveram o comportamento sugerido pelos trabalhos relacionados, para os três conjuntos de dados, 5 pistas linguísticas em dois conjuntos de dados (#p1, #p3, #p9, #p7 e #p13) e 6 pistas linguísticas em apenas um conjunto de dados (#p16, #p14, #p15, #p11, #p10 e #p8).

Nos estudos de (Zhou and Zhang 2008), escritores de conteúdo enganoso possuem maior afeto positivo e negativo, porém, a teoria da perspectiva auto-apresentacional diz que eles são menos positivos e agradáveis, o que foi verificado neste trabalho.

Esperava-se que a emotividade fosse maior em conteúdos enganaosos, mas observamos que o número de adjetivos e advérbios é menor em quantidade nas informações enganosas, sendo essas duas informações importantes para o cálculo da emotividade. Dessa forma, obtivemos menos emotividade no Fake.Br e no FakeTweetBr.

Nos trabalhos de (Zhou et al. 2003; Zhou et al. 2004; Zhou and Zhang 2008), foi possível observar maior informalidade em conteúdos enganosos. Porém, é importante ressaltar que os conjuntos de dados usados por estes autores são balanceados. Em nossos experimentos, o único conjunto de dados que resultou em mais informalidade foi o Fake.br, que foi o único conjunto de dados balanceado utilizado, os demais eram desbalanceados. O que pode ter contribuído com a baixa acertividade dessa pista linguística.

6. Conclusões

Este trabalho apresentou a ferramenta LC-Tool que extrai pistas linguísticas de corpus rotulados com conteúdo enganoso. Encontramos 6 pistas que alcançaram o comportamento esperado nos conjuntos de dados anônimo-1, Fake.br e FakeTweetBr.

O estudo da detecção de conteúdo enganoso é importante para a comunidade acadêmica e a sociedade em geral. Nessa perspectiva, estamos trabalhando para que em trabalhos futuros possamos realizar experimentos utilizando técnicas de aprendizado profundo. Para isso, pretendemos aplicar o modelo BERTimbau em conjuntos de dados

com conteúdo enganoso escritos em língua portuguesa.

Referências

Fernandez, A. C. and Devaraj, M. (2019). Computing the linguistic-based cues of fake news in the philippines towards its detection. pages 1–9.

Fuller, C., Biros, D., Twitchell, D., Burgoon, J., and Adkins, M. (2006). An analysis of text-based deception detection tools. volume 6, page 418.

Litvinova, O., Seredin, P., Litvinova, T., and Lyell, J. (2017). Deception detection in Russian texts. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 43–52, Valencia, Spain. Association for Computational Linguistics.

Zhou, L., Burgoon, J., and Douglas, T. (2004). A comparison of classification methods for predicting deception in computer-mediated communication. 20(4):139–165.

Zhou, L., Burgoon, J., Twitchel, D., Quin, T., and Jay, N. (2003). An exploratory study into deception detection in text-based computer-mediated communication. 20(4):1–10.

Zhou, L. and Sung, Y.-w. (2008). Cues to deception in online chinese groups. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, pages 146–153.

Zhou, L. and Zhang, D. (2008). Following linguistic footprints: Automatic deception detection in online communication. *Commun. ACM*, 51(9):119–122.