Análise do viés nas notícias sobre a Ucrânia através de processamento de linguagem natural

Johnatan Mucelini[1]*; Jéssica Peixoto de Araújo[2]

[1] Universidade de São Paulo. Doutorando no Departamento de Química de São Carlos. Rua Dr. Carlos de Camargo Salles, nº 306 – Jardin Lutfalla; 13560-550, São Carlos, São Paulo, Brasil.
[2] Nome da Empresa ou Instituição (opcional). Titulação ou função ou departamento. Endereço completo (pessoal ou profissional) – Bairro; 00000-000 Cidade, Estado, País
*autor correspondente: Johnatan.mucelini@gmail.com

Análise do viés nas notícias sobre a Ucrânia através de processamento de linguagem natural

Resumo

Através de web scraping, realizou-se a coleta de notícias de cinco mídias internacionais: Al Jazeera, BBC, Global Times, RT, New York Times e The Times. Após limpeza dos dados textuais e análises de sentimentos realizadas com léxicos e processamento de linguagem natural (PLN) verificou-se possíveis evidências de vieses dos veículos de mídia, como viés de cobertura, de declaração e de forças de oferta e demanda. O léxicon NRC se mostrou sensível para captar as diferentes características das notícias. O workflow desenvolvido aqui pode ser reproduzido para outros assuntos e servir de base para que especialistas analisem possíveis vieses em notícias da sua área.
Palavras-chave: NLP, análise de sentimentos, mineração de texto, léxicon, viés mídia.

Introdução

O viés midiático pode impactar a sociedade de várias formas, desde favorecer determinadas ideologias (Gerbner et al., 1982) até favorecer candidaturas ou partidos políticos. D'Alessio e Allen (2000), fizeram uma metanálise de 59 investigações quantitativas de viés midiático de eleições americanas para presidente olhando pela perspectiva de como o viés acontecia de fato. Eles classificaram os vieses em três categorias: gatekeeping (porteiro), quando ocorre a seleção de quais histórias serão cobertas e por consequência quais não serão; coverage (cobertura), quando a quantidade de coberturas jornalísticas tende a favorecer um lado em detrimento de outros; e statement (declaração), quando as pessoas trabalharam no veículo midiático colocam suas opiniões no conteúdo jornalístico. Hamilton (2011), por outro lado, procurou olhar o viés através da sua origem. Ele constatou que "notícias são commodity, não um espelho da realidade" e por isso elas são moldadas por forças de oferta e de demanda. A oferta parte dos veículos midiáticos enquanto a demanda parte público-alvo das notícias. Hamilton torna evidente como as questões econômicas e ideológicas podem gerar viés nas notícias.

Um caso interessante para estudar o viés midiático são as notícias relacionadas os recentes conflitos na Ucrânia, onde houveram diversas denúncias de viés racista.(Salacanin, 2000;Al Lawati e Ebrahim, 2022;Huaxia Group, 2022) As denúncias se concentram em veículos de mídias ocidentais com destaque para Europa e Estados Unidos da América. Por exemplo, "Para ser franco, esses não são refugiados da Síria, são refugiados da Ucrânia ... Eles são cristãos, são brancos, são muito semelhantes (a nós)", fala Kelly Cobiella para a NBC News. Phillipe Corbe fala na BFM TV que "Não estamos falando aqui de sírios fugindo do bombardeio do regime sírio apoiado por Putin, estamos falando de europeus saindo em carros parecidos com os nossos para salvar suas vidas"

Ao mesmo tempo, parece que a forma como os veículos de mídia ocidentais apresenta a Ucrânia mudaram ao longo dos conflitos armados na Ucrânia, que se iniciaram com uma crise em 2014 e 2015.(Bonet, 2015) Em fevereiro de 2015, o The Guardian noticiou “Bem-vindo a Ucrânia, o país mais corrupto da Europa” em. Em setembro de 2019, o Vox traz a manchete “Um comediante ucraniano que se tornou presidente está envolvido na confusão do impeachment de Trump”. Em abril de 2021, o New Europe notícia “O governo do presidente ucraniano se torna cada vez mais corrupto e autoritário”. Já em 28 de fevereiro de 2022, o The Guardian traz a manchete “A luta pela Ucrânia é uma luta pelos ideais liberais. Então, como Boris Johnson pode liderá-lo?”. Em 4 de março do mesmo ano o The Washington Post noticiou: “Zelensky: O presidente da TV virou herói de guerra”. Aparentemente, houve uma mudança de postura dessas mídias próximo o acirramento dos conflitos armados decorrente da intervenção Russa, em 24 de fevereiro de 2022, passando o foco das notícias de pontos negativos para positivos da Ucrânia e do presidente Zelensky (desde 2019).

O motivo da aparente mudança de postura desses veículos de mídia não é claro, mas qualquer que seja a explicação precisa ser suportada por dados. Surge a questão, seria possível descrever e caracterizar, através de dados, a mudança de postura dos veículos midiáticos ocidentais, em relação a Ucrânia, ao longo dos anos?

Para responder essa pergunta é necessário encarar a questão das notícias como um problema de ciência de dados. E para isso, é necessário coletar milhares de notícias sobre a Ucrânia, de vários jornais e ao longo de muitos anos. Uma análise da aparente mudança de postura precisa se basear em informações obtidas por metodologias sólidas baseadas em processamento de linguagem natural [PLN].(Silge e Robinson, 2017)

Ao encarar o problema dessa forma, surgem outras questões. A análises de sentimentos dessas notícias poderiam mensurar viés de declaração ou cobertura? (D’Alessio e Allen, 2000) Outra questão importante é saber quais foram as mudanças de viés dos jornais russos sobre o mesmo tema. Pois há denúncias de censura por parte do governo a veículos de mídia estatal e privados.(Troianovski e Safronova, 2022)

Dois trabalhos particularmente relevantes nesse tema abordaram notícias relacionadas a Ucrânia durante o período da crise de 2014 e 2015. Cremisini et al. (2019) analisaram 4,538 artigos relacionados ao período de classificando como “pro-Rússia”, “pro-Ocidente”, ou “Neutro” para treinar modelos com o objetivo de predizer o viés artigos novos. Entretanto, os resultados encontrados indicam que o modelo criado acabou aprendendo o estilo jornalísticos dos autores das notícias com viés, e não o que é o invés em si. Färber et al. (2020) usaram uma estratégia de combinar análises de especialistas e não especialistas para classificar, sentença por sentença, 90 artigos de Cremisini et al. (2020) de acordo com 12 dimensões de viés. Isso resultou em 43,197 sentenças anotadas por especialistas e não

especialistas que apontaram diversas tendencias de vieses nas notícias, por exemplo, a tendência de notícias que tem viés pro-ocidente positivo ter pro-Rússia negativo.

Material e Métodos

Coletou-se notícias lançadas entre 2016 e 2022 de seis mídias internacionais diferentes: Al Jazeera, Global Times, RT, BBC, New York Times, The Times. Utilizou-se técnicas de web scraping em python com auxílio das bibliotecas *selenium* para coletar links de notícias e o pacote *newspaper3k* (Ou-Yang) para baixar as notícias e extrair o texto do HTML. A coleta dos links foi feita de duas formas diferentes. A mídia Al Jazeera disponibiliza em seu site os links de todas as suas notícias, e por isso coletou-se todos diretamente, ignorando apenas as notícias que a própria mídia classifica como galeria de fotos, vídeos, podcasts e esportes, por trazer poucos dados utilizável para a análise de texto que o presente estudo propõe.

As outras mídias não disponibilizam links das notícias, e, portanto, construiu-se um script baseado em *selenium* que fazer pesquisas no google e coleta os links dos resultados. Nestas pesquisas foi buscado a palavra “Ucrânia” restringindo a busca para o site da mídia em investigação e para 200 intervalos de tempo que compreendiam o período total entre 2016 e 2022. Como após a invasão russa na Ucrânia a quantidade de notícias aumentou muito, os intervalos de tempo foram estipulados de forma que a distribuição de notícias da sobre a Ucrânia na Al Jazeera fosse homogênea. Ao todos, o script fez quase 1000 pesquisas no google, coletando os links até páginas 30 de resultados e evitando bloqueio do PI através de rotacionando de VPN.

Utilizou-se o pacote *language-detection* (Nakatani, 2010) para detectar e remover notícias que não fossem em inglês. A Tabela 1 apresenta a quantidade final de notícias coletadas.

Site	**Número de Notícias**	**Número de Notícias da Ucrânia**
Al Jazeera	98925	3328
Global Times	6889	382
RT	9803	605
BBC	14848	2790
New York Times	13687	2122
The Times	19912	1485
Total	**164064**	**10712**

Tabela 1. Quantidade de notícias coletadas que foram utilizados como dados desta pesquisa. Fonte: Resultados originais da pesquisa.

Verificou-se que a biblioteca *newspaper3k* utilizada para coletar o texto da notícia performou bem para Al Jazeera, Global Times e RT, enquanto que para o BBC, New York Times, The Times e algumas notícias foram coletadas incompletas. A quantidade média de caracteres das notícias por mídia foi: Global Times 4018, Al Jazeera 3197, RT 3440, New York Times 3131, The Times 2411 e BBC 2265.

Na limpeza dos dados textuais removeu-se caracteres problemáticos como aspas, emoji utilizando a biblioteca *cleantext*. A lista de s*topwords* foi coletadas dos pacotes *NLTK* (Bird, 2006), *wordcloud* (Mueller, 2020) *e Stopwords English (EN)*. Realizou-se a tokenização e lematização dos dados com base no NLTK e que usa uma lematização snowball (Porter, 1980) baseada na WordNet (Miller, 1995).

Utilizou-se análise de sentimentos usando lexicons para tentar verificar os "sentimentos" do texto das notícias. Os lexicons são dicionários que tentam relacionar termos presentes no texto a sentimentos de forma quantitativa, onde o sentimento mensurado não necessariamente é um sentimento literal. Utilizou-se três léxicons: VADER (Hutto e Gilbert, 2014) que estima os sentimentos positivo, negativo e score, que é um saldo entre positivo e negativo; AFINN (Nielsen, 2011), que mensurado um sentimento score; e NRC (Mohammad e Turney, 2013) que estima, além dos sentimentos positivo e negativo, os sentimentos surpresa, medo, raiva, ansiedade, confiança, tristeza, nojo, e alegria.

Resultados e Discussão

A Figura 1 mostra que, para todas as mídias, a quantidade de notícias sobre a Ucrânia aumenta em 2022 em relação ao todo o período entre 2016 e 2021. A mídia que mais aumenta é a Global Times, que aumentou quase 70 vezes na proporção anual, enquanto que a menos aumentou foi a RT, que aumentou pouco mais de 6 vezes. As demais mídias aumentaram entre 18 e 28 vezes.

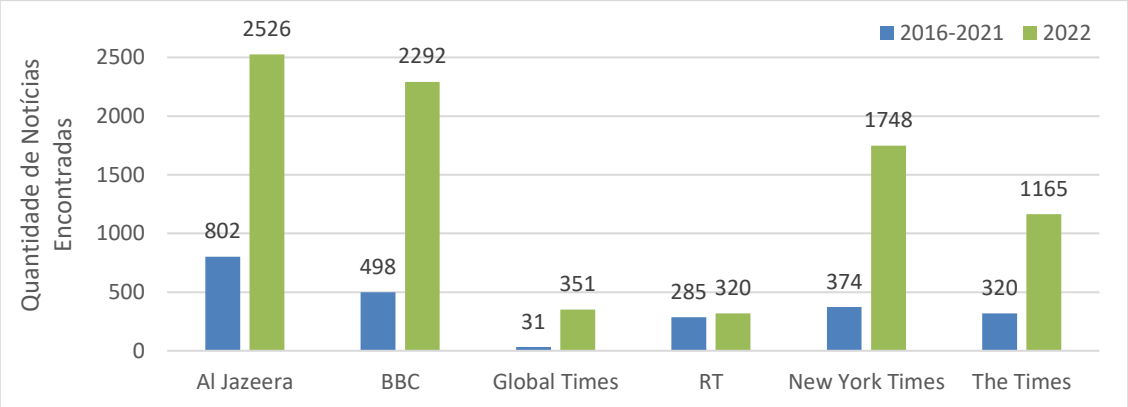

Figura 1. Quantidade de notícias por mídia e para os períodos 2016-2021 e 2022.
Fonte: Resultados originais da pesquisa.

Teste dos Lexicons para análise de notícias:

O pequeno aumento detectado na quantidade de notícias da mídia RT pode ser por ela ser uma mídia controlada pelo estado Russo que poderia estar publicizando menos a guerra do que as outras mídias. Em princípio isso poderia ser um viés de cobertura, viés de porteiro ou viés de força de oferta, mas não se pode excluir a possibilidade de que isso se deva a amostragem feita com pesquisas automatizadas no google não amostra bem. O grande aumento da Global Times se deve apenas a baixa quantidade de notícias sobre a Ucrânia que eram noticiadas antes.

Para verificar a capacidade dos diferentes lexicons em descrever as notícias, analisou-se os sentimentos das notícias da mídia Al Jazeera que apresenta grande volume e notícias tanto sobre a Ucrânia quanto sobre outros assuntos. A Figura 2 apresenta a distribuição dos sentimentos médios por notícia e separadas por ano para os scores dos léxicons VADER, AFINN e a diferença entre os sentimentos positivo e negativo do léxicon NRC. Para os scores de AFINN e NCR, a figura apresenta o valor estandardizado ao longo do período 2016-2022 para facilitar a visualização. Não há diferença significativa clara entre as distribuições com exceção da falta de normalidade nas distribuições das notícias.

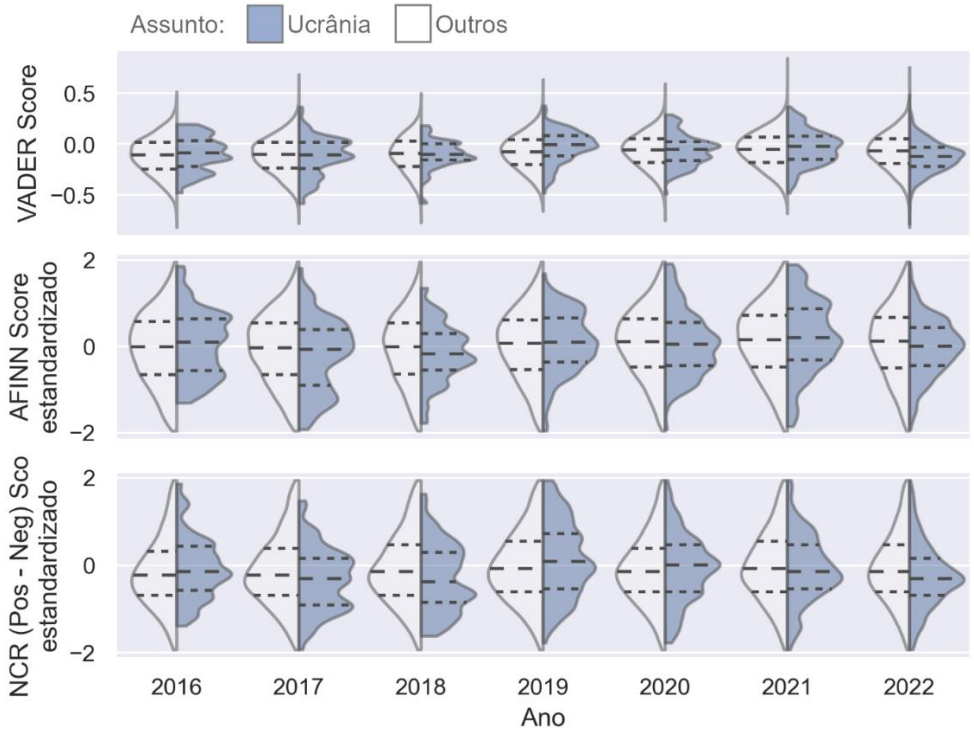

Figura 2. Distribuição dos sentimentos médios por notícia da mídia Al Jazeera, separadas por ano, para os sentimentos de score do léxicon VADER e para os valores estandardizado dos sentimentos score dos léxicons AFINN e da diferença entre os sentimentos positivo e negativo do léxicon NRC.

Fonte: Resultados originais da pesquisa.

A Figura 3 apresenta a média por ano das notícias da mídia Al Jazeera para valores dos sentimentos não binários do léxicon NCR, entre 2016 e 2022, diferenciando as notícias relacionadas a Ucrânia das outras notícias. Nota-se que até 2021 a diferença destas médias variava muito pouco, ou seja, em média as notícias relacionadas a Ucrânia e as não relacionados a Ucrânia apresentavam quantidades muito próxima dos mesmos sentimentos. Em 2022, entretanto, a quantidade média dos sentimentos emoções de medo, raiva, ansiedade e confiança aumentam significativamente. Uma vez que os resultados demostraram que o léxicon NRC demostra sensibilidade para a análise dos textos enquanto os outros não, utiliza-se apenas ele para outras análises de sentimentos.

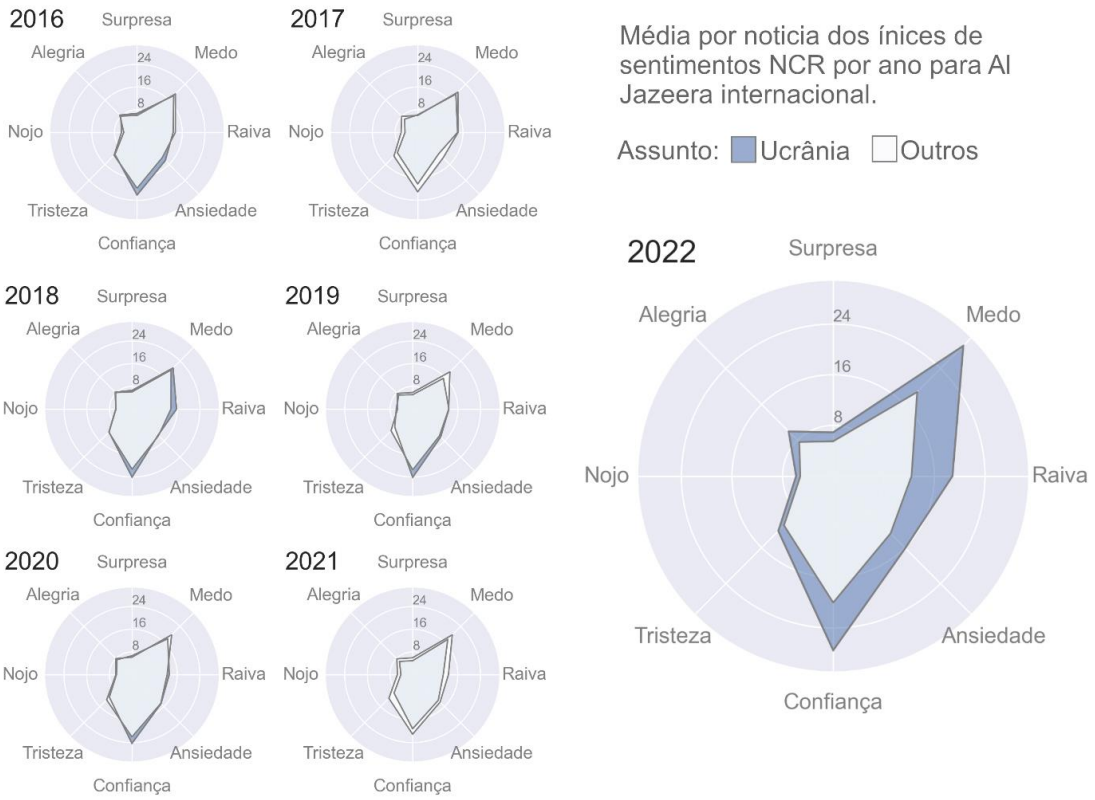

Figura 3. Média por artigo da Al Jazeera por ano dos sentimentos surpresa, medo, raiva, ansiedade, confiança, tristeza, nojo e alegria, obtidos com o léxicon NRC.
Fonte: Resultados originais da pesquisa.

Análise das notícias da Ucrânia

Para descrever os assuntos nos textos das notícias analisadas que tão origem aos sentimentos obtidos com o léxicon NRC, desenhou-se nuvens de palavras com a biblioteca *wordcloud* (Mueller, 2020), após a limpeza do texto e remoção de *stopwords*. A Figura 4 apresenta os sentimentos e palavras mais frequentes para todo o conjunto as 10712 notícias

da Ucrânia onde algumas palavras foram removidas da nuvem para melhorar a visualização. Visando analisar se de fato existem diferenças expressivas entre antes e depois da invasão russa, calculou-se os sentimentos NRC separando os conjuntos de dados em antes e depois das invasões, conforme mostra a Figura 5. Analogamente, calculou-se os sentimentos das notícias separando-as entre antes e depois da invasão e pela mídia que publicou a notícia, conforme mostram as Figuras 6 e 7.

Figure 4. Nuvem de palavras e sentimentos NCR obtidos com todas as 10712 notícias coletadas. Algumas palavras foram removidas da nuvem para facilitar a visualização e estão anotadas na figura em ordem de frequência descendente.
Fonte: Resultados originais da pesquisa.

Pode-se observar que após a invasão houve um aumento de 32% no medo. Em concordância, vê-se na Figura 5 que palavras como *war* (guerra), *force* (força/grupo de soldados), e *invasion* (invasão) ficam mais destacadas após a invasão. Apesar de todas as mídias indicam um aumento no sentimento de medo após a invasão, a BBC aumenta apenas 12%, enquanto New York Times e RT aumentam 69 e 60%, respectivamente, e as demais aumentaram entre 38 e 42%. Isso poderia estar relacionado a vieses de força de oferta, já que a mídia estatal britânica BBC tem muito menos interesse na atual guerra do que a mídia estadunidense New York Times e a mídia estatal russa RT.

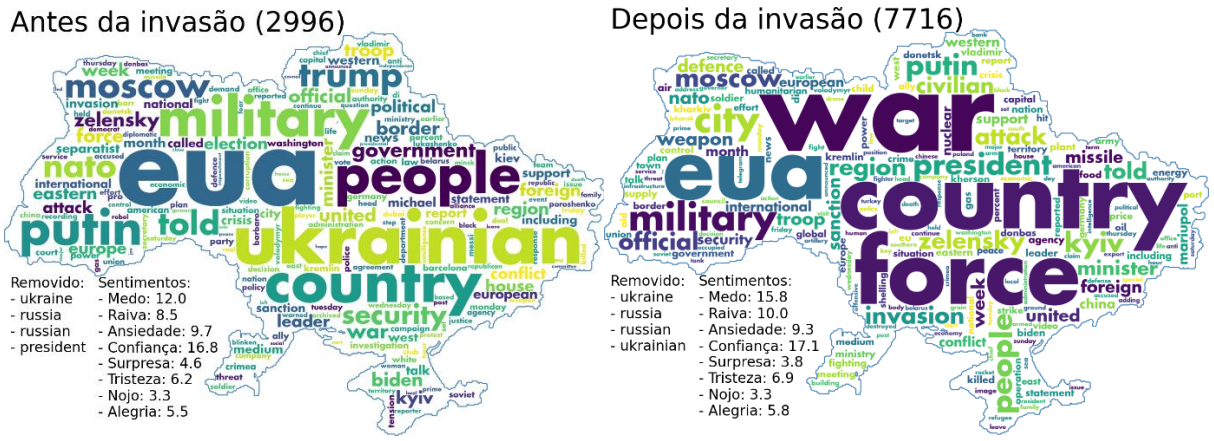

Figura 5. Duas nuvens de palavras com sentimentos NCR para as 2996 notícias sobre a Ucrânia de antes da invasão russa na Ucrânia e para as 7716 notícias sobre a Ucrânia de depois da invasão.

Fonte: Resultados originais da pesquisa.

Nota-se um aumento de 18% no sentimento de raiva após a invasão que se destaque para a mídias New York Times que aumenta 52%. A única exceção é a mídia BBC que tem seu sentimento de raiva reduzido em 5%.

Curiosamente, o sentimento de tristeza aumenta apenas 11% após a invasão considerando todas as notícias sobre a Ucrânia, com aumento destacado nas mídias RT, 31%, e Global Times, 81%. A quantidade de notícias na mídia Global Times antes da invasão, 52, é relativamente baixa e por isso pode indicar que essa mídia não tinha uma cobertura tão boa quanto as demais dos acontecimentos da Ucrânia antes da invasão. O aumento da RT talvez se deva a uma aparente mudança de editorial, evidenciado pela mudança na nuvem de palavras após a invasão, que passa de exportes, principalmente futebol, para a cobertura da guerra.

Nota-se que após a invasão todas as mídias, com exceção da RT, passaram a usar com altíssima frequência os termos *war* (guerra) e *conflict* (conflito). As palavras aparecem na nuvem de palavras da Figura 6, mas com baixa frequência quando comparada com as demais Mídias. A princípio isso poderia ser tanto um viés de força de oferta da mídia estatal russa RT, ou um viés de declaração.

As outras emoções mudaram muito pouco. Nota-se pequenas alterações pequenas, em geral reduções, no sentimento de ansiedades. A exceção são o New York Times e o Global Times que aumentaram em 13 e 30%, respectivamente. Isso pode ser devido ao fato da pena amostragem de 52 notícias para o Global Times antes da invasão.

Figure 6. Nuvens de palavras com sentimentos NCR para notícias da Al Jazeera, Global Times e RT sobre a Ucrânia de antes da invasão russa na Ucrânia para. O número entre parênteses indica a quantidade de notícias analisadas. Algumas palavras foram removidas das nuvens pra facilitar a visualização e estão anotadas.

Fonte: Resultados originais da pesquisa.

Figure 7. Nuvens de palavras com sentimentos NCR para notícias da BBC, New York Times e The Times sobre a Ucrânia de antes da invasão russa na Ucrânia para. O número entre parênteses indica a quantidade de notícias analisadas. Algumas palavras foram removidas das nuvens pra facilitar a visualização e estão anotadas.
Fonte: Resultados originais da pesquisa.

Considerações Finais

O léxicon NRC se mostrou mais sensível para as notícias em ralação ao léxicons VADER, AFINN. Foi possível descrever várias mudanças no texto das notícias após a invasão

russa na Ucrânia através da combinação de análise de sentimentos baseados em léxicons e de nuvens de palavras para veículos de mídia. Com essas mudanças pode-se verificar indícios de vieses nas mídias estudadas. Diferente dos trabalhos de Cremisini et al. (2019) e Färber et al. (2020), as evidências de vieses encontradas neste trabalho dificilmente podem ser confundida com estilos de escrita de autores já que se baseia em analisamos grandes quantidades de notícias de muitos autores, ao invés de uma notícia por vez como foi realizado no estudo destes autores. Ao mesmo tempo, a necessidade de grandes volumes de notícias é uma limitação do método. A análise de dados empregada por especialistas em questões geopolíticas e o método em si pode ser extrapolado para outros assuntos.

Referências

Gerbner, G.; Gross, L.; Morgan, M.; Signorielli, N. 1982. Charting the mainstream: Television's contributions to political orientations. Journal of communication. 32: 100-127.

D'Alessio, D.; Allen, M. 2000. Media bias in presidential elections: A meta-analysis. Journal of communication. 50: 133-156.

Hamilton, J. T. 2011. All the News That's Fit to Sell: How the Market Transforms Information into News. 1ed. Princeton University Press, Princeton, New Jersey, United States.

Salacanin, S. Ukraine Coverage Exposes Western Media Bias. Inside Arabia. 2022. Disponível em: <https://insidearabia.com/ukraine-coverage>. Acesso em: 4 nov. 2022

Al Lawati, A.; Ebrahim, N. How the Ukraine war exposed Western media bias. CNN. 2022. Disponível em: <https://edition.cnn.com/2022/03/04/media/mideast-summary-04-03-2022-intl/index.html>. Acesso em: 4 nov. 2022

Huaxia Group. Western coverage of Ukraine exposes deep-seated racist bias, double standards. Xinhuanet. 2022. Disponível em: <http://english.news.cn/europe/20220317/7d592a03633e4c0f9c89e1cc27a69221/c.html>. Acesso em: 4 nov. 2022

Bonet, P. Em 2014, Ucrânia viveu seu pior ano desde a independência em 1991. El Pais. 2014. Disponível em: <https://brasil.elpais.com/brasil/2015/01/01/internacional/1420136723_852421.html>. Acesso em: 4 nov. 2022

Norton, B. 2022. New York Times' ridiculous attack on me exposes its deceitful propaganda tactics. Multipolarista. Disponível em: <https://multipolarista.com/2022/04/14/new-york-times-attack-propaganda/>. Acesso em: 4 nov. 2022

Silge, J.; Robinson, D. 2017. Text mining with R: A tidy approach. 1ed. O'Reilly Media, United States of America.

Troianovski, A.; Safronova, V. Russia Takes Censorship to New Extremes, Stifling War Coverage. New York Times. 2022. Disponível em: <https://www.nytimes.com/2022/03/04/world/europe/russia-censorship-media-crackdown.html>. Acesso em: 4 nov. 2022

Cremisini, A.; Aguilar, D.; Finlayson, M. A. 2019. A Challenging Dataset for Bias Detection: The Case of the Crisis in the Ukraine. p. 173-183. In:Social, Cultural, and Behavioral Modeling. 1 ed. Springer. Cham, Germany.

Färber, M.; Burkard, V.; Jatowt, A.; Lim, S. 2020. A Multidimensional Dataset Based on Crowdsourcing for Analyzing and Detecting News Bias. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020. Virtual Event, Ireland. p. 3007-3014.

Ou-Yang, L. Disponível em: <https://newspaper.readthedocs.io/en/latest/#>. Acessado em: 4 nov. 2022.

Hutto, C.J.; Gilbert, E.E. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In: Proceedings of the 14th International Conference on Weblogs and Social Media. 2020. Ann Arbor, USA.

Nielsen, F. Å. 2011. Disponível em: <https://arxiv.org/abs/1103.2903> Acessado em: 4 nov. 2022.

Mohammad, S. M.; Turney, P. D. 2013. NRC emotion lexicon. National Research Council, Canada. 2: 234.

Bird, S. 2006. NLTK: the natural language toolkit. In: Proceedings of the COLING/ACL 2006.

Miller, G. A. 1995. WordNet: a lexical database for English. p. 39-41. In: Communications of the ACM 38.

Mueller, A. 2020. Wordcloud. Disponível em: <https://github.com/amueller/word_cloud> Acessado em: 24 de abril de 2023.

Nakatani, S. 2010. Language-detection. Disponível em <https://github.com/shuyo/language-detection> Acessado em: 24 de abril de 2023.

Porter, M. F. 1980. An algorithm for suffix stripping. Program. 14: 130-137