

An Analysis of Subjectivity in Brazilian News

D. F. Lima¹, A. S. C. Melo², L. B. Marinho³

Universidade Federal de Campina Grande, Brazil

diogo.florencio@copin.ufcg.edu.br, allanmelo@copin.ufcg.edu.br, lbmarinho@dsc.ufcg.edu.br

Abstract. With the advent of digital journalism, the democratization of information has become a reality, since news articles are published as soon as the facts occur and are accessible from any device connected to the internet. It is common sense the perception that some newspapers are more biased than others when it comes to the way of exposing the facts. However, automatic ways of measuring such biases is still an open research challenge. Under the premise that journalistic texts must have objective and unbiased language, news with high levels of subjectivity may indicate bias. In this paper, we propose to use subjectivity lexicons to characterize subjectivity in five news portals that are popular in Brazil. To better understand the results found, we performed a correlation analysis between the levels of subjectivity found and readability and news popularity metrics. We believe that the methods we used along with our findings contribute to a better understanding of the linguistic characteristics of the news we consume daily.

Categories and Subject Descriptors: I.2.7 [Natural Language Processing]: Text analysis; I.2.6 [Artificial Intelligence]: Learning

Keywords: Bias, Machine Learning, Natural Language Processing, News, Subjectivity

1. INTRODUÇÃO

De acordo com uma pesquisa de 2017 da Fundação *Knight and Gallup*, os norte americanos acreditam que 62% das notícias que eles consomem são enviesadas¹. No livro *Bias: A CBS Insider Exposes How the Media Distort the News* [Goldberg 2001], Bernard Goldberg, que trabalhou cerca de trinta anos como repórter da rede de televisão americana CBS, expõe que a cobertura jornalística é frequentemente enviesada para beneficiar partidos, grupos, pessoas ou ideias que se alinham à ideologia dos jornalistas.

De modo geral, notícias jornalísticas devem utilizar uma linguagem formal, clara, objetiva e em terceira pessoa, marcando a impessoalidade. Em contrapartida, notícias tendenciosas frequentemente utilizam uma linguagem mais subjetiva e pouco objetiva visando influenciar ou persuadir leitores através do apelo à emoção e/ou técnicas de persuasão [Wiebe et al. 2005; Mihalcea et al. 2007]. Sob a premissa de que legibilidade e imparcialidade são características do gênero jornalístico, notícias com altos níveis de subjetividade e menor legibilidade podem indicar algum tipo de viés.

Em trabalho prévio, léxicos de subjetividade construídos manualmente por especialistas linguistas [Amorim et al. 2018] foram utilizados para medir, por meio de *word embeddings*, a subjetividade de textos linguísticos [Sales et al. 2019]. Nesse artigo, nós estendemos esse trabalho caracterizando a subjetividade de notícias em cinco portais de notícias populares no Brasil: *Estadão*, *Folha de São Paulo*, *O Antagonista*, *O Globo* e *Veja*. Para melhor entender os resultados encontrados, correlacionamos os níveis de subjetividade encontrados com o tamanho da notícia, métricas de legibilidade e popularidade das notícias. Consideramos popularidade como a quantidade de comentários das notícias e para legibilidade usamos as seguintes métricas: Coleman Liau Index [Coleman and Liau 1975],

¹<https://www.knightfoundation.org/reports/perceived-accuracy-and-bias-in-the-news-media>

Smog Index [Nigam et al. 2016], Lix Index [Anderson 1983] e Dale Chall Score [Klare 1952]. Para guiar nossa pesquisa definimos as seguintes questões de pesquisa (QP):

- QP1: Como se comparam os níveis de subjetividade entre os portais considerados?
- QP2: Há correlação entre subjetividade e legibilidade?
- QP3: Há correlação entre subjetividade e tamanho da notícia?
- QP4: Há correlação entre subjetividade e popularidade?

Acreditamos que o método usado, juntamente com as respostas obtidas a essas questões, contribuem para um melhor entendimento sobre as características linguísticas das notícias que consumimos diariamente.

2. TRABALHOS RELACIONADOS

Subjetividade é um conceito que vem sendo largamente estudado em textos jornalísticos [Wilson et al. 2005; Chaturvedi et al. 2015; Yaqub et al. 2018]. Wilson et al. [Wilson et al. 2005], estudaram subjetividade em notícias identificando automaticamente opiniões, sentimentos e especulações presentes no texto, por meio de um classificador *Naive Bayes*. Chaturvedi et al. [Chaturvedi et al. 2015], por sua vez, trabalharam com detecção de subjetividade através de redes neurais convolucionais baseadas em *word embeddings*.

Por vezes o conceito de subjetividade é associado a popularidade. Bae e Lee [Bae and Lee 2012] analisaram a popularidade de *tweets* através do viés presente no texto, neste caso o viés foi definido em termos da influência de sentimentos positivos e negativos; Flaounas et al. [Flaounas et al. 2013] exploraram a predição de popularidade de notícias por meio do viés presente no fato noticiado.

Nossa análise de subjetividade é automática e baseada em léxicos de subjetividade. Essa metodologia é usada em alguns trabalhos relacionados [Amorim et al. 2018; Sales et al. 2019; Moraes et al. 2016; Jha et al. 2016]. Amorim et al. [Amorim et al. 2018], por exemplo, utilizam a abordagem de cálculo de subjetividade através de léxicos para avaliar comentários de avaliadores de redações do ENEM (Exame Nacional do Ensino Médio). Para avaliar a subjetividade, a pesquisadora utiliza léxicos de argumentação, sentimento, pressuposição, modalização e valoração. Sales et al. [Sales et al. 2019], por sua vez, se baseiam nos mesmos léxicos para abordar subjetividade como uma forma de viés midiático ao analisar notícias políticas. Nós estendemos esse trabalho de duas formas: (i) incluindo mais portais na análise e (ii) aprofundando a análise por meio de uma análise de correlação entre subjetividade, legibilidade e popularidade de notícias.

3. COLETA DE DADOS

Foram coletadas notícias e comentários dos cadernos de esporte e política dos portais *Estadão*², *Folha de São Paulo*³, *O Antagonista*⁴, *O Globo*⁵ e *Veja*⁶ durante o período de janeiro à dezembro de 2018. Em 2018, os cadernos de esporte e política cobriram eventos como a copa do mundo de futebol masculino⁷ e as eleições presidenciais brasileiras⁸, momentos especialmente interessantes de serem analisados devido a uma maior tendência das notícias serem apresentadas de forma mais subjetiva, de acordo com o posicionamento esportivo/político do seu editor.

²<https://www.estadao.com.br>

³<https://www.folha.uol.com.br>

⁴<https://www.oantagonista.com>

⁵<https://oglobo.globo.com>

⁶<https://veja.abril.com.br>

⁷https://en.wikipedia.org/wiki/2018_FIFA_World_Cup

⁸https://en.wikipedia.org/wiki/2018_Brazilian_general_election

Os portais *Estadão*, *Folha de São Paulo* e *O Globo* são portais *mainstream* acessados por leitores de todo posicionamento esportivo/político. O *O Antagonista* é um portal declaradamente de direita no Brasil⁹ enquanto a *Veja* se declara como uma revista de oposição ao governo independente de ideologia política¹⁰. Utilizamos a *Veja* e *O Antagonista* como *sanity check* da nossa análise de subjetividade, uma vez que é esperado que esses portais apresentem maiores índices de subjetividade em suas notícias devido a seus posicionamentos políticos bem definidos.

Para cada notícia foram coletadas as seguintes informações: data de publicação, autor, seção, título, texto, url e comentários. Para cada comentário, coletamos: data do comentário, autor e texto. A Tabela I apresenta a quantidade de notícias e comentários coletados por portal.

	Folha de São Paulo	Estadão	O Antagonista	O Globo	Veja
Notícias	30,085	58,702	33,131	35,391	38,076
Comentários	37,492	52,637	2,196,993	45,191	68,789

Table I. Quantidade de notícias e comentários por portal.

4. FUNDAMENTAÇÃO TEÓRICA

4.1 Subjetividade

Subjetividade pode ser definida como a capacidade do interlocutor se propor como sujeito [Benveniste 1971], introduzindo sua opinião ao que é dito, podendo resultar tanto em marcas singulares na opinião dos leitores quanto na construção de crenças e valores compartilhados pela opinião pública.

Em trabalho prévio Sales et al. [Sales et al. 2019] propuseram o uso de léxicos de subjetividade para medir, por meio de *word embeddings*, a subjetividade de textos jornalísticos. Estes léxicos foram construídos por [Amorim et al. 2018] através da análise manual de expressões que frequentemente aparecem em textos quando o interlocutor aparenta expressar alguma subjetividade. Cada léxico encapsula um aspecto de subjetividade, mais especificamente esses aspectos são:

- **Argumentação (arg)** representa palavras e expressões que estão relacionados a um discurso argumentativo, como: “aliás”, “como consequência”, “de certa forma”, “apesar de”.
- **Sentimento (sen)** reúne palavras e expressões que indicam a presença do estado de espírito ou sentimento do interlocutor na notícia. Palavras como “infelizmente”, “felizmente” ou “preferencialmente” são exemplos deste conjunto.
- **Pressuposição (pre)** contém expressões que sugerem que o interlocutor toma como verdade suposições anteriores. Exemplos desse conjunto léxico são: “hoje em dia”, “continuar a”.
- **Modalização (mod)** engloba palavras utilizadas pelo interlocutor quando este tem uma postura estabelecida por algo ou alguém, tais como “inegável”, “fundamental” e “justo”.
- **Valoração (val)** inclui palavras relacionadas a quantidade e/ou classificação, geralmente adjetivando alguma coisa, por exemplo: “absolutamente” e “aproximadamente”.

Através desses aspectos, estimamos o percentual de subjetividade presente no texto da notícia de interesse.

⁹https://pt.wikipedia.org/wiki/O_Antagonista

¹⁰<https://pt.wikipedia.org/wiki/Veja>

4.2 Legibilidade

Para o cálculo de legibilidade, consideramos 4 métricas estado-da-arte: *Coleman Liau Index* [Coleman and Liau 1975], *Smog Index* [Nigam et al. 2016], *Lix Index* [Anderson 1983] e *Dale Chall Score* [Klare 1952]. A seguir explicamos brevemente cada uma das métricas.

4.2.1 *Smog Index*. O *Smog Index* estima o nível de instrução necessário para entender um texto. Ela baseia-se na quantidade de palavras com 3 ou mais sílabas dentro de um total de 30 sentenças do texto, sendo as sentenças escolhidas como as 10 primeiras, 10 do meio e as 10 finais do texto. A métrica é definida por:

$$SMOG = 1.0430 \times \sqrt{P \times \frac{30}{S}}$$

onde P refere-se ao número de palavras com 3 ou mais sílabas dentre as sentenças escolhidas e S indica o total de sentenças texto.

4.2.2 *Coleman Liau Index*. Assim como *Smog*, o *Coleman Liau Index* é uma métrica de legibilidade projetada para avaliar a dificuldade de compreensão do texto. *Coleman Liau Index* baseia-se na quantidade caracteres (letra ou dígito) e sentenças empregadas a cada 100 palavras utilizadas. Assim, a métrica é determinada da seguinte forma:

$$CLI = 0.0588 \times \bar{L} - 0.296 \times \bar{S} - 15.8$$

onde \bar{L} representa a quantidade de caracteres utilizados e \bar{S} o número de sentenças formadas a cada 100 palavras.

4.2.3 *Lix Index*. *Lix Index* indica a dificuldade de ler um texto a partir da quantidade de palavras longas (compostas por mais de 6 letras), total de palavras e a quantidade de períodos utilizados no texto, onde um período é definido pela presença de ponto final, aspas ou primeira letra maiúscula. Sua fórmula é apresentada a seguir:

$$LIX = \frac{L}{B} + \frac{C \cdot 100}{L}$$

B é o número de períodos do texto e C é o número de palavras longas, e L o total de palavras do texto.

4.2.4 *Dale Chall Score*. O *Dale Chall Score*, por sua vez, avalia a dificuldade de leitura do texto a partir do total de palavras difíceis, total de palavras e total de sentenças do texto. As palavras difíceis são identificadas a partir de sua ausência em uma lista pré-definida de 3000 palavras consideradas de fácil compreensão para estudantes do ensino fundamental. Qualquer palavra que não esteja contida na lista é considerada difícil. A métrica é calculada da seguinte maneira:

$$SCORE = 0.1579 \times \left(\frac{D}{L} \times 100\right) + 0.04696 \times \left(\frac{L}{S}\right)$$

D referencia a quantidade palavras difíceis presentes no texto, L é o total de palavras e S indica a quantidade de sentenças.

5. METODOLOGIA

Para melhor representar as relações semânticas das palavras e expressões das notícias, um modelo *word embedding* foi treinado baseado no algoritmo *word2vec skip-gram* [Mikolov et al. 2013], com tamanho de janela 5 e *negative samples* 5, utilizando como entrada um corpus de 195.385 notícias coletadas dos 5 portais selecionados (ver I).

A partir do modelo treinado, calculamos a distância *Word Mover's Distance* (WMD) [Kusner et al. 2015] de toda notícia $n \in N$ para cada léxico de subjetividade $s \in S$, onde $S = \{arg, pre, sen, val, mod\}$. O WMD baseia-se nas palavras de dois documentos pré-definidos, e suas posições no espaço vetorial definido por um modelo *word embedding*, para calcular o menor custo (distância euclidiana) necessário para sobre-posicionar todas as palavras do primeiro documento às do segundo. Assim, a quantidade de subjetividade do tipo s em uma notícia n se dá por $WMD(n, s)$.

Para cada notícia n , $\|S\|$ tipos de subjetividade são calculados. Dessa forma, a subjetividade sub_n de uma notícia n pode ser representada por um vetor de tamanho $\|S\|$:

$$sub_n = \{WMD(n, s) | s \in S\} \quad (1)$$

Utilizando a média aritmética, torna-se possível transformar o vetor de subjetividades em um único valor:

$$perc_n = 1 - \frac{\sum_{s \in S} WMD(n, s)}{\|S\|} \quad (2)$$

Note que a subtração na equação 2 é um artifício matemático para transformar a distância em similaridade. Dessa forma, quanto maior a métrica $perc_n$, maior a subjetividade na notícia.

5.1 Validação

Com o objetivo de validar o uso dos léxicos no nosso contexto, rotulamos cada notícia do *dataset* como opinativa ou informativa e comparamos a subjetividade associada às notícias de cada rótulo. Espera-se que notícias opinativas possuam maior subjetividade que as informativas, dado que esse tipo de notícia tem por objetivo expressar explicitamente a opinião e perspectiva do interlocutor sobre o fato noticiado. Consideramos como notícias opinativas aquelas publicadas nas seções de opinião, blogs e colunas de cada portal; as demais notícias são consideradas como informativas.

A Figura 1 exibe os intervalos de confiança da diferença média de subjetividade entre notícias opinativas e informativas, com 95% de confiança. Cada intervalo foi calculado a partir de 5000 reamostragens de tamanho aleatório, computadas por meio de *bootstrapping* com reposição. Intervalos de confiança contendo o valor 0 indicam que não há uma diferença significativa entre as subjetividades dos portais.

Com exceção do portal *O Antagonista*, o intervalo estimado para os portais são positivos e não incluem 0. Desse modo, observamos que o método utilizado é capaz de estimar maiores índices de subjetividade em notícias opinativas, apresentando o resultado esperado.

No caso do *O Antagonista*, que é declaradamente alinhado a uma ideologia política (direita), o que pode explicar, em algum senso, o fato de não encontramos diferença significativa entre os textos opinativos e informativos.

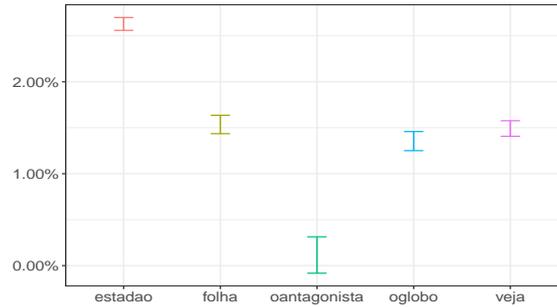


Fig. 1. Diferença média de subjetividade por portal.

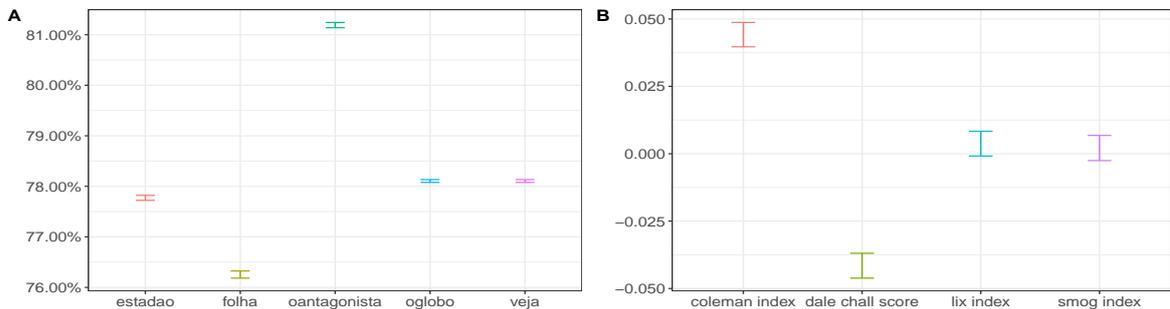


Fig. 2. (A) Média de subjetividade por portal (B) Correlação entre subjetividade e legibilidade.

6. EXPERIMENTOS E RESULTADOS

Nesta seção, são apresentados os experimentos e seus resultados considerando as diferentes questões de pesquisa introduzidas anteriormente. Todos os resultados dos experimentos a seguir são retratados por intervalos com 95% de confiança estimados a partir de 2000 re-amostragens de tamanho aleatório, computadas por meio de *bootstrapping* com reposição.

6.1 Subjetividade por Portal

O primeiro experimento está relacionado à **QP1**. A Figura 2(A) retrata o intervalo de confiança do percentual médio de subjetividade para cada portal.

O Antagonista é o portal com maior índice de subjetividade e, por se tratar de um portal que possui alinhamento político declarado (direita), esse resultado condiz com o esperado. Em seguida, *Veja* e *O Globo* são retratados com valores de subjetividade similares, sem apresentarem diferença significativa. Por último, *Estadão* e *Folha de São Paulo* manifestam, respectivamente, os menores índices de subjetividade.

6.2 Legibilidade vs. Subjetividade

A legibilidade de um texto diz respeito à sua qualidade tipográfica, refletindo em sua facilidade de leitura. Considerando a **QP2**, neste experimento estimamos a correlação de Spearman [Zar 1972] entre subjetividade e as métricas de legibilidade definidos na seção 4.2. Os resultados são retratados na Figura 2(B).

Para qualquer métrica de legibilidade, os intervalos estimados são próximos de 0. Ou seja, a

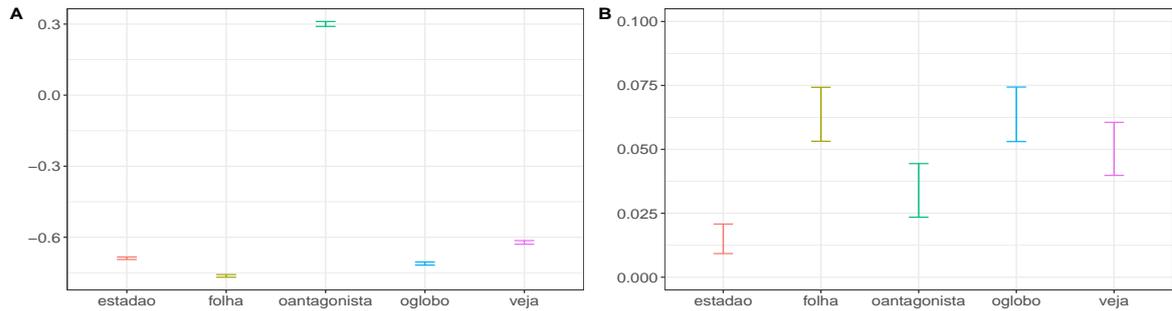


Fig. 3. (A) Correlação entre subjetividade e tamanho (B) Correlação entre subjetividade e popularidade.

correlação existente entre as variáveis é desprezível. Com isso temos indícios de que a subjetividade de uma notícia independe de sua legibilidade.

6.3 Tamanho vs. Subjetividade

Neste experimento, respondemos a **QP3** verificando a correlação entre o tamanho da notícia e sua subjetividade. Nosso interesse é analisar se a estimativa de subjetividade de uma notícia pode ser influenciada por seu tamanho. A Figura 3(A) exibe o intervalo de confiança estimado do coeficiente de correlação de Spearman entre as variáveis tamanho do texto e popularidade da notícia para cada portal.

O portal *O Antagonista* apresenta correlação positiva entre tamanho e subjetividade enquanto os demais portais apresentam correlações negativas. Dentre os resultados, destacam-se os valores de correlação da *Folha de São Paulo*, *Estadão* e *O Globo*, com valores aproximados de -0.75 , o que pode ser considerado uma correlação negativa forte, isso mostra uma tendência de que notícias menores são mais subjetivas.

6.4 Popularidade vs. Subjetividade

Por fim, respondemos a **QP4** analisando a existência de correção entre subjetividade e popularidade. Aqui, estamos interessados em investigar se notícias mais subjetivas são mais populares, o que pode estar relacionado ao apelo à emoção e/ou técnicas de persuasão [Wiebe et al. 2005; Mihalcea et al. 2007], comumente empregadas para influenciar os leitores. A Figura 3(B) exibe o intervalo de confiança estimado do coeficiente de correlação de Spearman por portal. Todos os portais apresentam intervalo próximo de 0. Assim, há indícios de que a popularidade de uma notícia independe de sua subjetividade.

7. CONCLUSÕES E TRABALHOS FUTUROS

É senso comum a percepção de que alguns jornais são mais enviesados que outros na forma de expor os fatos. Notícias com maior subjetividade podem gerar marcas singulares na opinião dos leitores e até mesmo na construção de crenças e valores compartilhados pela opinião pública.

Neste artigo, realizamos uma análise de subjetividade em notícias de 5 portais de notícias brasileiros: *Estadão*, *Folha de São Paulo* e *O Globo*, portais *mainstream*; *O Antagonista* e *Veja*, portais com posições ideológicas declaradas. Relacionamos subjetividade a diferentes aspectos da notícia através da aplicação de métodos estado-da-arte (p.ex. tamanho, legibilidade e popularidade). Nossas principais descobertas são:

— *O Antagonista* e *Veja* foram classificados como os portais mais subjetivos. Intuitivamente, este é

- o resultado esperado, já que são portais com alinhamento político declarado, o que nos dá mais indícios sobre a eficácia do método introduzido;
- O percentual de subjetividade de uma notícia independe da clareza de seu texto e também não impacta em sua popularidade;
 - Os resultados mostram evidências de que notícias menores tendem a ser mais subjetivas. Isso pode estar associado à motivação da notícia, intuitivamente notícias que tem por objetivo informar se preocupam em fornecer mais detalhes sobre o fato noticiado, resultando em maior clareza ao leitor.

Em trabalhos futuros, pretendemos incluir mais portais de notícias (p.ex. Carta Capital) afim de cobrir todos os posicionamentos do espectro político-ideológico. Também pretendemos estender e validar nossa abordagem em outros gêneros textuais. Além disso, pretendemos incluir nossas descobertas como *features* em um sistema de recomendação de notícias, buscando reduzir a subjetividade das notícias recomendadas.

REFERENCES

- AMORIM, E., CANÇADO, M., AND VELOSO, A. Automated essay scoring in the presence of biased ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. pp. 229–237, 2018.
- ANDERSON, J. Lix and rix: Variations on a little-known readability index. *Journal of Reading* 26 (6): 490–496, 1983.
- BAE, Y. AND LEE, H. Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers. *Journal of the American Society for Information Science and Technology* 63 (12): 2521–2535, 2012.
- BENVENISTE, E. Subjectivity in language. *Problems in general linguistics* vol. 1, pp. 223–230, 1971.
- CHATURVEDI, I., CAMBRIA, E., ZHU, F., QIU, L., AND NG, W. K. Multilingual subjectivity detection using deep multiple kernel learning. *Proceedings of Knowledge Discovery and Data Mining, Sydney*, 2015.
- COLEMAN, M. AND LIAU, T. L. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60 (2): 283, 1975.
- FLAOUNAS, I., ALI, O., LANSDALL-WELFARE, T., DE BIE, T., MOSDELL, N., LEWIS, J., AND CRISTIANINI, N. Research methods in the age of digital journalism: Massive-scale automated analysis of news-content—topics, style and gender. *Digital Journalism* 1 (1): 102–116, 2013.
- GOLDBERG, B. *Bias: A CBS Insider Exposes How the Media Distort the News*. Regnery Publishing, 2001.
- JHA, V., SHREDEVI, G., SHENOY, P. D., AND VENUGOPAL, K. Generating multilingual subjectivity resources using english language. *Int. J. Comput. Appl* 152 (9): 41–47, 2016.
- KLARE, G. R. A table for rapid determination of dale-chall readability scores. *Educational Research Bulletin*, 1952.
- KUSNER, M., SUN, Y., KOLKIN, N., AND WEINBERGER, K. From word embeddings to document distances. In *International Conference on Machine Learning*. pp. 957–966, 2015.
- MIHALCEA, R., BANEJA, C., AND WIEBE, J. Learning Multilingual Subjective Language via Cross-Lingual Projections. *Proceedings of ACL* 1 (1): 14–21, 2007.
- MIKOLOV, T., LE, Q. V., AND SUTSKEVER, I. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.
- MORAES, S. M., SANTOS, A. L., REDECKER, M., MACHADO, R. M., AND MENEGUZZI, F. R. Comparing approaches to subjectivity classification: A study on portuguese tweets, 2016.
- NIGAM, S., KUMAR, N., MANDAL, N., PADMA, B., AND RAO, S. Real time ambient air quality status during diwali festival in central, india. *Journal of Geoscience and Environment Protection* vol. 4, pp. 162–172, 2016.
- SALES, A., BALBY, L., AND VELOSO, A. Media bias characterization in brazilian presidential elections. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*. HT '19. ACM, New York, NY, USA, pp. 231–240, 2019.
- WIEBE, J., WILSON, T., AND CARDIE, C. Annotating Expressions of Opinions and Emotions in Language. *Empirical Methods in Natural Language Processing* 1 (1): 164–210, 2005.
- WILSON, T., HOFFMANN, P., SOMASUNDARAN, S., KESSLER, J., WIEBE, J., CHOI, Y., CARDIE, C., RILOFF, E., AND PATWARDHAN, S. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*. pp. 34–35, 2005.
- YAQUB, U., SHARMA, N., PABREJA, R., CHUN, S., ATLURI, V., AND VAIDYA, J. Analysis and visualization of subjectivity and polarity of twitter location data. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. ACM, pp. 67, 2018.
- ZAR, J. H. Significance testing of the spearman rank correlation coefficient. *Journal of the American Statistical Association* 67 (339): 578–580, 1972.