UNIVERSIDADE DE SÃO PAULO

Escola de Artes, Ciências e Humanidades

Relatório Técnico PPgSI-005/2021

Detecção de alinhamento partidário em redes sociais: o córpus GovBR

Samuel Caetano da Silva, Ivandré Paraboni

Setembro - 2021

Série de Relatórios Técnicos

PPgSI-EACH-USP
Rua Arlindo Béttio, 1000 – Ermelino Matarazzo
03828-000 – São Paulo, SP.
TEL: (11) 3091-8197 http://www.each.usp.br/ppgsi

Detecção de alinhamento partidário em redes sociais: o córpus GovBR

Samuel Caetano da Silva, Ivandré Paraboni[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo
São Paulo – SP, Brazil

{samuel.caetano.silva,ivandre}@usp.br

Resumo. *Polarização e orientação política são fenômenos de interesse para um grande número de aplicações computacionais. Não por acaso, estes temas têm sido também o foco de estudos na área de Processamento de Línguas Naturais (PLN). Estudos deste tipo são entretanto majoritariamente desenvolvidos com base em córpus no idioma inglês, havendo uma nítida carência de recursos deste tipo para o português. Como forma de preencher esta lacuna, o presente relatório descreve a construção de um recurso linguístico-computacional básico, aqui denominado córpus GovBR, de* tweets *favoráveis e contrários ao presidente brasileiro. O córpus faz uso de um método de rotulação distante baseada em* hashtags*, e contempla um conjunto balanceado de* timelines *públicas de usuários em extremos opostos da polarização política para uso em estudos dedicados ao processamento computacional de textos deste tipo em português.*

1. Introdução

A detecção computacional de alinhamento partidário em textos pode ser entendida como a tarefa de identificar se um texto contém alguma ideologia política evidente. De forma mais específica, esta questão pode ser observada sob dois ângulos a serem discutidos na sequência, que são (i) polarização política (hiperpartidarismo), que é observável em textos politicamente enviesados defendendo ou acusando um ponto de vista específico Kiesel et al. [2019] e (ii) orientação política, que é observável em textos contendo opinião pessoal acerca de determinados assuntos (e.g. imigração, aborto, etc.) [Iyyer et al. 2014].

A polarização política (i) acontece quando há uma forte divisão entre grupos ideologicamente opostos, fazendo com que o debate político seja feito por meio de discursos hiperpartidários cujo objetivo é atacar e prejudicar outro grupo ou outra ideologia. Discursos hiperpartidários reforçam um processo de formação de grupos antagônicos na sociedade [Kiesel et al. 2019, Potthast et al. 2018].

A diferença entre um texto hiperpartidário (politicamente enviesado) e um neutro é sutil. Considere os dois exemplos de frases a seguir.

- *"(...) diz que a NFL deveria demitir jogadores que se ajoelharam durante hino nacional"* (notícia neutra)
- *"(...) vira as costas aos trabalhadores americanos"* (notícia hiperpartidária)

Nestes exemplos, é possível apontar o tom de cada texto como a diferença entre ser hiperpartidário ou neutro. A primeira frase *"(...)* ***diz*** *(...)"* possui tom neutro, enquanto a segunda *"(...)* ***vira as costas*** *(...)"* possui tom acusativo. O uso de certos termos ou expressões podem indicar se um texto é hiperpartidário Srivastava et al. [2019].

Já no caso da orientação política (ii), esta é inerente a qualquer indivíduo e é subjetiva. Muitas das opiniões expressas por pessoas acerca de determinados assuntos como aborto,

porte de armas, entre outros, contêm ideias que podem identificar sua inclinação política. As frases a seguir exemplificam casos de opinião sobre o tema redução da maioridade penal.

- *"A redução da maioridade penal, se implantada, poderia ter alguns efeitos positivos, (...)."* (opinião positiva com relação ao assunto)
- *"Ineficiente. Provavelmente não diminuiria a quantidade de crimes cometidos enquanto lotaria os presídios (...)."* (opinião negativa com relação ao assunto)

Nos exemplos apresentados, a polaridade das opiniões expressas (positivo/negativo) sobre o assunto maioridade penal pode indicar a orientação política contida no texto. Neste sentido, detectar a orientação política em texto pode implicar analisar os sentimentos contidos nas palavras desse texto com relação a um tópico de debate [Iyyer et al. 2014, Srivastava et al. 2019, dos Santos e Paraboni 2019]. Assim, com relação ao tema redução da maioridade penal, pode-se supor que um sentimento positivo indica uma orientação alinhada à direita, enquanto que uma opinião negativa indica orientação alinhada à esquerda política.

É importante, entretanto, observar que (i) polarização e (ii) orientação política são conceitos distintos. Uma opinião expressa sobre assuntos de debate social (e.g. aborto, legalização de drogas, redução da maioridade penal, etc.) não implica necessariamente em hiperpartidarismo. Entretanto, essas opiniões podem ser distorcidas e usadas como um elemento do discurso hiperpartidário [Iyyer et al. 2014, Kulkarni et al. 2018, Bhatia e P 2018].

Tarefas de Processamento de Língua Natural (PLN) como mineração, agrupamento ou classificação de textos têm recebido muita atenção por conta da disponibilidade de documentos a partir de diferentes fontes (e.g. redes sociais, portais de notícias, fichas médicas, etc.). Em especial, a classificação de textos é uma tarefa relevante por causa de suas múltiplas aplicações práticas, que incluem a análise de sentimentos e posicionamentos [Mohammad et al. 2016a, 2017, Siddiqua et al. 2019], detecção de fraude ou violações autorais [Custódio e Paraboni 2018, 2021], caracterização demográfica [dos Santos et al. 2017, Preotiuc-Pietro et al. 2017, Silva e Paraboni 2018a,b, Hsieh et al. 2018, Takahashi et al. 2018, Ramos et al. 2018, Pizarro 2019, dos Santos et al. 2020b, Rangel et al. 2020], análise discursiva [Paraboni 1997, Paraboni e de Lima 1998, Preiss 2001, Poesio et al. 2016], detecção de *cyberbulling* ou de discurso ofensivo [Wich et al. 2020], e muitas outras[1].

Polarização e orientação política, especialmente em redes sociais, são fenômenos de interesse para um grande número de aplicações computacionais e, não por acaso, têm sido o foco de estudo na área de PLN. Pela perspectiva de polarização política, detectar automaticamente textos hiperpartidários permite ao usuário ter algum conhecimento prévio do tipo de texto que irá consumir [Potthast et al. 2018, Kiesel et al. 2019]. Já sob a ótica de orientação política, identificar o posicionamento de um indivíduo permite que sistemas reconheçam de maneira automática a inclinação política do autor do texto dos Santos e Paraboni [2019], Mohammad et al. [2016b].

Problemas de polarização e orientação política são majoritariamente são dedicados ao

[1]Também estudadas no contexto da geração de língua natural em, e.g., Paraboni e van Deemter [1999], Paraboni [2003].

idioma inglês, ou seja, desenvolvidos a partir de córpus neste idioma. No caso do português, por outro lado, não foi possível identificar a disponibilidade de recursos deste tipo. Assim, como forma de preencher esta lacuna, o presente relatório descreve a construção de um recurso linguístico-computacional básico, aqui denominado córpus GovBR, de *tweets* favoráveis e contrários ao governo federal do Brasil. O córpus faz uso de um método de rotulação distante baseada em *hashtags*, e contempla um conjunto balanceado de *timelines* públicas de usuários em extremos opostos da polarização política para uso em estudos da área. No total, o conjunto de dados é constituído de cerca de 13 milhões de *tweets* e 190 milhões de tokens.

O restante deste relatório é organizado como segue,. A seção 2 sumariza uma série de estudos recentes na área de detecção de informações de natureza política a partir de textos. A seção 3 descreve a construção do córpus GovBR e apresenta algumas estatísticas descritivas. Finalmente, a seção 4 discute o uso do córpus em projetos da área de PLN.

2. Trabalhos relacionados

Esta seção apresenta alguns trabalhos relacionados à detecção de fenômenos políticos em textos. Na tabela 1, é apresentada uma visão geral dos trabalhos abordados por meio das colunas de citação ao estudo, contexto, atributos, método, gênero e idioma utilizados pelo trabalho.

Para cada trabalho na tabela 1, é apresentado o contexto político encapsulado em três grupos (i) ideologia (contendo ideologia política, posicionamento ideológico, alinhamento político, perspectiva política), (ii) polarização (contendo hiperpartidarismo, viés político) e (iii) ódio (contendo discurso ofensivo), os principais atributos utilizados (txt = textual, grafo = rede de relações), o método utilizado para classificação (RNN = rede recorrente, U = estilometria, CNN = rede convolutiva, RegLog = regressão logística, SVM = *Support Vector Machines*, RGC = rede de grafo, NPOV = *neutral point-of-view* etc.), o gênero linguístico (D = discurso/debate, N = notícias, T = Twitter, W = Wikipédia) e o idioma do córpus utilizado (In = inglês, Al = alemão).

Tabela 1. Resumo dos trabalhos relacionados

Trabalho	Contexto político	Atributos	Método	Gênero	Idioma
Iyyer et al. [2014]	ideologia	txt	RNN	D	In
Potthast et al. [2018]	polarização	txt	U	N	In
Kulkarni et al. [2018]	ideologia	txt, grafo	CNN	N	In
Bhatia e P [2018]	ideologia	txt	RegLog	N	In
Jiang et al. [2019]	polarização	txt	CNN	N	In
Srivastava et al. [2019]	polarização	txt	RegLog	N	In
Drissi et al. [2019]	polarização	txt	BERT	N	In
Bestgen [2019]	polarização	txt	SVM, RegLog	N	In
Lee et al. [2019]	polarização	txt	BERT	N	In
Patankar et al. [2019]	polarização	txt	NPOV	N	In
Stefanov et al. [2020]	alinhamento	txt, grafo	FastText	T	In
Wich et al. [2020]	ódio	txt	SHAP	T	Al
Baly et al. [2020]	ideologia	txt	LSTM, BERT	N, T	In
Feng et al. [2021]	ideologia	grafo	RGC	W	In
Li e Goldwasser [2021]	polarização	txt	biLSTM	N	In
Grimminger e Klinger [2021]	ódio	txt	BERT	T	In

A maioria dos estudos apresentados são de 2019, uma justificativa é a própria competição SemEval 2019 que estimulou estudos abordando o tema e proporcionou córpus em comum para todos os participantes. O estudo em Iyyer et al. [2014], mostra como o tema de detecção de fenômenos políticos em textos é relevante desde 2014, intensificando-se em anos recentes.

Tendo-se em vista que parte dos estudos apresentados são referentes à competição Kiesel et al. [2019], a abordagem mais comum é a de hiperpartidarismo, entretanto outros trabalhos abordam fenômenos políticos diferentes, quase todos tendo a política norte-americana como pano de fundo. O idioma Inglês é a língua mais abordada por esses trabalhos.

De acordo com os resumos apresentados, é possível sumarizar que o fenômeno político mais evidente é o de hiperpartidarismo, em virtude da competição SemEval 2019, outros fenômenos são posicionamento e alinhamento político. A maioria dos trabalhos tratam do idioma Inglês e demais línguas como, em especial, o Português, carecem de trabalhos semelhantes aos citados.

Além disso, alguns pontos levantados por meio desta revisão é que combinar diferentes fontes de conhecimento e realizar a classificação a nível de sentença parecem ser alternativas na tarefa de detecção de fenômenos políticos em textos.

Diversas técnicas são utilizadas para gerar representações textuais dos documentos, sendo o modelo de língua pré-treinado BERT o mais comum. As técnicas de classificação mais comuns foram baseadas em redes neurais como CNN e LSTM, além disso, também houve o uso de técnicas mais tradicionais como Regressão Logística e *Support Vector Machines*.

3. O córpus GovBR

GovBR é um recurso linguístico-computacional ao estilo de outras iniciativas do gênero para o PLN em português já desenvolvidas pelo grupo responsável, como os córpus Stars [Teixeira et al. 2014] e Stars2 [Paraboni et al. 2017], utilizados predominantemente nas tarefas de seleção de conteúdo [Paraboni 2000, Paraboni e van Deemter 2002, dos Santos Silva e Paraboni 2015] e realização superficial [de Lucena et al. 2010, Pereira e Paraboni 2008] de língua natural, ou ainda como no caso do córpus SetembroBR para detecção de transtornos de saúde mental em redes sociais [dos Santos et al. 2020a].

O córpus consiste de uma coleção de *timelines* públicas de usuários Twitter que fazem uso de determinadas *hashtags* indicativas de um posicionamento favorável ou contrário ao atual presidente do Brasil. Estas *hashtags*, selecionadas a partir de tópicos frequentes na plataforma Twitter e complementadas com expressões similares, são relacionadas a seguir.

- **Favoráveis:**
- RespeitemOPresidente
- JairNaoCaiNemaPau
- ForçaBolsonaro
- FechadoComBolsonaroAte2026
- fechadocombolsonaro
- DeusBrasilBolsonaro
- bolsonarotemrazao
- BolsonaroOrgulhoDoBrasil
- BolsonaroEstamosContigo
- Bolsonaro2022

- **Contrárias:**
- stopbolsonaromundial
- StopBolsonaro
- Somos70porcento
- ImpeachmentDeBolsonaro
- forabozo
- ForaBolsonaroGenocida
- ForaBolsonaroEseuBandodeCriminosos
- forabolsonaro
- EleNao
- BolsonaroVagabundo
- bolsonarotraidor
- bolsonarogenocida
- BolsonaroEnloqueceu
- BolsonaroAssassino
- BolsonaroAcabou
- BicudaNoBozo

Neste conjunto de *hashtags*, é importante observar que o uso de uma determinada expressão não necessariamente representa um posicionamento específico, podendo inclusive representar o seu oposto (por exemplo, se empregada em sentido irônico). O presente

método de rotulação, que pode ser visto como uma forma de supervisão distante, é assim apenas uma aproximação do sentido real do texto, é há portanto uma certa margem de imprecisão.

Para cada indivíduo que publicou ao menos um *tweet* contendo uma *hashtag* de interesse, foram coletados todos os seus *tweets* públicos até o limite máximo de 3200 *tweets* por usuário permitido pela plataforma. A partir das publicações coletadas, foram removidas aquelas que não eram escritas em português, os *retweets* e as mensagens com menos de cinco caracteres de extensão. Além disso, foram removidas as *hashtags* propriamente ditas, e as menções @ foram convertidas em nomes (sem @). Finalmente, *timelines* contendo menos de 100 *tweets* válidos foram desconsideradas.

Para a composição final do córpus, foram excluídos todos os usuários com posicionamento ambivalente (ou seja, que tinham simultaneamente publicações contendo *hashtags* favoráveis e contrárias). Além disso, como forma de obter um conjunto balanceado, foram mantidos 2726 usuários de cada classe. A Tabela 2 apresenta estatísticas descritivas do conjunto de posicionamentos rotulados do córpus.

Tabela 2. Estatísticas descritivas dos posicionamentos rotulados.

posição	usuários	tweets	tokens
contrária	2726	6.826.071	96.931.824
favorável	2726	6.714.740	93.828.262
total	5452	13.540.811	190.760.086

4. Considerações

Este documento descreveu a construção do corpus GovBR de *tweets* potencialmente contrários e favoráveis ao atual governo brasileiro, e algumas de suas características. O córpus foi coletado por meio de rotulação distante baseada em *hashtags* populares na plataforma Twitter, e constitui um conjunto de dados textuais em português de proporções consideráveis, da ordem de 190 milhões de tokens.

O córpus GovBR encontra-se atualmente em uso em estudos de classificação e predição de posicionamentos a partir de dados textuais [Pavan et al. 2020], e de detecção automática de orientação política, discurso de ódio e outros. Uma abordagem típica adotada em estudos deste tipo consiste em explorar apenas o subconjunto de tweets diretamente associados a estes posicionamentos políticos, desprezando-se assim boa parte dos dados coletados.

A disponibilidade das *timelines* completas de cada usuário do córpus GovBR, entretanto, pode também permitir que sejam exploradas abordagens centradas no indivíduo, ou seja, buscando classificar seu alinhamento político geral (ou fenômeno correlato) com base em mensagens de diferentes graus de teor político (ou mesmo sem teor político nenhum). Um estudo desta natureza é deixado como sugestão de possível trabalho futuro.

Referências

Baly, R., Martino, G. D. S., Glass, J., e Nakov, P. (2020). We can detect your bias: Predicting the political ideology of news articles. Em *Proceedings of the 2020 Conference*

on Empirical Methods in Natural Language Processing (EMNLP), páginas 4982–4991, Online. Association for Computational Linguistics.

Bestgen, Y. (2019). Tintin at SemEval-2019 task 4: Detecting hyperpartisan news article with only simple tokens. Em *Proceedings of the 13th International Workshop on Semantic Evaluation*, páginas 1062–1066, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Bhatia, S. e P, D. (2018). Topic-specific sentiment analysis can help identify political ideology. Em *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, páginas 79–84, Brussels, Belgium. Association for Computational Linguistics.

Custódio, J. E. e Paraboni, I. (2018). EACH-USP ensemble cross-domain authorship attribution. Em *Working Notes Papers of the Conference and Labs of the Evaluation Forum (CLEF-2018) vol.2125*, Avignon, France.

Custódio, J. E. e Paraboni, I. (2021). Stacked authorship attribution of digital texts. *Expert Systems with Applications*, 176:114866.

de Lucena, D. J., Paraboni, I., e Pereira, D. B. (2010). From semantic properties to surface text: The generation of domain object descriptions. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, 14(45):48–58.

dos Santos, V. G., Paraboni, I., e Silva, B. B. C. (2017). Big five personality recognition from multiple text genres. Em *Text, Speech and Dialogue (TSD-2017) Lecture Notes in Artificial Intelligence vol. 10415*, páginas 29–37, Prague, Czech Republic. Springer-Verlag.

dos Santos, W. R., Funabashi, A. M. M., e Paraboni, I. (2020a). Searching Brazilian Twitter for signs of mental health issues. Em *12th International Conference on Language Resources and Evaluation (LREC-2020)*, páginas 6113–6119, Marseille, France. ELRA.

dos Santos, W. R. e Paraboni, I. (2019). Moral Stance Recognition and Polarity Classification from Twitter and Elicited Text. Em *Recents Advances in Natural Language Processing (RANLP-2019)*, páginas 1069–1075, Varna, Bulgaria.

dos Santos, W. R., Ramos, R. M. S., e Paraboni, I. (2020b). Computational personality recognition from facebook text: psycholinguistic features, words and facets. *New Review of Hypermedia and Multimedia*, 25(4):268–287.

dos Santos Silva, D. e Paraboni, I. (2015). Generating spatial referring expressions in interactive 3D worlds. *Spatial Cognition & Computation*, 15(03):186–225.

Drissi, M., Segura, P. S., Ojha, V., e Medero, J. (2019). Harvey mudd college at SemEval-2019 task 4: The clint buchanan hyperpartisan news detector. Em *Proceedings of the 13th International Workshop on Semantic Evaluation*, páginas 962–966, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Feng, S., Chen, Z., Yu, P., e Luo, M. (2021). Encoding heterogeneous social and political context for entity stance prediction. *arXiv preprint arXiv:2108.03881*.

Grimminger, L. e Klinger, R. (2021). Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. Em *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, páginas 171–180, Online. Association for Computational Linguistics.

Hsieh, F. C., Dias, R. F. S., e Paraboni, I. (2018). Author profiling from facebook corpora.

Em *11th International Conference on Language Resources and Evaluation (LREC-2018)*, páginas 2566–2570, Miyazaki, Japan. ELRA.

Iyyer, M., Enns, P., Boyd-Graber, J., e Resnik, P. (2014). Political ideology detection using recursive neural networks. Em *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 1113–1122, Baltimore, Maryland. Association for Computational Linguistics.

Jiang, Y., Petrak, J., Song, X., Bontcheva, K., e Maynard, D. (2019). Team bertha von suttner at semeval-2019 task 4: Hyperpartisan news detection using elmo sentence representation convolutional network. Em *Proceedings of the 13th International Workshop on Semantic Evaluation*, páginas 840–844, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., Stein, B., e Potthast, M. (2019). SemEval-2019 task 4: Hyperpartisan news detection. Em *Proceedings of the 13th International Workshop on Semantic Evaluation*, páginas 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Kulkarni, V., Ye, J., Skiena, S., e Wang, W. Y. (2018). Multi-view models for political ideology detection of news articles. Em *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, páginas 3518–3527, Brussels, Belgium. Association for Computational Linguistics.

Lee, N., Liu, Z., e Fung, P. (2019). Team yeon-zi at SemEval-2019 task 4: Hyperpartisan news detection by de-noising weakly-labeled data. Em *Proceedings of the 13th International Workshop on Semantic Evaluation*, páginas 1052–1056, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Li, C. e Goldwasser, D. (2021). Using social and linguistic information to adapt pretrained representations for political perspective identification. Em *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, páginas 4569–4579, Online. Association for Computational Linguistics.

Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., e Cherry, C. (2016a). A dataset for detecting stance in tweets. Em *10th Language Resources and Evaluation Conference (LREC-2016)*, Portoroz, Slovenia.

Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., e Cherry, C. (2016b). Semeval-2016 Task 6: Detecting Stance in Tweets. Em *International Workshop on Semantic Evaluation*.

Mohammad, S. M., Sobhani, P., e Kiritchenko, S. (2017). Stance and sentiment in tweets. *ACM Transactions on Internet Technology on Argumentation in Social Media*, 17(3).

Paraboni, I. (1997). Uma arquitetura para a resolução de referências pronominais possessivas no processamento de textos em língua portuguesa. Master's thesis, PUCRS, Porto Alegre.

Paraboni, I. (2000). An algorithm for generating document-deictic references. Em *Procs. of workshop Coherence in Generated Multimedia, associated with First Int. Conf. on Natural Language Generation (INLG-2000), Mitzpe Ramon*, páginas 27–31.

Paraboni, I. (2003). *Generating references in hierarchical domains: the case of Document Deixis*. Tese de Doutorado , University of Brighton.

Paraboni, I. e de Lima, V. L. S. (1998). Possessive pronominal anaphor resolution in Portuguese written texts. Em *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, páginas 1010–1014. Association for Computational Linguistics.

Paraboni, I., Galindo, M., e Iacovelli, D. (2017). Stars2: a corpus of object descriptions in a visual domain. *Language Resources and Evaluation*, 51(2):439–462.

Paraboni, I. e van Deemter, K. (1999). Issues for the generation of document deixis. Em *Procs. of workshop on Deixis, Demonstration and Deictic Belief in Multimedia Contexts, in association with the 11th European Summers School in Logic, Language and Information (esslli99)*, páginas 44–48.

Paraboni, I. e van Deemter, K. (2002). Generating easy references: the case of document deixis. Em *INLG-2002, New York*, páginas 113–119.

Patankar, A., Bose, J., e Khanna, H. (2019). A bias aware news recommendation system. Em *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, páginas 232–238. IEEE.

Pavan, M. C., dos Santos, W. R., e Paraboni, I. (2020). Twitter Moral Stance Classification using Long Short-Term Memory Networks. Em *9th Brazilian Conference on Intelligent Systems (BRACIS). LNAI 12319*, páginas 636–647. Springer.

Pereira, D. B. e Paraboni, I. (2008). Statistical surface realisation of Portuguese referring expressions. Em *Gotal-2008, Lecture Notes in Artificial Intelligence 5221*, páginas 383–392, Gothenburg, Sweden. Springer-Verlag.

Pizarro, J. (2019). Using N-grams to detect Bots on Twitter. Em Cappellato, L., Ferro, N., Losada, D., e Müller, H., editores , *CLEF 2019 Labs and Workshops, Notebook Papers*, página 10. CEUR-WS.org.

Poesio, M., Stuckardt, R., e Versley, Y., editores (2016). *Anaphora Resolution - Algorithms, Resources, and Applications*. Theory and Applications of Natural Language Processing. Springer.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., e Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. Em *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 231–240, Melbourne, Australia. Association for Computational Linguistics.

Preiss, J. (2001). Anaphora resolution with word sense disambiguation. Em *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, páginas 143–146, Toulouse, France. Association for Computational Linguistics.

Preotiuc-Pietro, D., Liu, Y., Hopkins, D., e Ungar, L. (2017). Beyond binary labels: Political ideology prediction of twitter users. Em *55th Annual Meeting of the Association for Computational Linguistics*, páginas 729–740, Vancouver. Association for Computational Linguistics.

Ramos, R. M. S., Neto, G. B. S., Silva, B. B. C., Monteiro, D. S., Paraboni, I., e Dias, R. F. S. (2018). Building a corpus for personality-dependent natural language understanding and generation. Em *11th International Conference on Language Resources and Evaluation (LREC-2018)*, páginas 1138–1145, Miyazaki, Japan. ELRA.

Rangel, F., Rosso, P., Zaghouani, W., e Charfi, A. (2020). Fine-grained analysis of language varieties and demographics. *Natural Language Engineering*, página 1–21.

Siddiqua, U. A., Chy, A. N., e Aono, M. (2019). Tweet stance detection using an attention based neural ensemble model. Em *NAACL-HLT 2019*, páginas 1868–1873, Minneapolis, USA.

Silva, B. B. C. e Paraboni, I. (2018a). Learning personality traits from Facebook text. *IEEE Latin America Transactions*, 16(4):1256–1262.

Silva, B. B. C. e Paraboni, I. (2018b). Personality recognition from Facebook text. Em *13th International Conference on the Computational Processing of Portuguese (PROPOR-2018) LNCS vol. 11122*, páginas 107–114, Canela. Springer-Verlag.

Srivastava, V., Gupta, A., Prakash, D., Sahoo, S. K., R.R, R., e Kim, Y. H. (2019). Vernonfenwick at SemEval-2019 task 4: Hyperpartisan news detection using lexical and semantic features. Em *Proceedings of the 13th International Workshop on Semantic Evaluation*, páginas 1078–1082, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Stefanov, P., Darwish, K., Atanasov, A., e Nakov, P. (2020). Predicting the topical stance and political leaning of media using tweets. Em *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, páginas 527–537, Online. Association for Computational Linguistics.

Takahashi, T., Tahara, T., Nagatani, K., Miura, Y., Taniguchi, T., e Ohkuma, T. (2018). Text and image synergy with feature cross technique for gender identification. Em *Working Notes Papers of the Conference and Labs of the Evaluation Forum (CLEF-2018) vol.2125*, página 12, Avignon, France.

Teixeira, C. V. M., Paraboni, I., da Silva, A. S. R., e Yamasaki, A. K. (2014). Generating relational descriptions involving mutual disambiguation. Em *Computational Linguistics and Intelligent Text Processing (CICLing-2014), Lecture Notes in Computer Science 8403*, páginas 492–502, Kathmandu, Nepal. Springer.

Wich, M., Bauer, J., e Groh, G. (2020). Impact of politically biased data on hate speech classification. Em *Proceedings of the Fourth Workshop on Online Abuse and Harms*, páginas 54–64, Online. Association for Computational Linguistics.