Abordagens Baseadas em Léxicos para a Classificação de Sentimentos Orientada aos Alvos de Opinião em Comentários do Domínio Político

Lucas Lazarini[1], **Fábio S. Igarashi Anno** [1], **Eloize R. Marques Seno**[1], **Helena M. Caseli**[2]

[1]Instituto Federal de São Paulo (IFSP) – São Carlos, SP

[2]Departamento de Ciência da Computação – Universidade Federal de São Carlos (UFSCar) São Carlos, SP

{lazarini.lucas, fabio.seyiji}@aluno.ifsp.edu.br,

eloize@ifsp.edu.br, helenacaseli@ufscar.br

Resumo. *O enorme volume de textos opinativos produzido nas mídias sociais têm levado a uma busca cada vez maior por algoritmos capazes de analisar os sentimentos de pessoas em relação à produtos, entidades políticas, etc. Muitos modelos de análise de sentimento (AS) foram propostos para o português nos últimos anos. Contudo, a maioria deles consiste em analisar o sentimento geral de uma sentença, não considerando, portanto, o sentimento individual relacionado a cada alvo de opinião no texto. Dado este contexto, este artigo investigou o uso de léxicos de sentimentos na classificação de sentimento orientado ao alvo de opinião em comentários sobre debate político em português.*

1. Introdução

O crescimento explosivo das mídias sociais tem possibilitado aos seus usuários expressarem com facilidade opiniões e emoções/sentimentos sobre produtos, eventos políticos, indivíduos etc. por meio de *blogs*, fóruns de discussão e redes sociais. Como consequência, algoritmos capazes de analisar o sentimento e a opinião pública compartilhada por meio de textos opinativos têm ganhado cada vez mais importância. A Análise de Sentimentos (AS) é uma subárea do Processamento de Língua Natural (PLN), que utiliza ferramentas e recursos linguísticos-computacionais para analisar sentimentos em textos a respeito de indivíduos, produtos, serviços, entre outros [Liu and Zhang 2012].

Nos últimos anos muitos trabalhos foram proposto na literatura sobre AS. Contudo, no que se refere à língua portuguesa, a maior parte deles consiste em analisar o sentimento em relação à polaridade (positiva, negativa ou neutra) geral conduzida por um documento ou sentença (vide, por exemplo, [Chaves et al. 2012, França and Oliveira 2014, Capellaro and Caseli 2021]), não considerando as opiniões particulares relacionadas a cada entidade presente no texto, chamadas de alvos de opinião. Por exemplo, no comentário "Foi de facto um debate cordato, civilizado, em que Jerónimo se mostrou um senhor e o Louçã meteu a viola no saco.", extraído do córpus utilizado nesta pesquisa (vide Seção 3), o sentimento em relação ao alvo "Jerónimo" é positivo, enquanto o sentimento em relação ao alvo "Louçã" é negativo.

Dado o contexto apresentado, este trabalho tem por objetivo investigar o uso de abordagens baseadas em léxicos na classificação de polaridade orientada para os alvos

de opinião em comentários sobre debate político em português. A escolha do domínio político é motivada pela quantidade limitada de trabalhos nesse domínio para o português.

O restante deste artigo está organizado da seguinte forma. A seção 2 descreve os principais trabalhos da literatura relacionados a este. Na seção 3 é apresentada a metodologia que está sendo adotada no desenvolvimento desta pesquisa (ainda em andamento), bem como os recursos e ferramentas linguísticos computacionais usados. Na seção 4 são apresentados alguns resultados preliminares. Por fim, a seção 5 apresenta algumas conclusões do trabalho.

2. Trabalhos Relacionados

Na literatura a análise de sentimentos (AS) tem sido aplicada a documentos, sentenças e aspectos [Medhat et al. 2014, Schouten and Frasincar 2016, Pereira 2021]. No nível de documentos, o sentimento é atribuído ao documento como um todo. No nível de sentenças, o sentimento é atribuído a cada sentença de um documento, enquanto no nível de aspectos os sentimentos relacionados a aspectos/atributos específicos das entidades mencionadas no texto (alvos de opinião) são identificados e posteriormente classificados de acordo com a opinião ou emoção. Por exemplo, em um comentário sobre um produto qualquer (por exemplo, *smartphone*) o produto em si costuma ser o alvo de opinião, enquanto os atributos/características relacionadas ao produto são os aspectos (como preço, qualidade da câmera e duração da bateria). Neste artigo, o foco de interesse é a AS orientada aos alvos de opinião em textos. Ou seja, uma tarefa que se aproxima da AS em nível sentencial, porém é mais específica e desafiadora do que a análise geral do sentimento expresso em uma sentença. Por outro lado, é uma tarefa menos refinada do que a análise de sentimentos baseada em aspectos.

Uma abordagem clássica de AS usada na literatura baseia-se no uso de léxicos de sentimentos, ou seja, dicionários que contêm palavras anotadas com suas respectivas polaridades. Esse tipo de abordagem utiliza a contagem e/ou a soma das polaridades das palavras presentes no texto para determinar sua polaridade geral (vide [Taboada et al. 2011, Liu and Zhang 2012, Costa and Pardo 2020]). Além das abordagens baseadas em léxicos, técnicas baseadas em aprendizado de máquina também têm se destacado em análise de sentimentos, incluindo tanto as técnicas tradicionais de aprendizado supervisionado e não supervisionado (por exemplo, [França and Oliveira 2014, Cristiani et al. 2020, Capellaro and Caseli 2021, Jain et al. 2021]) como técnicas mais sofisticadas baseadas em aprendizado profundo (*deep learning*) (por exemplo, [Zhang et al. 2018, Souza and Oliveira e Souza Filho 2022]). Abordagens híbridas combinando várias técnicas também são comuns [Appel et al. 2016].

Em [Carvalho et al. 2017], por exemplo, os autores compararam três diferentes classificadores (Naive Bayes, SVM e MaxEnt) e três métodos de seleção de atributos na classificação de polaridade em parágrafos de textos relacionados às eleições de 2014 no Brasil. O melhor classificador (MaxEnt) obteve uma acurácia em torno de 85%. Já o modelo de [Capellaro and Caseli 2021], baseado no BERT pré-treinado para o português, alcançou um *F1-score* de 96,6% na classificação de polaridade associada ao sentimento geral de *tweets* relacionados às eleições de 2018 no Brasil.

Se por um lado as abordagens baseadas em aprendizado de máquina podem levar a resultados bastante precisos na classificação de polaridade de sentimentos de textos para

o domínio no qual ele foi treinado, por outro lado esses modelos costumam ter um desempenho muito ruim quando aplicados a outros domínios [Taboada et al. 2011]. Nesse sentido, as abordagens baseadas em léxicos de sentimentos e as abordagens híbridas que combinam esse tipo de recurso com métodos de aprendizado de máquina são interessantes e podem resultar em modelos menos dependentes de domínio. Nesta pesquisa, ainda em andamento, iniciamos a investigação com o uso de abordagens baseadas em léxicos, para no futuro tratar a tarefa de análise de polaridade como um problema de aprendizado de máquina.

3. Metodologia de Desenvolvimento

A metodologia de desenvolvimento deste trabalho é baseada em corpus de comentários do domínio político e no uso de léxicos de sentimentos. Mais especificamente, como corpus de trabalho foi escolhido o SentiCorpus-PT [Carvalho et al. 2011], composto por comentários sobre debates televisivos referentes às eleições de 2009 do Parlamento Português. O corpus contêm 1.082 comentários, totalizando 3.868 sentenças. Cada sentença no corpus pode ter diferentes alvos de opinião. Os alvos de opinião são entidades humanas, nomeadamente políticos, organizações políticas (geralmente utilizadas para se referir aos seus membros), personalidades da mídia (por exemplo, jornalistas) ou usuários (comentadores). Cada sentença dispõe de anotações sobre cada alvo de opinião mencionado na sentença e a polaridade relacionada a cada alvo. A polaridade varia de -2 (o valor negativo mais forte) até 2 (o valor positivo mais forte). 94,3% das sentenças possuem pelo menos um alvo anotado, sendo que a maioria delas (79%) tem exatamente um alvo de opinião.

Como léxicos de sentimentos tem sido utilizado o LIWC [Balage Filho et al. 2013] e o SentiLex-PT [Carvalho and Silva 2015]. O LIWC é um léxico geral do português composto por 127.149 instancias organizadas em categorias. As categorias *posemo* e *negemo* indicam polaridade positiva e negativa, respectivamente. O SentiLex-PT, por sua vez, é um léxico de sentimentos sobre entidades humanas. Ele é composto por 7.014 lemas e 82.347 formas flexionadas organizadas em adjetivos, substantivos, verbos e expressões idiomáticas.

Para o pré-processamento do corpus foi utilizada a biblioteca Python spaCy[1] com o modelo português "pt_core_news_lg". O pré-processamento consistiu nas seguintes etapas: (i) tokenização, (ii) lematização, (iii) extração de *PoS - Part of Speech* e (iv) análise sintática de dependência.

Após o pré-processamento do córpus, um primeiro estudo foi realizado para atribuir a polaridade de sentimento (positiva, negativa ou neutra) para cada sentença em relação a cada alvo de opinião presente na mesma. Mais especificamente, a estratégia usada consistiu em atribuir ao alvo de opinião do comentário a polaridade resultante da soma das polaridades das palavras presentes no léxico (LIWC ou SentiLex-PT). Para cada léxico, duas diferentes estratégias foram usadas: a primeira considera apenas as palavras associadas ao alvo do comentário, via dependência sintática e a segunda que considera o comentário como um todo.

[1]https://spacy.io/models/pt (acesso em: 02/07/2023).

4. Resultados

Table 1. Resultados obtidos por cada estratégia.

Estratégia	Precisão	Cobertura	Medida-F
SL	45,1%	31,2%	**36,9%**
LW	36,6%	26,8%	30,9%
SL+LW	39,8%	**34,0%**	**36,7%**
SL-DEP	**55,1%**	5,5%	10,1%
LW-DEP	44,5%	5,1%	9,1%
SL+LW-DEP	49,5%	7,5%	13,0%

Em um experimento preliminar, cada estratégia de classificação de polaridade foi avaliada em termos de Precisão, Cobertura e Medida-F. A Tabela 1 resume os resultados obtidos com cada estratégia. Em termos de Precisão, o melhor valor (55,1%) foi obtido pela estratégia SL-DEP que se baseia no uso do SentiLex e nas relações de dependência sintática para associar cada alvo de opinião a cada palavra de sentimento identificada pelo léxico. Porém, as três estratégias que usam dependência sintática apresentaram os piores valores de Cobertura e, consequentemente, os piores desempenhos globais medidos em termos de Medida-F. Uma possível explicação para isso pode estar relacionada ao fato de que vários comentários no córpus não apresentam a estrutura esperada de sujeito-verbo-objeto, o que pode ter impactado na qualidade da análise de dependência gerada, fazendo com que a palavra com polaridade não fosse corretamente associada ao alvo do comentário. O melhor desempenho termos de Cobertura (34%) foi obtido ao combinar o uso do SentiLex e do LIWC em um mesma estratégia. No que se refere ao desempenho global, a estratégia baseada apenas no uso do SentiLex e a estratégia que combinou o uso dos dois léxicos (SL+LW) obtiveram os melhores desempenhos, ou seja, 36,9% e 36,7%, respectivamente. Contudo, testes estatísticos precisam ser realizados a fim de confirmar se essa pequena diferença entre essas duas abordagens é estatisticamente significativa.

5. Conclusões

Este trabalho descreveu um estudo preliminar com o objetivo de avaliar o uso de léxicos de sentimentos na classificação de sentimento (polaridade) relacionado a cada alvo de opinião em comentários sobre debate político em português. A tarefa de atribuir polaridade orientada aos alvos de opinião é mais desafiadora do que a classificação de sentimento geral de um comentário.

Como próximos passos deste trabalho pretende-se investigar outros léxicos disponíveis para o português e o uso de heurísticas definidas a partir de corpus que permitam associar palavras que expressam sentimento a cada alvo de opinião no texto. Pretende-se, ainda, combinar o uso de léxicos de sentimentos com técnicas de aprendizado de máquina.

References

Appel, O., Chiclana, F., Carter, J., and Fujita, H. (2016). A hybrid approach to the sentiment analysis problem at the sentence level. *Knowledge-Based Systems*, 108:110–124. New Avenues in Knowledge Bases for Natural Language Processing.

Balage Filho, P. P., Pardo, T. A. S., and Aluísio, S. M. (2013). An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 215–219.

Capellaro, L. and Caseli, H. M. (2021). Análise de polaridade e de tópicos em tweets no domínio da política no Brasil. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 47–55, Porto Alegre, RS, Brasil. SBC.

Carvalho, C. M. A., Nagano, H., and Barros, A. K. (2017). A comparative study for sentiment analysis on election Brazilian news. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 103–111, Uberlândia, Brazil. Sociedade Brasileira de Computação.

Carvalho, P., Sarmento, L., Teixeira, J., and Silva, M. J. (2011). Liars and saviors in a sentiment annotated corpus of comments to political debates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 564–568, Portland, Oregon, USA. Association for Computational Linguistics.

Carvalho, P. and Silva, M. (2015). SentiLex-PT: Principais características e potencialidades. *Linguística, Informática e Tradução: Mundos que se Cruzam, Oslo Studies in Language*, 7(1):425–438.

Chaves, M., Freitas, L., Souza, M., and Vieira, R. (2012). Pirpo: An algorithm to deal with polarity in portuguese online reviews from the accommodation sector. volume 7337, pages 296–301.

Costa, R. and Pardo, T. (2020). Métodos baseados em léxico para extração de aspectos de opiniões em português. In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 61–72, Porto Alegre, RS, Brasil. SBC.

Cristiani, A., Lieira, D., and Camargo, H. (2020). A sentiment analysis of brazilian elections tweets. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*, pages 153–160, Porto Alegre, RS, Brasil. SBC.

França, T. and Oliveira, J. (2014). Análise de sentimento de tweets relacionados aos protestos que ocorreram no Brasil entre junho e agosto de 2013. In *Anais do III Brazilian Workshop on Social Network Analysis and Mining*, pages 128–139, Porto Alegre, RS, Brasil. SBC.

Jain, P. K., Pamula, R., and Srivastava, G. (2021). A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. *Computer Science Review*, 41:100413.

Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. *Mining Text Data*, pages 415–463.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: a survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.

Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115.

Schouten, K. and Frasincar, F. (2016). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830.

Souza, F. D. and Oliveira e Souza Filho, J. B. (2022). BERT for sentiment analysis: Pre-trained and fine-tuned alternatives. In Pinheiro, V., Gamallo, P., Amaro, R., Scarton, C., Batista, F., Silva, D., Magro, C., and Pinto, H., editors, *Computational Processing of the Portuguese Language*, pages 209–218, Cham. Springer International Publishing.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37:267–307.

Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis : A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8.