Criação automatizada de dicionário léxico para mineração de opiniões

Aluno: Airton Bordin Junior

Subárea do conhecimento do CNPq: Metodologia e Técnicas de Computação.
Subsubárea do conhecimento: Banco de Dados
Linha de pesquisa: Inteligência Computacional
Cidade/Estado: Goiânia/GO

1 Contextualização

[Liu, 2010] apresenta uma visão multifacetada sobre a mineração de opiniões. Neste trabalho, o autor conceitualiza o problema e propõe uma forma estruturada de organização dos dados não estruturados, característica instínseca dos textos em linguagem natural, objeto de entrada da pesquisa. A definição de opinião como uma quíntupla (entidade, aspecto da entidade, sentimento, autor e tempo) é utilizada em grande parte dos trabalhos na área, caracterizando-se, portanto, como elemento fundamental nas pesquisas sobre o assunto. Visão geral sobre o tema e principais desafios e técnicas são vistos também em [Mohammad, 2016], [Ghaleb and Vijendran, 2016].

[Taboada et al., 2011] aborda a análise e mineração de opiniões baseada em dicionários léxicos. Apresenta o SO-CAL (Semantic Orientation Calculator), que usa lista de palavras já consolidadas para a geração de dicionários com novas entradas e suas polaridades de forma não supervisionada. Durante a descrição do trabalho, apresenta conceitos de intensificação e negação, amplamente utilizadas nas técnicas de geração de novos léxicos. Apesar de ser feita de forma automática, o autor utilizou uma etapa de verificação humana para a validação da consistência das palavras geradas pela técnica, fazendo uso de um serviço de *Mechanical Turk* da Amazon.

Na mesma linha, [Eisenstein, 2016] e [Bandhakavi et al., 2016] apresentam técnicas de análise de opiniões fazendo uso de dicionários léxicos. O primeiro apresenta uma abordagem usando a técnica de *Naive Bayes* para a classificação dos aspectos e cita problemas de estimativas de palavras e avaliação dos léxicos criados. O segundo faz uma comparação de algumas técnicas de avaliação em 4 conjuntos de dados diferentes, apresentando uma análise quantitativa do mesmo. Abordagens e comparações semelhantes, com algumas modificações no domínio e no idioma do problema abordado, podem ser vistos em [Khoo and Johnkhan, 2017], [Asghar et al., 2014] e [Ding et al., 2008].

A criação automatizada de dicionários léxicos, tema central do presente trabalho, é tratada em sua forma geral em [Widdows and Dorow, 2002] e [Duwairi et al., 2015]. O primeiro utiliza uma estratégia de criação e análise de uma estrutura de grafos, por meio uma base padronizada de palavras semente, que contém diversas entradas previamente avaliadas em suas polaridades e, também, a descrição de seus sinônimos. Apesar de fazer uma abordagem focada em substantivos, que representam os vértices do grafo, a ideia principal pode ser utilizada em outras estratégias de geração léxica automatizada que incorporem verbos, adjetivos, entre outros. [Duwairi et al., 2015] dá uma visão geral da criação de um dicionário de palavras, usando como base *tweets* em árabe. Importante destaque desse último foi a inclusão de *emoticons* na análise, característica amplamente utilizada, principalmente, em escritas informais na Internet.

A maior parte das estratégias de criação de dicionários léxicos utiliza como base de palavras semente o banco de dados *WordNet* - disponível em https://wordnet.princeton.edu/ - que fornece uma lista de palavras, sua polaridade e seus sinônimos. Importante destacar, também, que as bases utilizadas nos trabalhos supracitados consideram palavras no idioma inglês. Mesmo os trabalhos que utilizam estratégias em idiomas diferentes fizeram uso dessas bases por meio de um processo de tradução automatizada.

2 Problema

Como se pode observar, a análise de dados não estruturados - como o Processamento de Linguagem Natural (PLN) - é uma linha de pesquisa abrangente e que vem sendo tema de diversos trabalhos nos últimos anos, principalmente devido ao aumento no número de usuários de Internet e o consequente crescimento da produção de conteúdo na rede, como opiniões, avaliações, entre outros.

Nesse sentido, a mineração de opiniões, muitas vezes chamada de análise de sentimento, apresenta-se como uma das principais linhas de pesquisa no contexto de PLN. Essa subárea tem como principal desafio a análisa de opiniões descritas em linguagem natural para a identificação da polaridade implícita ou explícita no texto. Essa polaridade é, na maior parte das vezes, identificada como uma escala de pontuação de sua característica positiva, negativa ou neutra.

Como visto no capítulo anterior, uma das principais técnicas da análise de opiniões é a realizada por meio de dicionários de dados. Esses dicionários contêm palavras previamente avaliadas por especialistas humanos, principalmente quanto à sua polaridade. Algumas bases, como a *WordNet* - utilizada na maior parte dos trabalhos - apresenta, também, os sinônimos para cada palavra.

Porém, é evidente a limitação inerente à estratégia de utilização do dicionário léxico - a própria lista de palavras disponíveis. Esse fato muitas vezes limita a realização de uma análise mais profunda sobre determinado contexto. Nesse sentido, um dos principais desafios na área de mineração de opiniões é a criação do dicionário léxico de forma não supervisionada, tema central do presente trabalho. A maior parte desses dicionários é feita de forma manual, fato que caracteriza uma limitação óbvia para a maior parte dos contextos e domínios.

Consciente desse problema de pesquisa, a ideia principal do presente trabalho é a criação de um processo automatizado de avaliação de dados inéditos (nesse contexto, palavras que não estão contidas no dicionário) e suas polaridades (negativa, positiva ou neutra) que consequentemente comporão um novo dicionário léxico. Todo esse processo será feito a partir de um conjunto mínimo de palavras contidas em algum dicionário disponibilizado publicamente, como o amplamente citado e utilizado WordNet, por exemplo.

Devido à característica intrínseca do próprio problema, serão utilizadas na pesquisa bases de textos em inglês. Ao mesmo tempo, muitas técnicas utilizadas durante o trabalho também poderão ser utilizadas para resolver problemas em português, com as devidas alterações.

Espera-se que esses padrões possam ser utilizados como entrada de processos de avaliação em diversas áreas de PNL e mineração de opiniões como, por exemplo, análise de sentimentos em redes sociais. Algumas dessas pesquisas são realizadas na própria instituição, apoiando, assim, o trabalho de outros pesquisadores. Além disso, devido ao caráter automatizado dessa solução proposta, o mesmo processo poderá ser utilizado, avaliado e melhorado para outras situações, contextos e idiomas.

Por fim, a pesquisa e a utilização de diversas técnicas de PNL e criação não supervisionada de padrões poderão servir como um *benchmark* dos principais métodos, auxiliando na escolha de ferramentas e abordagens para trabalhos futuros em contextos específicos.

Referências

[Asghar et al., 2014] Asghar, M. Z., Khan, A., Ahmad, S., and Kundi, F. M. (2014). A review of feature extraction in sentiment analysis. *Journal of Basic and Applied Scientific Research*, 4(3):181–186.

[Bandhakavi et al., 2016] Bandhakavi, A., Wiratunga, N., Padmanabhan, D., and Massie, S. (2016). Lexicon based feature extraction for emotion text classification. *Pattern Recognition Letters*, pages –.

[Ding et al., 2008] Ding, X., Liu, B., and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240. ACM.

[Duwairi et al., 2015] Duwairi, R. M., Ahmed, N. A., and Al-Rifai, S. Y. (2015). Detecting sentiment embedded in arabic social media - a lexicon-based approach. *Journal of Intelligent and Fuzzy Systems*, 29(1):107–117.

[Eisenstein, 2016] Eisenstein, J. (2016). Unsupervised learning for lexicon-based classification. *CoRR*, abs/1611.06933.

[Ghaleb and Vijendran, 2016] Ghaleb, O. A. M. and Vijendran, A. S. (2016). Survey and analysis of recent sentiment analysis schemes relating to social media. *Indian Journal of Science and Technology*, 9(41).

[Khoo and Johnkhan, 2017] Khoo, C. S. and Johnkhan, S. B. (2017). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 0(0):0165551517703514.

[Liu, 2010] Liu, B. (2010). Sentiment analysis: A multifaceted problem. *IEEE Intelligent Systems*, 25(3):76–80.

[Mohammad, 2016] Mohammad, S. M. (2016). Challenges in sentiment analysis. *A Practical Guide to Sentiment Analysis, D. Das, E. Cambria, and S. Bandyopadhyay, Eds. Springer*.

[Taboada et al., 2011] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307.

[Widdows and Dorow, 2002] Widdows, D. and Dorow, B. (2002). A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.