Modelos de Classificação de Sentimentos em *Tweets* usando Programação Genética

Airton Bordin Junior[1]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)
Caixa Postal 131 – 74690-900 – Goiânia – GO – Brazil

Abstract. *The increase in the number of Internet users in recent years has resulted in a growing content production by its users. Often, the WEB is used as a platform for debates, opinions, evaluations, etc. This fact, in line with the ease of obtaining the information, made the area of Sentiment Analysis, also called Opinion Mining, a growing interest on the part of reserachers.*
One of the most used strategies in the process of Sentiment Analysis is the Lexical Dictionaries - a set of words and their polarities, generally defined as positive, negative or neutral. Although widely used, this approach has some challenges to overcome, such as identifying the domain of the text, for example - a word can have a completely different meaning depending on the context in which it is found.
The present work presents a Systematic Review of the literature to identify the main strategies adopted in the automatic creation and expansion of Lexical Dictionaries.

Resumo. *O aumento no número de usuários de Internet nos últimos anos teve como consequência uma crescentre produção de conteúdo por seus usuários. Frequentemente, a WEB é utilizada como plataforma para debates, opiniões, avaliações, etc. Esse fato, alinhado a facilidade de obtenção dessas informações, fez com que a área de Análise de Sentimentos, também chamada de Mineração de Opiniões, tivesse um interesse crescente por parte de pesquisadores.*
Uma das estratégias mais utilizadas no processo de Análise de Sentimentos é a utilização de Dicionários Léxicos - conjunto de palavras e suas polaridades, geralmente definidas como positiva, negativa ou neutra. Apesar de amplamente utilizada, essa abordagem possui alguns desafios a serem superados, como a identificação do domínio do texto, por exemplo - uma palavra pode ter um significado completamente diferente, dependendo do contexto em que se encontra.
O presente trabalho apresenta uma Revisão Sistemática da literatura para identificar as principais estratégias adotadas na criação e expansão automatizada de Dicionários Léxicos.

1. Análise de Sentimentos

A Análise de Sentimentos, também chamada de Análise de Opiniões ou Mineração de Opiniões, é uma linha de pesquisa abrangente e que vem sendo tema de diversos trabalhos nos últimos anos. Como observado em [Liu 2010], esse crescente interesse sobre o assunto ocorre principalmente devido ao aumento no número de usuários de Internet e o

consequente crescimento da produção de conteúdo independente na rede, como opiniões, avaliações, entre outros.

Essa área de estudo tem como principal desafio a Análise 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.

Uma das principais técnicas para aumentar a acurácia a Análise de Sentimentos é a utilização de Dicionários de Dados. Esses dicionários contêm palavras previamente avaliadas por especialistas humanos, principalmente quanto à sua polaridade. Neste contexto, esse conjunto de palavras, juntamente com suas polaridades, é chamado de Dicionário Léxico ou Dicionário de Sentimentos.

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 e ampliação do Dicionário Léxico de forma automatizada, tema central do presente trabalho. Grande parte desses dicionários são construídos de forma manual, fato que caracteriza uma limitação óbvia para a maior parte dos contextos e domínios, como observado em [Duwairi et al. 2015].

Existem, basicamente, 3 formas de criação e expansão de um Dicionário Léxico: manual - processo realizado por especialistas humanos que analisam cada palavra, atribuindo uma Orientação Semântica para cada uma delas - e duas formas (semi) automatizadas: baseada em Dicionário e baseada em Corpus. Frequentemente, essas técnicas são utilizadas em conjunto, principalmente a validação manual de Dicionários criados de forma automatizada. Criações de Dicionários utilizando somente abordagem manual, por sua característica limitante, são menos utilizadas e não serão abordadas de forma mais aprofundada no decorrer deste trabalho.

Consciente do problema de criação e expansão de Dicionários Léxicos para a Análise de Sentimentos, a ideia principal do presente trabalho é a [descrever aqui]

2. Programação Genética

Programação Genética é uma técnica de programação evolucionária que busca resolver problemas, de forma automatizada, sem demandar conhecimentos detalhados sobre a solução. De forma geral, podemos definir a Programação Genética como um método sistemático, não dependente de um domínio específico, usado para permitir que computadores solucionem problemas de forma automática, inciando com um conhecimento de alto nível sobre as regras gerais das possíveis soluções.

asdfasdf

asdfasdfasdfasf

asdfasdfasdf

3. Problema

4. Dicionários (pode ser uma subsessão)

Para o presente trabalho, [foram utilizados - colocar algo que queira dizer isso] os dicionários de palavras positivas e negativas de [Hu and Liu 2004][1]. Os dicionários fornecem um conjunto de 4783 palavras negativas e 2006 palavras positivas para apoiar no processo de Análise de Sentimentos.

Utilizou-se, também, o dicionário de emoticons SentiStrength[2], que fornece 46 emoticons positivos e 58 negativos.

5. Datasets (pode ser uma subsessão)

6. Resultados

7. Conclusão

A Análise de Sentimentos é uma área que vem ganhando cada vez mais a atenção dos pesquisadores. Por ser uma linha de pesquisa multidisciplinar, possui vários desafios que devem ser enfrentados pelas pesquisas futuras.

No que tange à criação e expansão do Dicionário Léxico, assunto principal do presente trabalho, novas técnicas podem ser aplicadas, como Algoritmos Evolucionários, de forma a tentar incrementar os resultados positivos das pesquisas. Além disso, a mudança de domínio continua sendo um desafio considerável.

Dicionários Léxicos de contexto geral vem sendo criados e disponibilizados para futuros trabalhos. Técnicas para a expansão e adequação para domínios e contextos específicos deverão ser criadas para aumentar ainda mais os resultados positivos.

Com o contínuo aumento da produção de conteúdo na WEB nos próximos anos, a área de Mineração de Opiniões terá um papel fundamental no apoio à tomada de decisões estratégicas e a criação de valor para empresas, governos e pesquisadores em geral.

Sem dúvidas, os desafios e resultados da área de Análise de Sentimentos são grandes motivadores para que novos projetos, pesquisas e produtos sejam desenvolvidos.

References

Becker, L., Erhart, G., Skiba, D., and Matula, V. (2013). Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 333–340.

D'Andrea, A., Ferri, F., Grifoni, P., and Guzzo, T. (2015). Article: Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications*, 125(3):26–33. Published by Foundation of Computer Science (FCS), NY, USA.

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.

[1]Disponível em https://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html#lexicon

[2]Disponível em http://sentistrength.wlv.ac.uk/

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

Ferreira, L., Dosciatti, M., Nievola, J. C., and Paraiso, E. C. (2015). Using a genetic algorithm approach to study the impact of imbalanced corpora in sentiment analysis. In *FLAIRS Conference*, pages 163–168.

Gilbert, C. H. E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Govindarajan, M. (2013). Sentiment analysis of movie reviews using hybrid method of naive bayes and genetic algorithm. *International Journal of Advanced Computer Research*, 3(4):139.

Graff, M., Tellez, E. S., Escalante, H. J., and Miranda-Jiménez, S. (2017). Semantic genetic programming for sentiment analysis. In *NEO 2015*, pages 43–65. Springer.

Guimaraes, N., Torgo, L., and Figueira, A. (2016). Lexicon expansion system for domain and time oriented sentiment analysis. In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016)*, pages 463–471.

Haddi, E., Liu, X., and Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, 17:26 – 32. First International Conference on Information Technology and Quantitative Management.

Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, pages 174–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, New York, NY, USA. ACM.

Iqbal, M., Karim, A., and Kamiran, F. (2015). Bias-aware lexicon-based sentiment analysis. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, SAC '15, pages 845–850, New York, NY, USA. ACM.

Kaji, N. and Kitsuregawa, M. (2007). Building lexicon for sentiment analysis from massive collection of html documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1075–1083, Prague, Czech Republic. Association for Computational Linguistics.

Keshavarz, H. and Abadeh, M. S. (2017). Alga: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs. *Knowl.-Based Syst.*, 122:1–16.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(TR/SE-0401):28.

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

Musto, C., Semeraro, G., and Polignano, M. (2014). A comparison of lexicon-based approaches for sentiment analysis of microblog posts. *Information Filtering and Retrieval*, 59.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2009). Sentiful: Generating a reliable lexicon for sentiment analysis. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–6.

Pantel, P. and Pennacchiotti, M. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 113–120, Stroudsburg, PA, USA. Association for Computational Linguistics.

Perez-Rosas, V., Banea, C., and Mihalcea, R. Learning sentiment lexicons in spanish.

Pinto, D., Rosso, P., and Jiménez-Salazar, H. (2007). Upv-si: Word sense induction using self term expansion. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 430–433, Stroudsburg, PA, USA. Association for Computational Linguistics.

Silva, M. J., Carvalho, P., Costa, C., and Sarmento, L. (2010). Automatic expansion of a social judgment lexicon for sentiment analysis.

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

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

Zhou, Z., Zhang, X., and Sanderson, M. (2014). *Sentiment Analysis on Twitter through Topic-Based Lexicon Expansion*, pages 98–109. Springer International Publishing, Cham.