Análise de Sentimentos utilizando Dicionários Léxicos Uma Revisão Sistemática

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. Introdução

A Mineração de Opiniões, também chamada de Análise de Opiniões ou Análise de Sentimentos, 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 realização de uma Revisão Sistemática da Literatura, conforme apresentada em [Kitchenham 2004], de forma a identificar e analisar os principais trabalhos disponíveis sobre o tema. Esses trabalhos apoiarão na resolução das questões de pesquisa, apresentadas em detalhes na próxima seção. A análise dos artigos também auxiliará na identificação dos *gaps* de pesquisa, que poderão ser explorados em trabalhos futuros.

2. Estratégia da Revisão Sistemática

A Revisão Sistemática da Literatura fornece uma forma estruturada, objetiva e reprodutível de identificar, avaliar e interpretar trabalhos relevantes em uma determinada área de conhecimento. A análise desses trabalhos apoia a resoluções das questões de pesquisa, que devem ser respondidas pelo projeto [Kitchenham 2004].

A definição das questões da pesquisa é uma parte crítica da Revisão Sistemática. Essas mesmas questões são utilizadas de forma a orientar a estratégia de busca e palavras-chave dos artigos nas bases de dados escolhidas.

As questões norteiam, também, os dados e informações que serão relevantes e extraídos dos trabalhos selecionados. Para este trabalho, as questões de pesquisa são as

seguintes:

- Quais as principais estratégias utilizadas para a classificação de sentimentos?
- Quais as principais técnicas empregadas na criação e expansão de Dicionários Léxicos?
- Como a diferença de contexto e domínio vem sendo tratada no contexto de Análise de Sentimentos?
- Que métricas estão sendo utilizadas para avaliar a qualidade dos algoritmos de criação e expansão de Dicionários Léxicos?

Para a elaboração desta Revisão Sistemática, foram realizadas buscas nas seguintes bases de dados: *Science Direct*, *IEEEXplore*, *ACM Digital Library*, *Research Gate*, *Semantic Scholar*. Essas buscas tem por objetivo coletar os trabalhos primários sobre o tema proposto para a identificação do estado obre o assunto.

Alguns critérios de seleção dos trabalhos devem ser adotados de forma a realizar a filtragem dos artigos mais relevantes. Para esta revisão, consideramos somente pesquisas realizadas a partir do ano 2000 e que foram publicadas no idioma inglês. Para o presente trabalho, consideramos apenas os artigos disponíveis gratuitamente. A tabela 1 apresenta as bases de dados utilizadas e os critérios básicos de filtragem.

Base de dados	**Anos cobertos na busca**	**Idioma**
Science Direct http://www.sciencedirect.com/	2000 até 2017	Inglês
IEEEXplore http://ieeexplore.ieee.org/	2000 até 2017	Inglês
ACM Digital Library http://dl.acm.org/	2000 até 2017	Inglês
Research Gate https://www.researchgate.net/	2000 até 2017	Inglês
Semantic Scholar https://www.semanticscholar.org/	2000 até 2017	Inglês

Table 1. Relação de bases de dados consultadas

Baseando-se nas questões de pesquisa demonstradas, foram formuladas palavras-chave para orientar a criação de *strings* de busca nas ferramentas disponibilizadas pelas bases de dados. As palavras escolhidas para o conjunto foram:

- Sentiment Analysis;
- Opinion Mining;
- Lexicon Expansion;
- Genetic Algorithms;
- Semantic Orientation.

Essas palavras-chave representam os principais tópicos da pesquisa, auxiliando na busca e escolha de trabalhos relevantes ao tema. De posse dessas palavras, foram criadas as chaves de busca para a seleção dos artigos. Cada uma das bases de dados apresentadas na tabela 1 fornece uma ferramenta *online* avançada de busca. Há algumas diferenças

String de busca	Base de dados
pub-date >1999 and (Sentiment Analysis OR Opinion Mining) AND (Lexicon Expansion) AND (Genetic Programming OR Genetic Algorithm) AND (Semantic Orientation)[All Sources(Computer Science)]	*Science Direct*
pub-date >1999 and (Sentiment Analysis OR Opinion Mining) AND (Lexicon Expansion OR Lexicon) AND (Semantic Orientation)[All Sources(Computer Science)]	*IEEEXplore*
((Sentiment Analysis OR Opinion Mining) AND (Lexicon Expansion OR Lexicon) AND (Semantic Orientation)) and refined by Year: 2000-2017	*ACM Digital Library*
”query”: (+Sentiment +Analysis Lexicon Genetic Algorithm) ”filter”: ”publicationYear”: ”gte”:2000	*Research Gate*

Table 2. Detalhamento das palavras chave da busca e respectivo idioma

entre elas, de forma que foi necessário adequar a *string* para cada uma, conforme apresentado na tabela 2

Caso a base de dados permita, será feita uma ordenação por relevância, baseada na quantidade de citações e importância do trabalho. Após a aplicação das buscas nas respectivas bases, foi realizada uma segunda filtragem dos dados, de forma a selecionar trabalhos aderentes às questões de pesquisa.

Como segunda estratégia de filtragem foi feita a leitura dos títulos e, posteriormente, dos resumos de cada trabalho, de forma a aceitar ou rejeitar a pesquisa para a próxima fase da Revisão. Os trabalhos rejeitados não são levados em consideração para o desenvolvimento do projeto.

Para auxiliar nesse processo, foi utilizada a ferramenta StArt *(State of the Art through Systematic Reviews)*, desenvolvida pela Universidade Federal de São Carlos, que fornece funcionalidades para apoiar o pesquisador em todas as fases da Revisão Sistemática da Literatura.

O principal critério utilizado para assegurar a qualidade dos estudos primários é a quantidade de citações que o mesmo possui. Via de regra, um trabalho com muitas citações na literatura pode ser considerado um trabalho de qualidade, uma vez que é utilizado como referência por vários outros.

A avaliação da aderência do trabalho, como comentada anteriormente, é feita por meio da leitura do resumo. Caso o mesmo não seja conclusivo o suficiente para a avaliação, é feita a leitura da introdução e conclusão.

Importante salientar que esta análise foi realizada por apenas um pesquisador e pode acarretar em uma baixa segurança em relação à validade das avaliações feitas.

Após a escolha dos estudos primários, os artigos foram analisados quanto ao seu conteúdo. Resultados quantitativos foram recuperados e analisados. Dados de um mesmo domínio em trabalhos diferentes foram agrupados de forma a facilitar o estudo comparativo das pesquisas.

Os dados foram tabulados de forma a facilitar a visualização e comparação das abordagens utilizadas nos trabalhos primários consultados. Além disso, as informações mais relevantes, como técnicas utilizadas pela maior parte dos trabalhos, serão apresentadas de forma a demonstrar sua importância, descrevendo detalhes importantes caso mostre-se necessário. Detalhes do resultado da Revisão serão apresentados nas próximas seções.

3. Análise da Revisão Sistemática

De forma a nortear o processo de Revisão Sistemática da Literatura, a análise dos artigos selecionados teve como objetivo principal responder as questões de pesquisa apresentadas na seção anterior.

Considerando a primeira questão de pesquisa definida, podemos organizar a Classificação de Sentimentos em 3 grupos principais: Baseadas em Dicionário Léxico, *Machine Learning* e estratégia híbrida. Abordagens usando algoritmos de *Machine Learning* fazem uso de um conjunto de dados para o treinamento e um conjuntos de teste para validar os resultados do processo de aprendizado. A abordagem baseada em Dicionários Léxicos não demanda um treinamento anterior para proceder com a classificação, pois faz uso de um conjunto de palavras e suas respectivas polaridades (frequentemente definidas como positivo, negativo ou neutro). A abordagem híbrida combina as duas estratégias anteriores, de forma a potencializar a classificação dos sentimentos.

Estratégias de aprendizado de máquina, como discutido anteriormente, fazem uso de dados para treinamento de um modelo que representa o conhecimento sobre determinado contexto. Dados de testes são aplicados de forma a validar e melhorar as estruturas, potencializando o aprendizado. Abordagens usando técnicas de *Machine Learning* são tratadas em [Gilbert 2014], [Haddi et al. 2013], [Keshavarz and Abadeh 2017] e [Taboada et al. 2011].

Uma das formas comumente utilizadas para realizar a Análise de Sentimentos, conforme argumenta [Guimaraes et al. 2016], faz uso de um Dicionário Léxico (algumas vezes chamado de Dicionário de Sentimentos), um conjunto de palavras e suas orientações semânticas (também chamadas de polaridades), frequentemente representadas como positiva, negativa ou neutra. A obtenção de um Dicionário consistente é essencial para uma correta Mineração de Sentimentos.

O uso de Dicionário Léxico não demanda treinamento prévio, pois utiliza como referência para a classificação dos sentimentos o próprio conjunto de palavras e suas orientações semânticas. Dicionários de contexto geral estão disponíveis para a utilização em trabalhos na área de Mineração de Opiniões. Dentre os principais, podemos citar o *SentWordNet* e discutido em [Zhou et al. 2014], [D'Andrea et al. 2015], [Iqbal et al. 2015], [Keshavarz and Abadeh 2017], [Neviarouskaya et al. 2009] e [Guimaraes et al. 2016], o *General Inquirer*, tratado em [Taboada et al. 2011], [Zhou et al. 2014] e [D'Andrea et al. 2015] e o MPQA (*Multi-perspective Question Answering*), abordado em [Wilson et al. 2005], [Becker et al. 2013], [Taboada et al. 2011] e [Musto et al. 2014]. [Hu and Liu 2004] também disponibilizam em seu *website* uma lista de palavras positivas e negativas para utilização na classificação de sentimentos. Esses dicionários e respectivos endereços são organizados na tabela 3.

Como discutido anteriormente, a utilização dos Dicionários Léxicos disponíveis

Dicionário Léxico	url	palavras
SentiWordNet	*sentiwordnet.isti.cnr.it*	117000
General Inquirer	*www.wjh.harvard.edu/ inquirer*	4206
MPQA Lexicon	*mpqa.cs.pitt.edu/*	8221
Hu and Liu's Lexicon	*www.cs.uic.edu/ liub/FBS/sentiment-analysis.html*	6789

Table 3. Lista de Dicionários Léxicos

tem um resultado satisfatório quando aplicados à classificadores de propósito geral. Há, porém, uma limitação quanto à mudança de contexto, pois as palavras podem ter significados diferentes, dependendo do domínio ao qual se refere. Por exemplo, a palavra 'câncer' pode não possuir um significado negativo em um contexto técnico.

Por conta da limitação inerente ao domínio, muitos trabalhos criam seus próprios Dicionários Léxicos, ligados do contexto da Análise de Sentimentos em que estão inseridos. A criação e expansão de Léxicos, nossa segunda questão de pesquisa definida neste trabalho, é abordada em diversos trabalhos, com diferentes técnicas, como apresentadas no decorrer dessa seção. Responderemos, também, nossa terceira questão de pesquisa, que busca analisar como a diferença de domínios e contextos vem sendo tratadas no contexto da Análise de Sentimentos.

[Guimaraes et al. 2016] argumenta que podemos classificar a criação e expansão de um Dicionário Léxico de 3 formas: 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.

No contexto de prognóstico automatizado de Orientação Semântica de palavras, um dos primeiros trabalhos apresentados foi [Hatzivassiloglou and McKeown 1997], focando na previsão de polaridade de adjetivos.

Uma forma de prever a polaridade sentimental de palavras desconhecidas é levar em consideração aspectos sintáticos e semânticos do texto. [Turney 2002] apresenta uma abordagem de expansão léxica fazendo uso da técnica de Pointwise Mutual Infomation (PMI), com o objetivo de calcular a co-ocorrência de palavras e, com isso, comparar a polaridade de novas palavras com outras já conhecidas. Nesse trabalho, amplamente referenciado por outras pesquisas, o autor compara o conjunto de palavras de Orientação Semântica desconhecida com as palavras *"excellent"* e *"poor"*, representando Orientações Semânticas positiva e negativa, respectivamente. Essas palavras previamente conhecidas utilizadas como base para a expansão do Dicionário são chamadas de palavras semente (*seed words*, em inglês). Como exemplo de trabalhos que utilizam o PMI para a criação e expansão do Dicionário Léxico podemos citar [Becker et al. 2013], [Zhou et al. 2014], [Pinto et al. 2007]. [Pantel and Pennacchiotti 2006] e [Kaji and Kitsuregawa 2007].

Existem alguns dicionários disponíveis para se trabalhar com criação e expansão de Dicionários Léxicos. A maior parte utiliza como base de palavras-semente o banco

de dados *WordNet* - disponível em https://wordnet.princeton.edu/ - que fornece outras facilidades, como sinônimos e antônimos. Importante destacar, também, que as bases utilizadas na maior parte dos trabalhos citados consideram palavras no idioma inglês. Mesmo alguns trabalhos que abordaram idiomas diferentes fizeram uso dessas bases por meio de um processo de tradução automatizada.

Entre os principais trabalhos da área de Análise de Sentimentos podemos citar [Taboada et al. 2011], que apresenta uma abordagem de Mineração de Opiniões baseada em Léxico combinada com uma verificação manual. Esse trabalho 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 Dicionários. 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.

Quanto às métricas utilizadas para a avaliação dos algoritmos de criação e expansão de Dicionários Léxicos, nossa quarta questão de pesquisa...

[Fazer tabelas para melhorar a representação dos dados. Ideias para tabelas: estratégias utilizadas para a expansão do léxico, dicionários para a expansão, etc]

3.1. Subsections

4. Figures and Captions

Figure and table captions should be centered if less than one line (Figure 1), otherwise justified and indented by 0.8cm on both margins, as shown in Figure **??**. The caption font must be Helvetica, 10 point, boldface, with 6 points of space before and after each caption.

Figure 1. A typical figure

In tables, try to avoid the use of colored or shaded backgrounds, and avoid thick, doubled, or unnecessary framing lines. When reporting empirical data, do not use more

decimal digits than warranted by their precision and reproducibility. Table caption must be placed before the table (see Table 1) and the font used must also be Helvetica, 10 point, boldface, with 6 points of space before and after each caption.

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.

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

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.

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.

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.

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.