

Análise de Sentimentos utilizando Dicionários Léxicos

Uma Revisão Sistemática

Airton Bordin Junior¹

¹Instituto 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 crescente 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 reproduzível de identificar, avaliar e interpretar trabalhos relevantes em uma determinada área de conhecimento. A análise desses trabalhos apoia a resolução 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*, *IEEEExplore*, *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
<i>Science Direct</i> http://www.sciencedirect.com/	2000 até 2017	Inglês
<i>IEEEExplore</i> http://ieeexplore.ieee.org/	2000 até 2017	Inglês
<i>ACM Digital Library</i> http://dl.acm.org/	2000 até 2017	Inglês
<i>Research Gate</i> https://www.researchgate.net/	2000 até 2017	Inglês
<i>Semantic Scholar</i> 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

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

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 conjunto 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.

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 2. 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

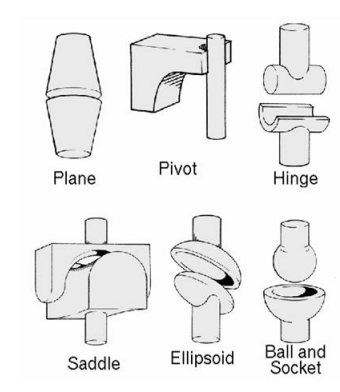


Figure 2. This figure is an example of a figure caption taking more than one line and justified considering margins mentioned in Section 4.

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.

Table 3. Variables to be considered on the evaluation of interaction techniques

	Chessboard top view	Chessboard perspective view
Selection with side movements	6.02 ± 5.22	7.01±6.84
Selection with in- depth movements	6.29±4.99	12.22±11.33
Manipulation with side movements	4.66± 4.94	3.47±2.20
Manipulation with in- depth movements	5.71 ±4.55	5.37 ±3.28

5. References

Bibliographic references must be unambiguous and uniform. We recommend giving the author names references in brackets, e.g. [Knuth 1984], [Boulic and Renault 1991], and [Smith and Jones 1999].

The references must be listed using 12 point font size, with 6 points of space before each reference. The first line of each reference should not be indented, while the subsequent should be indented by 0.5 cm.

References

Boulic, R. and Renault, O. (1991). 3d hierarchies for animation. In Magnenat-Thalmann, N. and Thalmann, D., editors, *New Trends in Animation and Visualization*. John Wiley & Sons Ltd.

- 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.
- Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(TR/SE-0401):28.
- Knuth, D. E. (1984). *The T_EX Book*. Addison-Wesley, 15th edition.
- Liu, B. (2010). Sentiment analysis: A multifaceted problem. *IEEE Intelligent Systems*, 25(3):76–80.
- Smith, A. and Jones, B. (1999). On the complexity of computing. In Smith-Jones, A. B., editor, *Advances in Computer Science*, pages 555–566. Publishing Press.