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.
[continue]

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.
[continuar]

1. Introdução

2. 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]

3. Programação Genética

Programação Genética é um campo da computação evolucionária que busca resolver problemas, de forma automatizada, sem demandar conhecimentos detalhados sobre a solução [Koza 1992]. De forma geral, podemos definir a Programação Genética como um método sistemático, não dependente de um domínio específico, usado para permitir que computadores criem programas para solução de problemas de forma automática, inciando com um conhecimento de alto nível sobre as regras gerais dos possíveis modelos.

Nesse contexto, programa significa um modelo capaz de, à partir de uma ou mais entradas, produzir uma saída para as mesmas. Embora possam ser representadas por diversos tipos de estruturas, a forma mais comum de representação é por meio de árvores, onde os nós internos representam funções e os nós folha representam terminais do problema. Um exemplo de um programa pode ser visto na figura 1 [Sannino et al. 2015].

Na Programação Genética, assim como em outros algoritmos baseados na evolução humana, são criadas populações onde cada indivíduo representa uma possível solução para o problema. A forma mais comum de inicialização da população é fazê-las de forma aleatória, evoluindo as mesmas no decorrer dos ciclos, chamados de gerações. Para cada geração, programas possivelmente melhores são criados, evoluindo os programas gerados. Assim como a natureza, a Programação Genética é um processo aleatório, e não garante o resultado ótimo. Porém, essa aleatoriedade faz com que, muitas vezes, as

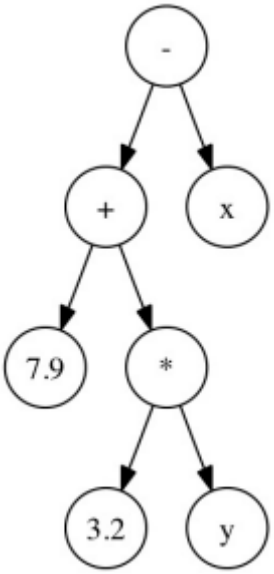

Figure 1. Programa representado a expressão (7.9 + (3.2 y) - x)

soluções fujam de problemas como soluções máximas locais, frequentemente enfrentado por métodos determiníticos gulosos [McPhee et al. 2008].

Uma das características mais importantes da Programação Genética é a função de aptidão, ou função *fitness*. Essa função busca representar, de forma quantitativa, a similaridade de cada indivíduo em relação ao resultado esperado. A escolha de uma boa função *fitness* está intimamente ligada ao tipo do problema.

Os responsáveis pela evolução da população de indivíduos são os operadores genéticos. Para a Programação Genética, os principais operadores são a seleção, mutação e *crossover*. Na seleção, um indivíduo é escolhido para fazer parte da próxima geração. A mutação modifica um nó da árvore, de forma a criar um indivíduo modificado em uma das partes escolhida aleatoriamente. O *crossover* realiza o cruzamento entre dois indivíduos (pais), gerando duas novas possíveis soluções para o problema (filhos). Além desses principais operadores, existem outros como a edição, encapsulamento e a destruição [PATELLI 2011].

[Colocar mais informações aqui?]

O processo de criação e evolução de indivíduos e gerações pode ser visto em detalhes no fluxograma apresentado na figura 2

4. Trabalhos relacionados

A Análise de Sentimentos é uma linha de pesquisa multidisciplinar, podendo ser considerada uma subárea de Processamento de Linguagem Natural (PNL), como afirma [Liu 2010]. O autor, um dos principais nomes sobre o assunto, 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

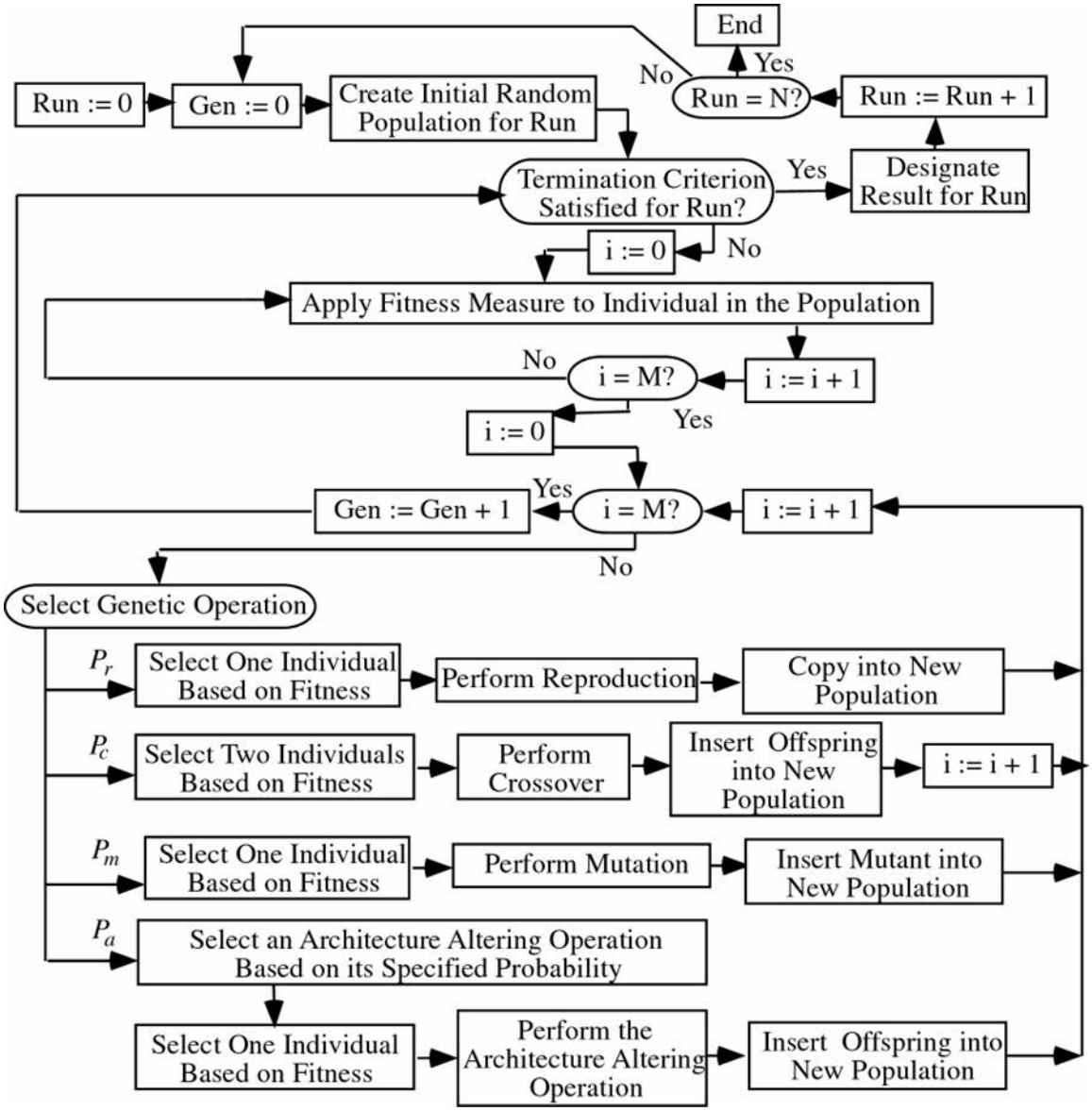

Figure 2. Fluxograma da Programação Genética

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], [Guimaraes et al. 2016], [Taboada et al. 2011], [Bandhakavi et al. 2016], [D'Andrea et al. 2015], entre outros trabalhos.

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], [Duwairi et al. 2015], entre outros.

Outro importante trabalho sobre Mineração de Opiniões, [Taboada et al. 2011] 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.

Na mesma linha, [Eisenstein 2016] e [Bandhakavi et al. 2016] apresentam outros procedimentos para apoiar a Análise de Sentimentos. 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].

Ainda tratando especificamente de Léxicos, [Kaji and Kitsuregawa 2007] aborda uma estratégia de criação e expansão de dicionários analisando uma coleção de páginas HTML. Apesar de trabalhar com o idioma japonês, a técnica pode ser adequada para outras línguas.

Sistemas Classificadores, como o criado em [Rodrigues et al. 2016], recebem como entrada um texto e retornam a Orientação Semântica do mesmo. Para esse processo, [Rodrigues et al. 2016] faz uso de um Dicionário próprio, com grande parte das polaridades anotadas manualmente. Outros classificadores são descritos em [Pang et al. 2002], [Zhou et al. 2014], [Silva et al. 2010], [Guimaraes et al. 2016], entre outros.

Cada classificador possui regras próprias para a avaliação da Orientação Semântica dos textos de entrada, dependendo do tipo de informação, contexto e outras características do domínio avaliado. A criação de modelo que faça uma classificação eficiente das entradas é um desafio, e demanda um conhecimento prévio sobre o assunto abordado.

A criação de modelos pode ser vista como um problema de otimização. Para essa classe de problemas, podemos fazer uso de estratégias de computação evolucionária, baseadas na teoria da evolução de *Darwin*. Dentre os trabalhos que abordam a Análise de Sentimentos, fazendo uso de Estratégias Evolutivas, podem citar [Ferreira et al. 2015], [Vohra and Teraiya 2013] e [Haddi et al. 2013] e [Silveira et al. 2014].

Conjuntos de dados previamente avaliados são amplamente utilizados para teste dos Sistemas de Classificação. Esses dados tem sua Orientação Semântica determinadas por especialistas humanos, e servem como entrada para a comparação da saída dos clas-

sificadores, ou seja, são dados considerados corretos e consistentes. Principais conjuntos de dados disponíveis para utilização no processo de Análise de Sentimentos podem ser vistos em [Hu and Liu 2004], [Iqbal et al. 2015], [Taboada et al. 2011], entre outros.

5. Análise do problema

6. Projeto da solução

6.1. Bibliotecas

Para apoiar no desenvolvimento da solução do problema, foi utilizada a biblioteca DEAP[1] (*Distributed Evolutionary Algorithms in Python*), escrita na linguagem *Python* e disponível para uso gratuito. Fornece abstrações para a implementação de várias classes de algoritmos evolucionários, como Algoritmos Genéticos, Programação Genética, entre outros. [Fortin et al. 2012]

Especificamente para o contexto de Programação Genética, DEAP fornece funcionalidades para controle de criação das estruturas de árvores, operadores genéticos, parametrização das operações, *logs*, entre outras.

6.2. Dicionários

Para o presente trabalho, [foram utilizados - colocar algo que queira dizer isso] os dicionários de palavras positivas e negativas de [Hu and Liu 2004][2]. 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[3], que fornece 46 emoticons positivos e 58 negativos.

6.3. *Datasets*

Há diversas bases de dados anotadas disponíveis na Internet para *download*. O escopo do presente trabalho é a Análise de Sentimentos em *Tweets*, por isso utilizou-se uma base disponibilizada para o evento SemEvaL 2016[4] (*International Workshop on Semantic Evaluation*), uma das principais referências na área de análise semântica.

O evento é dividido por *Tasks*, que possuem objetivos distintos dentro da área de análise semântica. Para este trabalho, utilizou-se a base de dados da *Task 4 - Sentiment Analysis in Twitter*. São disponibilizadas bases de treinamento e de testes para *download*[5] no site do evento.

A base de treinamento aplicada no trabalho possui 1421 *Tweets*, avaliados como positivo, negativo ou neutro. Por decisão de projeto, as mensagens neutras são ignoradas no decorrer do trabalho. Essa base possui 802 *Tweets* positivos e 130 negativos, além de 489 mensagens neutras ignoradas. [Base muito desbalanceada - problema? Verificar novamente esse número, baixar a base de novo]

[Falar sobre a base de teste agora]

[1]https://github.com/DEAP/deap

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

[3]http://sentistrength.wlv.ac.uk/

[4]http://alt.qcri.org/semeval2016/

[5]http://alt.qcri.org/semeval2016/task4/data/uploads/semeval2016-task4.traindevdevtest.v1.2.zip

7. Resultados

8. Conclusão

References

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

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.

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, 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, 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.

Fortin, F.-A., De Rainville, F.-M., Gardner, M.-A., Parizeau, M., and Gagné, C. (2012). DEAP: Evolutionary algorithms made easy. *Journal of Machine Learning Research*, 13:2171–2175.

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).

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.

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.

Koza, J. R. (1992). *Genetic programming: on the programming of computers by means of natural selection*, volume 1. MIT press.

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

McPhee, N. F., Poli, R., and Langdon, W. B. (2008). Field guide to genetic programming.

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

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

PATELLI, A. (2011). Genetic programming techniques for nonlinear systems identification.

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.

Rodrigues, R. G., das Dores, R. M., Camilo-Junior, C. G., and Rosa, T. C. (2016). Sentihealth-cancer: A sentiment analysis tool to help detecting mood of patients in online social networks. *International Journal of Medical Informatics*, 85(1):80 – 95.

Sannino, G., De Falco, I., and De Pietro, G. (2015). *Indirect Blood Pressure Evaluation by Means of Genetic Programming*, pages 75–92. Springer International Publishing, Cham.

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

Silveira, B. B., Leitão-Júnior, P. S., Ramada, M. S., and Martins, B. P. (2014). Geração de base de dados para o teste de aplicações de banco de dados pelo emprego da computação evolucionária.

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.

Vohra, S. and Teraiya, J. (2013). A comparative study of sentiment analysis techniques. *Journal JIKRCE*, 2(2):313–317.

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