Mineração de Textos: Análise de Sentimento em Redes Sociais - Revisão Sistemática

Elaine B. Figueiredo[1,3], Rita de Cássia Catini[2,3] e Leonardo Manoel Mendes[3]

[1]Universidade Anhembi Morumbi - São Paulo, SP - Brasil

[2]Faculdade de Tecnologia Arthur Azevedo (FATEC) - Mogi-Mirim, SP – Brasil

3Centro Universitário Campo Limpo Paulista (UNIFACCAMP) – Campo Limpo Paulista – SP – Brasil

efigueiredo13@gmail.com, ritacatini@gmail.com, leonardomanoelmendes@gmail.com

Abstract: *The sentiment analysis or opinion mining, is a subarea of data mining that uses PLN and text mining techniques to obtain the polarity of opinion. The present systematic review aims to identify and / or classify the main techniques used to express and analyze affectivity in comments on social networks*

Resumo: *A análise de sentimentos ou mineração de opiniões, é uma subárea da mineração de dados que utiliza PLN e técnicas de mineração de textos para obter a polaridade de opinião. A presente revisão sistemática tem por objetivo identificar e/ou classificar as principais técnicas utilizadas para expressar e analisar a afetividade em comentários nas redes sociais.*

1. Introdução

A Análise de sentimento ou Mineração de Opinião é uma área da mineração de dados que utiliza-se de técnicas de processamento de língua natural (PLN - NLP *Natural Language Processing*) e mineração de textos para classificar a polaridade da opinião: positiva, negativa ou neutra. As técnicas de processamento de língua natural (PLN) analisa o "*corpus*" do texto reconhecendo o contexto da informação, ou seja, analisando por meio de viés, na análise de sentimento a pln atua em conjunto com mineração de texto.

Técnicas de mineração de texto, quando aplicadas ao conteúdo publicado pelos usuários da Web, oferecem processos para obtenção de informações importantes de textos. Porém, como não existe padronização na forma como as pessoas se expressam na Internet, existem inúmeros desafios na aplicação das técnicas de mineração. Segundo Cheng (2016), a maioria dos problemas de mineração de texto se concentra em duas partes: (i) extração e seleção de recurso, e (ii) métodos de aprendizagem de máquina para classificação.

Há técnicas variadas para analisar a polaridade de um comentário e, de acordo com a necessidade intensificar o grau de abstração das informações, a seleção das técnicas podem variar bastante. É comum a combinação de mais de uma técnica de mineração de texto e muitas vezes é essencial para a melhora significativa do resultado.

Para esta análise, realizamos uma pesquisa em cinco (5) bases científicas, que resultaram em 628 trabalhos relacionados, sendo 53 incluídos e 575 excluídos. A fundamentação da pesquisa nos trabalhos relacionados (seção 2), levantamento dos trabalhos (seção 3) e a análise dos resultados (seção 4).

2. Contextualização

Esta seção apresenta uma visão geral referentes áreas de concentração da pesquisa. A

seguir serão abordados assuntos como afetividade, comunicação interpessoal na Web, mineração de texto e revisão sistemática, todos importantes para o claro entendimento deste trabalho.

2.1. Comunicação interpessoal na Web e Afetividade

A comunicação interpessoal é o processo pelo qual informações são trocadas e entendidas por duas ou mais pessoas. Muitas vezes o processo de comunicação ocorre por meio de recursos oferecidos pela Web, como redes sociais e blogs, por exemplo, e podem ter a intenção de motivar ou influenciar o comportamento. Entender e classificar de forma automatizada qual a afetividade, característica humana que nos permite demonstrar nossos sentimentos e emoções, envolvida em um comentário não é uma tarefa fácil, mas pode ter grande valor para pessoas ou empresas quando bem interpretadas.

Para Maynard e Funk (2011) As técnicas de classificação do sentimento podem ser aproximadamente divididas em abordagem de aprendizado de máquina, abordagem baseada em léxico e abordagem híbrida. A abordagem de aprendizagem de máquina (ML-*Machine Learning*) aplica os famosos algoritmos ML e usa recursos linguísticos. Já a abordagem baseada em léxico depende de um léxico de sentimento, uma coleção de termos de sentimento conhecidos e pré-compilados.

Aspectos relacionados a técnicas de processamento de linguagem natural (NLP) para detectar características em pequenos trechos de texto, conjunções textuais para separar palavras com sentimentos semelhantes ou opostos e as relações sintáticas entre palavras são também extremamente importantes para apoiar a classificação do sentimento. (MATSUMOTO et al; 2005)

2.2 Mineração de Sentimento

Segundo Gupta (2009) a mineração de texto, que também pode ser conhecida como Análise de Texto Inteligente, Mineração de Dados de Texto ou Descoberta de Conhecimento em Texto (KDT-*Knowledge Discovered in Texts*) subárea da Mineração de Dados, refere-se ao processo de extração de informações e conhecimento interessante e não trivial de dados não estruturados.

Muitos recursos e/ou “lugares” da Web, como blogs, fóruns, aplicativos, sites de comércio eletrônico e redes sociais, fornecem mecanismos para expressar opiniões. Ter recursos capaz de captar e interpretar as opiniões do público, pode ser uma vantagem competitiva que permite compreender o comportamento do consumidor e suas preferências de produtos, direcionando movimentos políticos, estratégias de empresas, campanhas de marketing, e monitoramento de reputação, por exemplo. A análise do sentimento é um estudo computacional de opiniões, sentimentos, emoções e atitude expressada em textos para uma entidade. (MEDHAT; HASSAN; KORASHY, 2014).

Para Ravi e Ravi (2015) o principal desafio da pesquisa é que os dados provenientes do mundo são não estruturados. Tal característica cria diversas dificuldades durante o processo de extração, seleção e classificação das mensagens. O autor ainda afirma que, além disso, a análise de sentimentos pode ser dividida em termos gerais em sete categorias. classificação de subjetividade, classificação de sentimentos, avaliação de medição de utilidade, criação de léxico, palavras de opinião e extração de aspecto de produto e detecção de spam de opinião. A escolha da categoria mais adequada está

diretamente ligada aos objetivos da mineração de texto.

2.3. Revisão Sistemática

A revisão sistemática é uma metodologia de pesquisa que apresenta uma avaliação justa a respeito de um tópico de pesquisa, fazendo uso de um modelo de revisão que seja confiável, rigoroso e permite auditagem [Kitchenham 2004]. A técnica define critérios claros de pesquisa e seleção de resultados claros, e permite aplicá-los em diferentes bases para obtenção de estudos de modo organizado,

Como em qualquer método de investigação científica, há uma preocupação em incluir o máximo possível de estudos, para evitar os vieses que podem levar a conclusões errôneas [Gonçalves 2009].

3. Levantamento de Trabalhos Existentes

Esta revisão sistemática segue as instruções sugeridas por [Kitchenham 2004], em que o processo é dividido em três fases: Planejamento, Execução e Controle.

3.1. Planejamento

A fase de planejamento é importante para definir a forma como a revisão sistemática será executada e os critérios que serão levados em consideração para a inclusão e exclusão de trabalhos [Biolchini et al. 2005].

Esta pesquisa tem por objetivo responder a questão de pesquisa **"*Quais técnicas de Mineração de texto são utilizadas atualmente para identificar e/ou classificar a afetividade em comentários nas redes sociais?*"**

Para tanto, foram identificadas as questões complementares abaixo, a serem avaliadas sobre o resultado da pesquisa, para a questão de pesquisa principal:

Q1. *Quais as técnicas de mineração de texto mais usadas atualmente para analisar redes sociais?*

Abaixo seguem os parâmetros definidos para a pesquisa (Tabela 1):

Tabela 1. Parâmetros da Pesquisa

Estratégia	Artigos disponíveis de 2010 a 2018, publicados em revistas e congressos, em inglês e português.
Fontes de pesquisa	IEEE, ACM, Scopus, SpringerLink, Science Direct. Pesquisa realizada na *web*.
Palavras-chave	Redes Sociais, Mineração de Texto e Afetividade (Análise de Sentimento)

Foram identificadas diversas formas de expressar as palavras-chave da pesquisa, por isso foi necessária a utilização de sinônimos para ampliar o resultado da pesquisa (Tabela 2). Para esta pesquisa foram utilizados somente sinônimos em inglês.

Tabela 2. Lista de Sinônimos

Palavra-Chave	Sinônimos
Mineração de Texto	*text mining* *kdt* *Discovering knowledge in text*
Afetividade	*Affectivity* *sentiment* *sentiment classification* *Sentiment Analysis* *expressions of positivity* *expressions of negativity*

Redes Sociais	*Social Network*

A seleção preliminar dos trabalhos foi realizadas por três principais ações: construção da *strings* de busca, realização das buscas, seleção preliminar dos trabalhos. A construção da *string* de busca de busca foi feita a partir da combinação dos sinônimos sobre a *string* principal utilizando-se operadores lógicos *"and"* e *"or"*, e respeitando as particularidades de busca de cada base selecionada. Segue a *string* de busca principal:

("social network") AND ("text mining" OR "kdt" OR "Discovering knowledge in text") AND ("affectivity sentiment" OR "sentiment classification" OR "Sentiment Analysis" OR "expressions of positivity" OR "expressions of negativity")

Uma vez realizadas as buscas nas bases foi necessário refinar a pesquisa. Para a seleção dos artigos, foram definidos critérios de inclusão e exclusão (Tabela 3) e aplicados à partir da leitura do **Título** e **Resumo** de cada artigo. Uma vez atendidos os critérios, o artigo era selecionado como **estudo primário** da revisão sistemática, para leitura de seu conteúdo integral.

Tabela 3. Critérios de Seleção

Critério	ID	Descrição
Inclusão	**I1**	Considerar pesquisas que envolvam classificação de texto usando mineração de texto em redes sociais.
	I2	Avaliar alternativas de técnicas de mineração de texto.
	I3	Considerar estudos que utilizem técnicas de classificação de sentimentos
Exclusão	**E1**	Artigos com idioma diferente de inglês
	E2	Artigos do mesmo tema e autor, em diferentes bases
	E3	Artigos que possuam apenas título e resumo, sem corpo
	E4	Abordam experimentação de técnicas para classificar afetividade fora do contexto textual.
	E5	Abordam experimentação de técnicas para classificar afetividade fora da Web.
	E6	Estudos que utilizem técnicas de classificação de texto sem mineração de texto.

3.2. Execução

Como resultado da pesquisa foram recuperados **628 trabalhos**. A partir deste resultado foi possível observar **53 estudos primários** a serem analisados de acordo com os critérios de seleção apresentados na (Tabela 3). O resultado consolidado é observado no Gráfico1.

Foram identificados 575 artigos que não atenderam aos critérios de seleção, e portanto não foram considerados adequados para a revisão sistemática.

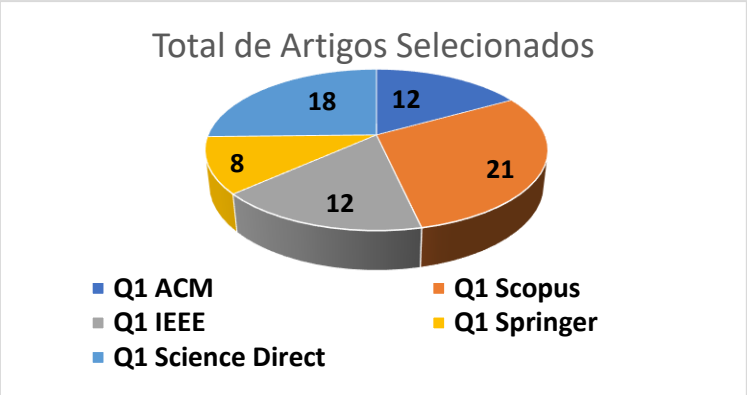

Gráfico 1. a) Total estudos selecionados para análise

Os estudos primários selecionados para análise seguem os respectivos critérios:

1- Realização das buscas em cada base de dados, avaliando apenas as 'palavras-chave' contidas na lista no 'Título' (*Title*) OU 'Abstract' OU "Corpo"(*full text*) de cada artigo.

2- Construção das strings de busca; utilizando "*OR*" e "*AND*".

3- Utilização de sinônimos para completar a busca, alimentando a base de "palavras-chave" para cada questão.

4- Para classificar os critérios de elegibilidade só serão lidos 'título' (*Title*) E '*abstract*'.

5- Pesquisa de acordo com os critérios de elegibilidade (inclusão e exclusão), classificando cada item encontrado com os respectivos critérios de inclusão e relevância.

6- Priorização dos artigos a serem avaliados, excluindo os duplicados e que estejam pouco aderentes a pesquisa.

7- Só serão lidos os artigos que atenderem aos critérios de inclusão, e não corresponderem aos critérios de exclusão: a partir destes serão realizadas as análises.

4. Análise dos Resultados

Na fase de análise dos resultados da revisão sistemática foram lidos todos os artigos com objetivo de responder a questão principal de pesquisa através da análise de questão complementar conforme a seguir:

Q1. Quais as técnicas de mineração de texto mais usadas atualmente para analisar redes sociais?

De acordo com os estudos primários selecionados as técnicas de mineração de textos mais utilizadas são: SVM (*Support Vector Machine*), Naïve Bayes, Entropia Máxima, técnicas de PNL (*process natural language*) como: *NGram*, *POSTagged Ngram*, *BiGram*, Percepção de Camada Múltipla (*multLayer Perceptron* -MLP), algumas técnicas híbrida alternando K-means+SVN, NLP *KeyWords*, *DeepLearning*, *Machine Learning*, Regressão Logística Multnomial, LDA, Análise Semântica, Web Semântica combinada com o *K-means* e LDA (*Lattent Dirichlet Allocation),* Índice Gini combinado com o SVM, Árvores de Decisão, dentre outras e técnicas combinadas.

5. Conclusão

Nos estudos primários selecionados percebe-se um enorme potencial de pesquisa em mineração de texto, há diversas técnicas, as que prevalecem na maioria dos estudos são: SVM *(Support Vector Machines),* Naïve Bayes e Máxima Entropia. Em estudos mais recentes, é notável a combinação de várias técnicas associadas como o algoritmo *k-means* associado a técnicas específicas para KDT.

Referências

CHENG, Ching-Hsue. *A Text Mining Based on Refined Feature Selection to Predict Sentimental Review*. In: Proceedings of the Fifth International Conference on Network, Communication and Computing. ACM, 2016. p. 150-154.

GUPTA, Vishal et al. *A survey of text mining techniques and applications*. Journal of emerging technologies in web intelligence, v. 1, n. 1, p. 60-76, 2009.

MATSUMOTO, Shotaro; TAKAMURA, Hiroya; OKUMURA, Manabu. *Sentiment Classification Using Word Sub-sequences and Dependency Sub-trees*. In: PAKDD. 2005.

p. 301-311.

MAYNARD, Diana; FUNK, Adam. *Automatic detection of political opinions in tweets*. In: Extended Semantic Web Conference. Springer, Berlin, Heidelberg, 2011. p. 88-99.

MEDHAT, Walaa; HASSAN, Ahmed; KORASHY, Hoda. *Sentiment analysis algorithms and applications: A survey*. Ain Shams Engineering Journal, v. 5, n. 4, p. 1093-1113, 2014.

RAVI, Kumar; RAVI, Vadlamani. *A survey on opinion mining and sentiment analysis: tasks, approaches and applications*. Knowledge-Based Systems, v. 89, p. 14-46, 2015.

Akaichi, J. (2013) 'Social networks' Facebook' statutes updates mining for sentiment classification', *Proceedings - SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013*, pp. 886–891. doi: 10.1109/SocialCom.2013.135.

Deshmukh, J. S. and Tripathy, A. K. (2018) 'Entropy based classifier for cross-domain opinion mining', *Applied Computing and Informatics*. King Saud University, 14(1), pp. 55–64. doi: 10.1016/j.aci.2017.03.001.

Han, H. *et al.* (2018) 'Generate domain-specific sentiment lexicon for review sentiment analysis', *Multimedia Tools and Applications*. Multimedia Tools and Applications, (145). doi: 10.1007/s11042-017-5529-5.

Hemmatian, F. and Sohrabi, M. K. (2017) 'A survey on classification techniques for opinion mining and sentiment analysis', *Artificial Intelligence Review*. Springer Netherlands, pp. 1–51. doi: 10.1007/s10462-017-9599-6.

J, A. K. and Abirami, S. (2018) 'Aspect-based opinion ranking framework for product reviews using a Spearman's rank correlation coefficient method', *Information Sciences*. Elsevier Inc., 460–461, pp. 23–41. doi: 10.1016/j.ins.2018.05.003.

Khan, F. H., Qamar, U. and Bashir, S. (2017) 'A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet', *Knowledge and Information Systems*. Springer London, 51(3), pp. 851–872. doi: 10.1007/s10115-016-0993-1.

Kim, K. (2018) 'An improved semi-supervised dimensionality reduction using feature weighting: Application to sentiment analysis', *Expert Systems with Applications*. Elsevier Ltd, 109, pp. 49–65. doi: 10.1016/j.eswa.2018.05.023.

Manek, A. S. *et al.* (2017) 'Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier', *World Wide Web*. World Wide Web, 20(2), pp. 135–154. doi: 10.1007/s11280-015-0381-x.

Riaz, S. *et al.* (2017) 'Opinion mining on large scale data using sentiment analysis and k-means clustering', *Cluster Computing*. Springer US, pp. 1–16. doi: 10.1007/s10586-017-1077-z.

Ristoski, P. and Paulheim, H. (2016) 'Semantic Web in data mining and knowledge discovery: A comprehensive survey', *Journal of Web Semantics*. Elsevier B.V., 36, pp. 1–22. doi: 10.1016/j.websem.2016.01.001.

Zhang, W. *et al.* (2018) 'Sentiment classification and computing for online reviews by a hybrid SVM and LSA based approach', *Cluster Computing*. Springer US, pp. 1–14. doi: 10.1007/s10586-017-1693-7.