6-2017

Lexicons in Sentiment Analytics

Bo Yuan
Missouri University of Science and Technology, byvf9@mst.edu

Keng L. Siau
Missouri University of Science and Technology, siauk@mst.edu

Lexicons in Sentiment Analytics

Bo Yuan
Missouri University of Science and Technology
byvf9@mst.edu

Keng Siau
Missouri University of Science and Technology
siauk@mst.edu

ABSTRACT
With the increasing amount of text data, sentiment analytics (SA) is becoming an important tool for text miners. An automated approach is needed to parse the online reviews and comments, and analyze their sentiments. Since lexicon is the most important component in SA, enhancing the quality of lexicons will improve the efficiency and accuracy of sentiment analysis. In this research, we study the effect of coupling a general lexicon with a specialized lexicon (for a specific domain) and its impact on sentiment analysis. Two special domains and one general domain were used. The two special domains are the petroleum domain and the biology domain. The general domain is the social network domain. The results, as expected, show that coupling a general lexicon with a specialized lexicon improves the sentiment analysis. However, coupling a general lexicon with another general lexicon does not improve the sentiment analysis.

Keywords
Lexicon, Sentiment Analysis, Text Mining, Machine Learning, Data Mining.

INTRODUCTION

Online reviews and comments are generating tones of textual data for text miners (Lee and Siau, 2001; Adeborna and Siau, 2014). Instead of simply classifying text based on keywords, it may be more valuable and insightful to analyze the sentiment of the text (Christos Troussas, Maria Virvou, Kurt Junshean Espinosa, Kevin Llaguno, Jaime Caro, 2013). Lexicon is an important component of sentiment analysis. Lexicon/corpus construction is generally viewed as a prerequisite for sentiment analysis. Some of the existing lexicons include the Harvard Inquirer, Linguistic Inquiry and Word Counts, MPQA Subjectivity Lexicon, Bing Liu's Opinion Lexicon, and SentiWordNet (Matheus Araújo; Pollyanna Gonçalves; Meeyoung Cha; Fabrício Benevenuto, 2014).

LITERATURE REVIEW

Lexicon plays an important role in SA. Among the lexicons mentioned above, SentiWordNet is the most frequently used and probably the most well-known for general analysis. SentiWordNet has three sentiment levels for each opinion word: positivity, negativity, and objectivity (dell'Informazione). SentiWordNet has developed from version 1.0 to version 3.0. There are some differences between SentiWordNet 1.0 and 3.0: (1) versions of WordNet and (2) algorithms used for annotating WordNet automatically. SentiWordNet 3.0 is working on improving part (2) (dell'Informazione).

CONCEPTUAL FOUNDATION

To establish a domain lexicon, the first step is data extraction. The technique used for data extraction is web crawler. Traditional search engines like AltaVista, Yahoo, and Google can also complete tasks which web crawler does. However, there are some limitations for those traditional search engines to complete crawler's work (Baike, 2010): 1) many non-related or less-related webpages are retrieved, 2) traditional search engines cannot handle some structured data, and 3) traditional search engines can only search according to key words but not semantic information. Web crawler is an Internet bot which systematically browses the World Wide Web, typically for the purpose of web indexing (Wikipedia, 2016).

Web crawler (Web crawler, 2016) can extract webpages from Internet automatically. In the process, web crawler needs to filter URLs which have no relation to our research according to specific web analysis algorithms, and extract and collect useful URLs into a waiting list. It then continues to extract URLs from the waiting list and downsize the waiting list at the same time until all URLs in the list satisfy web crawler system's specifications that were specified (ScienceDaily, 2016).

We apply LDA-based topic modeling method (Barber) to extract aspects. For LDA-based topic modeling, each document $d \in D$ of an unlabeled training corpus D is determined by a multinomial distribution θ. Given the topic z, a term t is characterized according to the multinomial distribution ϕ, determined by another hyper-parameter, a Dirichlet priori, β.

For the product aspect mining, we need to utilize a subset of the most informative topics to represent product aspects. We apply tf-idf measure to select the most informative topics to represent product aspects. For the experiments reported in this paper, we adopted topz = 15.

A set of user-related consumer reviews, domain knowledge reports, blogs, and articles are used to establish the relations between sentiments and aspects via learning process. Identifying the adjectives associated with the product aspects is a good step to establish pairs. The calculation (Lau, R. Y., Li, C., & Liao, S. S., 2014) is to give the pairs polarity scores to show how good they are and how bad they are. The polarity score of a sentiment-aspect pair sa is defined as follows:

$$WD(sa) = \tanh\left[\frac{df(sa)}{\omega_{pos}} \times \Pr(pos|sa) \times \log_2 \frac{\Pr(pos|sa)}{\Pr(pos)} - \frac{df(sa)}{\omega_{neg}} \times \Pr(neg|sa) \times \log_2 \frac{\Pr(neg|sa)}{\Pr(neg)}\right]$$

$$polarity_{Ont}(sa) = \begin{cases} \dfrac{WD(sa) - \omega_{\omega d}}{1 - \omega_{\omega d}} & if\ WD(sa) > \omega_{\omega d} \\ -\left(\dfrac{|WD(sa)| - \omega_{\omega d}}{1 - \omega_{\omega d}}\right) & if\ WD(sa) < -\omega_{\omega d} \\ 0 & otherwise \end{cases}$$

The central question of this research is whether coupling a general lexicon (e.g., SentiWordNet) with a domain specific lexicon will improve the results of sentiment analysis. Figure 1 illustrates the concept.

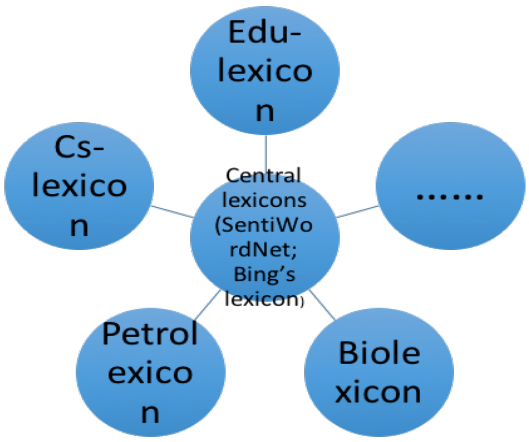

Figure 1. Lexicon Network

We can couple a general lexicon (e.g., SentiWordNet) with Biolexicon for biology domain, and couple a general lexicon with Petrolexicon for petroleum domain. With this modular design concept, there is no need for professionals in the petroleum industry to come up with a completely new and comprehensive lexicon for the petroleum industry. Rather, a small and specialized Petrolexicon can be constructed and this can be coupled with a general lexicon for sentiment analysis.

EVALUATIONS AND COMPARISONS

Three domains were selected in this research. One domain is petroleum industry and a Petrolexicon was constructed as part of this research. Another domain is the biology domain and a Biolexicon was used. We also used a SocialSent lexicon for the social network domain. The Petrolexicon and the Biolexicon are regarded as specialized domains. SocialSent lexicon, on the other hand, is not a very specialized domain and the text used in social media usually does not contain too many technical jargons. Figure 2 illustrates the analysis process.

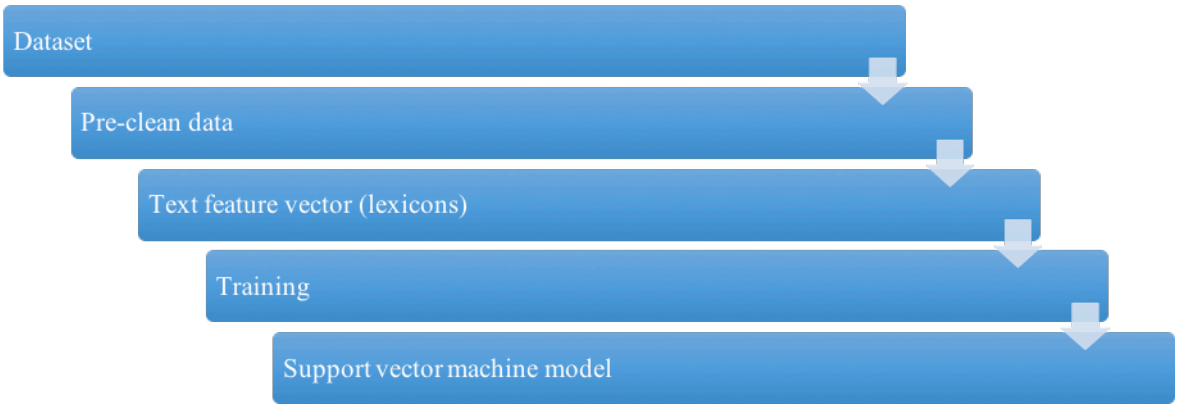

Figure 2. Analysis Procedure

The preliminary results are shown below for the three lexicons (Tables 1, 2, and 3). For example, for the Petrolexicon, we compared the SentiWordNet with the Petrolexicon, and we also compared the combination of SentiWordNet + Petrolexicon with SentiWordNet and Petrolexicon.

Lexicon	Product Reviews	Petroleum News, Reports, and Blogs	Journal Articles
SentiWordNet	0.7827477	0.7452156	0.6518541
Petrolexicon	0.8025648	0.8758446	0.9025464
SentiWordNet+Petrolexicon	0.8025486	0.9215569	0.9745665

Table 1. Results for Petrolexicon

Lexicon	Product Reviews	Biology News, Reports, and Blogs	Journal Articles (from Google Scholar)
SentiWordNet	0.8518152	0.8364654	0.7615454
BioLexicon	0.9016564	0.9453122	0.9815457
SentiWordNet+Biolexicon	0.9015666	0.9423321	0.9815956

Table 2. Results for Biolexicon

Lexicon	Product Reviews	Social Networking News, Reports, and Blogs	Journal Articles (from Google Scholar)
SentiWordNet	0.7648151	0.8084144	0.8186455
SocialSent	0.7695952	0.8448518	0.8318656
SentiWordNet+SocialSent	0.7628494	0.8485265	0.8326451

Table 3. Results for SocialSent

The results show that specialized lexicons (i.e., Petrolexicon and Biolexicon) seem to be performing better than SentiwordNet. Also, the combination of a general lexicon (i.e., in our case, SentiWordNet) and a specialized lexicon seems to produce better results for Petrolexicon. For Biolexicon, the combination of a general lexicon and a specialized lexicon produces about the same results as Biolexicon alone. For SocialSent, since it is not a specialized lexicon, there is hardly any difference between SentiWordNet and SocialSent.

EXPECTED CONTRIBUTIONS AND FUTURE RESEARCH

We hypothesize that specialized lexicons will improve the effectiveness of sentiment analyses, especially for specialized domains. Our preliminary results suggest that this hypothesis may be supported. We are in the process of conducting significant tests.

This study is expected to contribute to both academic researchers and practitioners. For academic research, we propose the idea of coupling specialized lexicons with general lexicons to enhance the quality of sentiment analyses. The modular approach also provides an efficient and effective ways to developing specialized lexicons for various industries and domains. For practitioners, this research suggests that coupling a general lexicon with specialized lexicon(s) can result in higher quality sentiment analyses.

There are a couple of directions for future research. For example, future research can investigate the best ways to couple the lexicons. Machine learning and deep learning approaches (Siau and Yang, 2017) to sentiment analysis can be explored and compared to existing sentiment analysis techniques. Sentiment analysis is focusing mainly on text at the moment. New efforts are studying the use of sentiment analysis on voice and video (e.g., Zhao and Siau, 2017).

REFERENCES

1. Adeborna, E., & Siau, K. (2014). An Approach to Sentiment Analysis-the Case of Airline Quality Rating. PACIS, (p. 363).
2. Baike (2010). Retrieved from Web Crawler: http://baike.baidu.com/link?url=vRXSRbTINNKhFO4ZlLMYMt1SYDfPCO9niSQU7U67As2sZGszEb_CDcovVSgHjuUp6U6ko4wji5258pwACRvtwh-J34quXfWXjwmN90TtoXX- PW5grbjNPlJCDkHzPBFw#ref_[1]_284853.
3. Barber, J. (n.d.). Latent Dirichlet Allocation (LDA) with Python. Retrieved from RSstudio: https://rstudio-pubs-static.s3.amazonaws.com/79360_850b2a69980c4488b1db95987a24867a.html.
4. Christos Troussas, Maria Virvou, Kurt Junshean Espinosa, Kevin Llaguno, Jaime Caro. (2013). Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning. Information, Intelligence, Systems and Applications (IISA).
5. dell'Informazione, I. d. (n.d.). SENTIWORDNET 3.0: An Enhanced Lexical Resource. Retrieved from http://www.researchgate.net/profile/Fabrizio_Sebastiani/publication/220746537_SentiWordNet_3.0_An_Enhanced_Lexical_Resource_for_Sentiment_Analysis_and_Opinion_Mining/links/545fbcc40cf27487b450aa21.pdf.
6. Lee, S., Siau, K. (2001). A Review of Data Mining Techniques. Industrial Management & Data Systems, 101(1), 41-46.
7. Matheus Araújo; Pollyanna Gonçalves; Meeyoung Cha; Fabrício Benevenuto. (2014). iFeel: a system that compares and combines sentiment analysis methods. International World Wide Web Conference (p. 1348). Seoul: InternationalWorld Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland.

8. Raymond Y.K. Lau, Stephen S.Y. Liao, Chunping Li. (2014). Social Analytics: Learning Fuzzy Product Ontologies for Aspect-Oriented Sentiment Analysis. Decision Support Systems, 65, 80-94.
9. ScienceDaily. (2016). Retrieved from Web crawler: https://www.sciencedaily.com/terms/web_crawler.htm.
10. Siau, K., Yang, Y. (2017). Impact of Artificial Intelligence, Robotics, and Machine Learning on Sales and Marketing. Twelve Annual Midwest Association for Information Systems Conference (MWAIS 2017), Springfield, Illinois, May 18-19, 2017.
11. Web crawler. (2016). Retrieved from ScienceDaily: https://www.sciencedaily.com/terms/web_crawler.htm.
12. Wikipedia. (2016, 11 22). Retrieved from Web crawler: https://en.wikipedia.org/wiki/Web_crawler.
13. Zhao, W., Siau, K. (2017). Machine Learning Approaches to Sentiment Analytics. Twelve Annual Midwest Association for Information Systems Conference (MWAIS 2017), Springfield, Illinois, May 18-19, 2017.