Spring 2017

Sentiment analytics: Lexicons construction and analysis

Bo Yuan

SENTIMENT ANALYTICS: LEXICONS CONSTRUCTION AND ANALYSIS

by

BO YUAN

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE IN INFORMATION SCIENCE AND TECHNOLOGY

2017

Approved by

Keng Siau, Advisor
Fiona Nah
Michael Gene Hilgers
Pei Yin

ABSTRACT

With the increasing amount of text data, sentiment analysis (SA) is becoming more and more important. 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, the effect of coupling a general lexicon with a specialized lexicon (for a specific domain) and its impact on sentiment analysis was presented. Two special domains and one general domain were studied. The two special domains are the petroleum domain and the biology domain. The general domain is the social network domain. The specialized lexicon for the petroleum domain was created as part of this research. 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.

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my advisor, Professor Keng Siau, who has the attitude and the substance of a genius: he continually and convincingly conveyed a spirit of adventure in regard to research and scholarship and an excitement in regard to teaching. Without his guidance and persistent help, this thesis would not have been possible.

I would like to thank my committee members, Professor Fiona Nah, Professor Michael Gene Hilgers, and Professor Pei Yin. They helped me in this journey and are concerned about my research progress and my well-being.

Finally, I would like to thank all my friends, IST staff, and my families for helping me survive all the stress during the last two years and not letting me give up.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

NOMENCLATURE

Symbol	Description
β	Dirichlet priori
θ	a multinomial distribution
ϕ	a multinomial distribution

1. INTRODUCTION

1.1. SENTIMENT ANALYSIS

Generally, data mining is the process of analyzing data in order to gain some goals and integrate it into useful information (Palace, 1996). Text mining is to use various mining algorithms to process useful information from the text (Text Mining, 2015). After text mining, sentiment analysis came out with more advanced technology for more accurate text mining. Sentiment analysis is to recognize and extract meaningful information using natural language processing (NLP) and computational linguistics from data. The application of sentiment analysis is happening in marketing, customer service, education and even energy fields (Sentiment analysis, 2015). Sentiment analysis is, undoubtedly, the advanced method in text mining, especially online social media data. As the Internet is developing rapidly, it is common to find reviews or comments of products, services, events, and brand names online (Matheus Araújo; Pollyanna Gonçalves; Meeyoung Cha; Fabrício Benevenuto, 2014). The goal of sentiment analysis is to identify the attitude of customers according to the polarity of the reviews and comments that they left online. Obviously, sentiment analysis created a new type of data. Data will be never only numerical digits but reviews and comments. It makes the contribution to gain what people think about the subject. This information may be from tweets, blogs, and new articles. A huge amount of sentences, conversations, product reviews and posts on social media are produced every second. They are all data which can be analyzed and provide much information to people. People here can refer to those in companies, costumers or users who experienced some products.

1.2. SENTIMENT LEXICON

Lexicon is an important part after cleaning data and before feature selection in sentiment analysis. So lexicon/corpus construction is generally viewed as a prerequisite for sentiment analysis. Since the middle of 20th century, many lexicons were built and developed such as 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).

However, there are few specialized lexicons for specialized domains. The two specialized lexicons are biolexicon and socialsent. As part of this research, a specialized lexicon, petrolexicon, was developed for the petroleum industry. The idea is to establish a SA lexicon network. The network where its center is SentiWordNet and SentiWordNet can be coupled with other domain lexicons such as business domain lexicon and petroleum domain lexicon. (Figure 1.1).

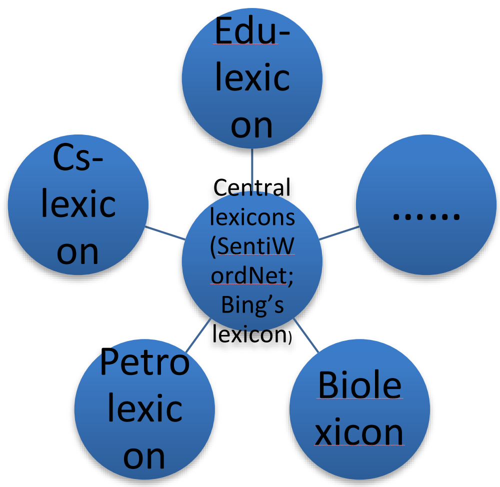

Figure 1.1. SA Lexicon Network

1.3. DESIGN SCIENCE

Design science research (DSR) focuses on exploring new methods for problems known or unknown (Alan R. Hevner, Salvatore T. March, Jinsoo Park, Sudha Ram, 2004). In this research, design science method will be used to structure methodology. The differences between DSR and widespread qualitative and quantitative methods have two key points: 1) DSR is trying to solve a generic problem and considered as an activity for testing hypothesis for future research. 2) The latter aims to explore real-life situations and come up with a theory that explains the current or past problems (Alan R. Hevner, Salvatore T. March, Jinsoo Park, Sudha Ram, 2004). Meanwhile, there are several steps to be followed if design science is used: 1) Start a specific space and find a solution. 2)

Generalize the problem and solution when moving to the generic space. (Alan R. Hevner, Salvatore T. March, Jinsoo Park, Sudha Ram, 2004).

In this paper, the design science method was used to guide the research. After a thorough literature review, the specialized lexicon, petrolexicon, was constructed for the petroleum industry. This is followed by an analysis of the three lexicons -- petrolexicon, biolexicon, and socialsent -- in text analysis. Finally, the suggestions on how to improve lexicon creation and the future research directions for sentiment analysis were presented.

2. LITERATURE REVIEW

2.1. SENTIMENT ANALYSIS

There are some main sentiment analysis techniques and methods such as machine learning, lexical dictionaries, natural language processing, psychometric scale, imagematics, and cloud-based technique (Matheus Araújo; Pollyanna Gonçalves; Meeyoung Cha; Fabrício Benevenuto, 2014). The machine learning needs a huge data resource due to the training part. Linguistic method is much easier than machine learning in the terms of operation and comprehension. Nowadays, these two methods are usually combined with each other. For example, in 'Sentiment Analysis-A Study on Product Features' (Meng, 2012), unsupervised and supervised machine learning include many linguistic rules and constraints that could improve the accuracy of calculations and classifications. Psychometric scale method is a more specific area. It mainly analyzes the mood of people and introduces the new smile or cry index as a formalized measure of societal happiness and sadness. Therefore, it is sometimes combined with lexical dictionaries. Lexical dictionary method is a development of lexical affinity and linguistic method to some extent. The simple method can be easy to operate if you are a beginner. It does not require too many data resources or calculations. Natural language processing is a technique that can implement the interaction between the human and computer. It can help us analyze the polarity of texts. SenticNet is based on the techniques. It is an approach that classifies texts as positive or negative (Matheus Araújo; Pollyanna Gonçalves; Meeyoung Cha; Fabrício Benevenuto, 2014).

Sentiment analysis techniques can be broadly classified into two categories – Machine Learning and Linguistic Method (as shown in Figure 2.1). Table 2.1 lists some papers in these two categories.

Machine learning is the most popular method right now in sentiment analysis area. In machine learning, there are also many techniques such as Support Vector Machine, Decision Tree, Neural Network Learning and so on. Also supervised machine learning and unsupervised machine learning are also playing an important role in machine learning.

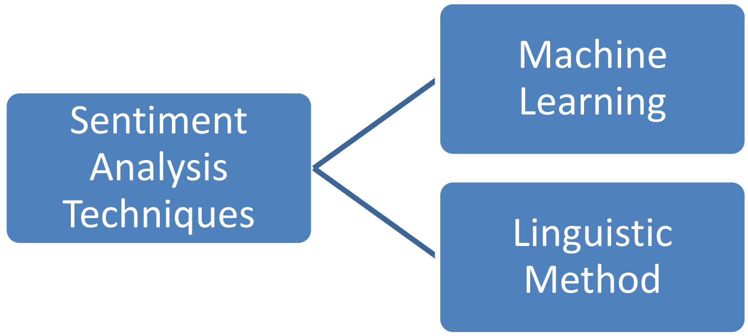

Figure 2.1. Sentiment Analysis Techniques

Table 2.1. Sentiment Analysis Techniques

	Paper Title	Techniques Used
Machine Learning	A Novel Hybrid HDP-LDA Model for Sentiment Analysis (Wanying Ding, Xiaoli Song, Lifan Guo, Zunyan Xiong, Xiaohua Hu, 2013)	This paper proposes a novel hybrid Hierarchical Dirichlet Process-Latent Dirichlet Allocation (HDP-LDA) model. This model can automatically determine the number of aspects, distinguish factual words from opinioned words, and effectively extracts the aspect specific sentiment words.
	Deep Learning for the Web (Kyomin Jung, Byoung-Tak Zhang, Prasenjit Mitra, 2015)	Deep learning is a machine learning technology that automatically extracts higher-level representations from raw data by stacking multiple layers of neuron-like units. The stacking allows for extracting representations of increasingly complex features without time-consuming, offline feature engineering.

Table 2.1. Sentiment Analysis Techniques (Cont.)

	Paper Title	Techniques Used
Machine Learning	iFeel: A Web System that Compares and Combines Sentiment Analysis Methods (Matheus Araújo; Pollyanna Gonçalves; Meeyoung Cha; Fabrício Benevenuto, 2014)	iFeel, a Web application system is introduced in this paper. iFeel can access seven existing sentiment analysis methods: Happiness Index, SentiWordNet, PANAS-t, Sentic-Net, and SentiStrength, SASA, Emoticons. iFeel can combine these methods to achieve high F-measure.
	A Comparative Study of Feature Selection and Machine Learning Techniques for Sentiment Analysis (Anuj Sharma, Shubhamoy Dey, 2012)	In this paper, machine learning based on Naïve Bayes, Support Vector Machine, Maximum Entropy, Decision Tree, K-Nearest Neighbor, Winnow, and Adaboost is applied.
	Sentence-based Plot Classification for Online Review Comments (Hidenari IWAI, Yoshinori HIJIKATA, Kaori IKEDA, Shogo NISHIDA, 2014)	Many shopping sites provide functions to submit a user review for a purchased item. Reviews of items, including stories such as novels and movies sometimes contain spoilers (undesired and revealing plot descriptions) along with the opinions of the review author. A system was proposed. Users see reviews without seeing plot descriptions. This system classifies each sentence in a user review as plot-reviews. Five common machine-learning algorithms were tested to ascertain the appropriate algorithm to address this problem.

Table 2.1. Sentiment Analysis Techniques (Cont.)

	Paper Title	**Techniques Used**
Machine Learning	Sentiment analysis in twitter using machine learning techniques (Neethu M S, Rajasree R, 2013)	The twitter posts about electronic products like mobiles, laptops and so on are analyzed by machine learning.
	Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning (Christos Troussas, Maria Virvou, Kurt Junshean Espinosa, Kevin Llaguno, Jaime Caro, 2013)	This paper uses Naïve Bayes Classifier to pattern the educational process and experimental results.
	Resolving Inconsistent Ratings and Reviews on Commercial Webs Based on Support Vector Machines (Xiaojing Shi, Xun Liang, 2015)	852,071 ratings and reviews from the Taobao website are the dataset. The support vector machine is used to solving inconsistent ratings and reviews.
	Sentiment Word Identification Using the Maximum Entropy Model (Xiaoxu Fei, Huizhen Wang, Jingbo Zhu, 2010)	The maximum-entropy classification model is constructed to detect sentiment words in an opinion sentence.

Table 2.1. Sentiment Analysis Techniques (Cont.)

	Paper Title	Techniques Used
Machine Learning	Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis (Geetika Gautam, Divakar yadav, 2014)	Dataset was preprocessed first, after that extracted the adjective from the dataset that has some meaning which is called feature vector, then selected the feature vector list and thereafter SVM, Naive Bayes, Maximum entropy corporation with WordNet are used to extract synonyms for the content feature.
Linguistic Method	Pathways for irony detection in tweets (Larissa A. de Freitas, Aline A. Vanin, Denise N. Hogetop, Marco N. Bochernitsan, Renata Vieira, 2014)	After observing the general data obtained and a corpus constituted by tweets, a set of patterns that might suggest ironic/sarcastic statements are proposed. The extracted texts for each pattern were analyzed by a judge in order to classify whether those texts represent ironic/sarcastic statements or not.
	Big Data Sentiment Analysis using Hadoop (Ramesh R, Divya G, Divya D, Merin K Kurian, Vishnuprabha V, 2015)	Sentiment Analysis on Big Data is achieved by collaborating Big Data with hadoop. The proposed approach is to identify texts into positive, negative and neutral position with Hadoop, which is a dictionary-based technique.

Figure 2.2 depicts the commonly used sentiment analysis methods. Representative papers are listed in Table 2.2.

As seen below, commonly used sentiment analysis methods are machine learning, lexical dictionaries, natural language processing, and psychometric scale. Natural language processing is not only applied to the big data area but also statistics and finance. It is useful to help researchers to recognize words, sentences, and paragraphs through computers. It has some popular tools here: OpenNLP, FudanNLP, Language Technology Platform (LTP). There are some difficult points during applying NLP. How to recognize every word is the first difficult. Since there are more than one meaning for many words. How to recognize the meaning of every word is another difficult.

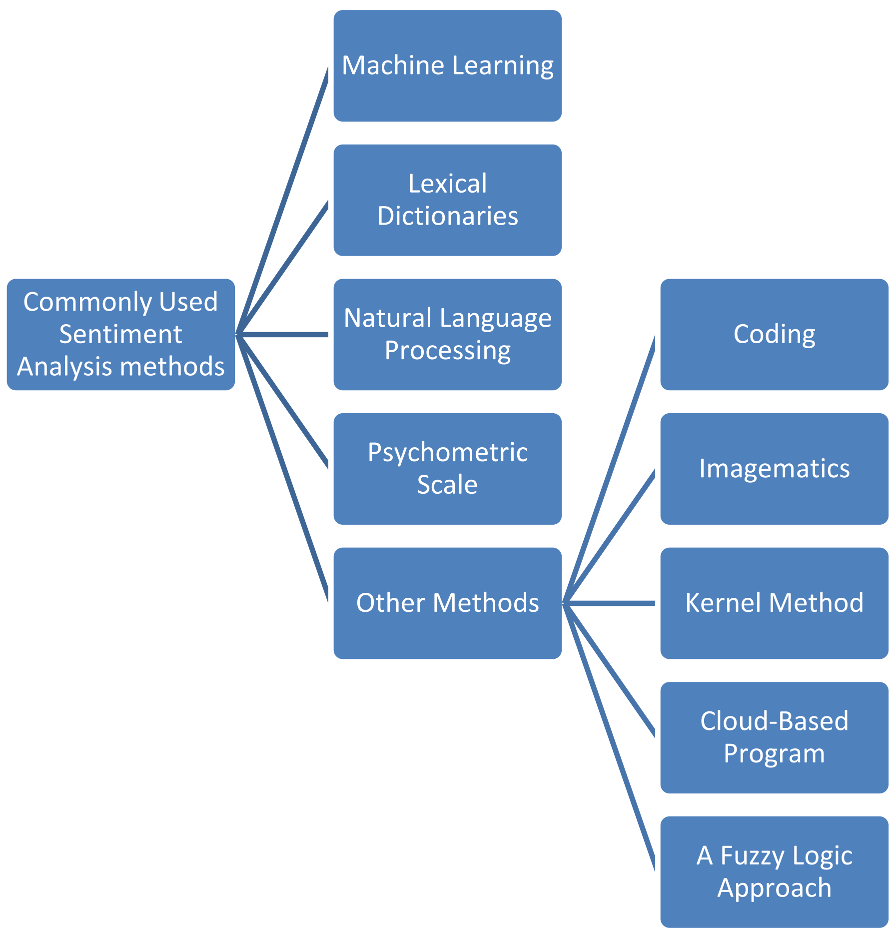

Figure 2.2. Commonly Used Sentiment Analysis Methods

Table 2.2. Commonly Used Sentiment Analysis Methods

	Paper Title	Techniques Used
Machine Learning	Same as those in Table 2.1.	
Lexical Dictionaries	Big Data Sentiment Analysis using Hadoop (Ramesh R, Divya G, Divya D, Merin K Kurian, Vishnuprabha V, 2015)	Sentiment Analysis on Big Data is achieved by collaborating Big Data with hadoop. The focus of this research was to device an approach that can perform Sentiment Analysis quicker because vast amount of data needs to be analyzed. Also, it had to ensure that accuracy is not compromised too much while focusing on speed.
	Microblogging sentiment analysis with lexical based and machine learning approaches (Maharani, 2013)	There are two main methods, which are lexical based machine learning and model based. This research is trying to classify tweets using those two methods.
	Chinese sentiment classification using a neural network tool — Word2vec (Zengcai Su, Hua Xu, Dongwen Zhang, Yunfeng Xu, 2014)	The neural network models based on word2vec is constructed to learn the vector representations in a higher dimension.

Table 2.2. Commonly Used Sentiment Analysis Methods (Cont.)

	Paper Title	**Techniques Used**
Lexical Dictionaries	Analysing market sentiment in financial news using lexical approach (Tan Li Im, Phang Wai San, Chin Kim On, Rayner, Patricia Anthony, 2013)	A lexicon-based approach to analyze financial news.
	Emotions on Facebook A Content Analysis of Mexico's Starbucks Page (Anatoliy Gruzd, Jenna Jacobson, Philip Mai, Barry Wellman, 2015)	Emoticons are the newly-developing language for sentiment analysis. It is simple to detect the polarity. But it is a huge project to establish a good-running emoticon-dictionary.
Natural Language Processing	iFeel: A Web System that Compares and Combines Sentiment Analysis Methods (Matheus Araújo; Pollyanna Gonçalves; Meeyoung Cha; Fabrício Benevenuto, 2014)	iFeel, a Web application system is introduced in this paper. iFeel can access to seven existing sentiment analysis methods: Happiness Index, SentiWordNet, PANAS-t, Sentic-Net, and SentiStrength, SASA, Emoticons. iFeel can combine these methods to achieve high F-measure.
	A Localization Toolkit for Sentic Net (Yunqing Xia, Xiaoyu Li, Erik Cambria, Amir Hussain, 2014)	A toolkit for creating non-English versions of SenticNet in a time- and cost-effective way is proposed.

Table 2.2. Commonly Used Sentiment Analysis Methods (Cont.)

	Paper Title	Techniques Used
Natural Language Processing	Enhanced SenticNet with Affective Labels for Concept-Based Opinion Mining (Soujanya Poria, Alexander Gelbukh, Amir Hussain, Newton Howard, Dipankar Das, Sivaji Bandyopadhyay, 2013)	Enhanced SenticNet with Affective Labels for Concept-Based Opinion Mining (Soujanya Poria, Alexander Gelbukh, Amir Hussain, Newton Howard, Dipankar Das, Sivaji Bandyopadhyay, 2013)
Psychometric Scale	Collective Smile: Measuring Societal Happiness from Geolocated Images (Saeed Abdullah, Elizabeth L. Murnane, Jean M.R. Costa, Tanzeem Choudhury, 2015)	This paper introduces the Smile Index as a standard measurement of general happiness in society.
	iFeel: A Web System that Compares and Combines Sentiment Analysis Methods (Matheus Araújo; Pollyanna Gonçalves; Meeyoung Cha; Fabrício Benevenuto, 2014)	iFeel, a Web application system is introduced in this paper. iFeel can access to seven existing sentiment analysis methods: Happiness Index, SentiWordNet, PANAS-t, Sentic-Net, and SentiStrength, SASA, Emoticons. iFeel can combine these methods to achieve high F-measure.

Table 2.2. Commonly Used Sentiment Analysis Methods (Cont.)

	Paper Title	Techniques Used
Psychometric Scale	Emotions on Facebook A Content Analysis of Mexico's Starbucks Page (Anatoliy Gruzd, Jenna Jacobson, Philip Mai, Barry Wellman, 2015)	Emoticons are the newly-developing language for sentiment analysis. It is simple to detect the polarity. But it is a huge project to establish a good-running emoticon-dictionary.
Current New Methods	Tweeting Live Shows: A Content Analysis of Live-Tweets from Three Entertainment Programs (Qihao Ji, Danyang Zhao, 2015)	In terms of the coding schema, each tweet was categorized by its *Language* (whether a tweet was written in English), *Relevancy* (whether it was relevant to the show), *Nature of Tweet* (whether it was a retweet, a tweet sent to a specific user, or a tweet sent to other users), and *Character Name* (whether the tweet contained any character's name from the show). Then coding procedure was processed.
	Towards Social Imagematics: sentiment analysis in social multimedia (Quanzeng You, Jiebo Luo, 2013)	This paper looks at not only textual but visual features in sentiment analysis.

Table 2.2. Commonly Used Sentiment Analysis Methods (Cont.)

	Paper Title	Techniques Used
Current New Methods	Enhanced Factored Sequence Kernel for Sentiment Classification (Luis Trindade, Hui Wang, William Blackburn, Philip S. Taylor, 2014)	A very active line of work focuses on the application of existing machine learning methods to sentiment analysis problems, for example support vector machine, which is a popular kernel method for text classification. This paper focuses on sequence kernels, which have been successfully employed for various natural language processing tasks including sentiment analysis.
	Tweeting Live Shows: A Content Analysis of Live-Tweets from Three Entertainment Programs (Qihao Ji, Danyang Zhao, 2015)	For data collection, Discovertext™, a cloud-based program was used.
	A Fuzzy Logic Approach for Opinion Mining on Large Scale Twitter Data (Li Bing, Keith C. C. Chan, 2014)	This paper proposes a novel matrix-based fuzzy algorithm, called the FMM system, to mine the defined multi- layered Twitter data.

2.2. LEXICON

Lexicon, as mentioned above, is an important tool that plays a role in sentiment analysis. Among existing lexicons, SentiWordNet is the most well-known and the most popular. 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, (2) algorithms used for annotating WordNet automatically, which now can refine the scores randomly. SentiWordNet 3.0 is trying to the improve part (2) (dell'Informazione).

2.3. APPLICATIONS OF SA

Same argue that sentiment analysis originates from customer products and services. Amazon.com is a representative example. Twitter and Facebook are also a hot and popular sites for many sentiment analysis applications.

The applications for sentiment analysis are many. Thousands of text documents can be processed by sentiment analysis in minutes, compared to the hours it would take a team of people to manually complete. The data can be words, sentences, or paragraphs. In China, sentiment analysis is called feeling analysis directly. It suggests that what feelings or mood people have can be analyzed. Digital numbers, on the other hand, cannot tell us what people feel. They can only tell us sales volume or the marketing distribution. Because SA can be efficient and can produce relatively high and reliable accuracy, many businesses and researchers are adopting text and sentiment analysis and combining them into their own research processes.

In business, the most widely used applications are in financial and sale marketing. For example, the Stock Sonar (www. Thestocksonar.com). It is a sentiment system where positive and negative assessments for each stock are updated every minute. In China, Yun Ma, Alibaba's CEO just created a miracle on Nov. 11th. There was a nation-wide shopping holiday on Taobao, Alibaba's shopping website, the biggest online shopping in China. There was 100 billion RMB sales volume in one minute after the online shopping holiday opened. Every product there has customer reviews and the customer reviews have already been summarized and separated into different groups: good product, bad product, nice looking, useful, and bad quality…customers can check them more easily than amazon. Because there are only raw data on Amazon, it is not easy for customers to find if there are some bad reviews. Sentiment applications in health care almost and mainly focus on reviews of drugs or health care service from patients. Figure and table 2.3 depicts some of the application areas for sentiment analysis.

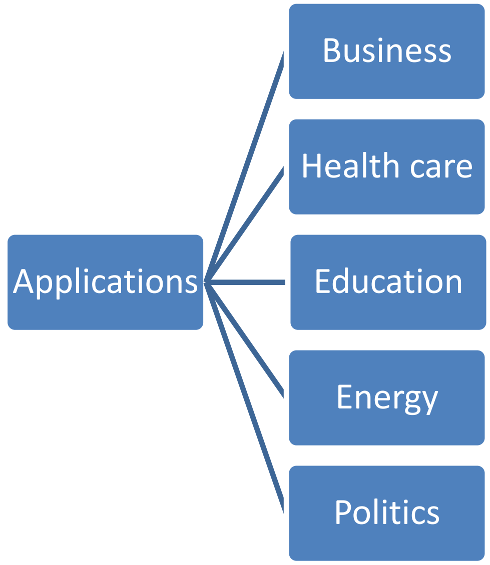

Figure 2.3. Applications of Sentiment Analysis

Table 2.3. Applications of Sentiment Analysis

	Paper Title	Applications
Business	A Large-Scale Sentiment Analysis for Yahoo! Answers (Onur Kucuktunc, B. Barla Cambazoglu, Ingmar Weber, Hakan Ferhatosmanoglu, 2012)	This paper uses a sentiment extraction tool to investigate the information like gender, education level, and age in a large online question-answering site. Analyzing what can affect the mood of customers will be applied in advertisement, recommendation, and search.
	Emotions on Facebook A Content Analysis of Mexico's Starbucks Page (Anatoliy Gruzd, Jenna Jacobson, Philip Mai, Barry Wellman, 2015)	Emoticons are the newly-developing language for sentiment analysis. It is simple to detect the polarity. But it is a huge project to establish a good-running emoticon-dictionary.

Table 2.3. Applications of Sentiment Analysis (Cont.)

	Paper Title	**Applications**
Business	Tweeting Live Shows: A Content Analysis of Live-Tweets from Three Entertainment Programs (Qihao Ji, Danyang Zhao, 2015)	In terms of the coding schema, each tweet was categorized by its *Language* (whether a tweet was written in English), *Relevancy* (whether it was relevant to the show), *Nature of Tweet* (whether it was a retweet, a tweet sent to a specific user, or a tweet sent to other users), and *Character Name* (whether the tweet contained any character's name from the show). Then coding procedure was processed. From this process, this paper explores whether live-tweets vary across different entertainment television programs in terms of the tweets' content.
Health Care	Tweet Analysis for User Health Monitoring (Ranjitha Kashyap, Ani Nahapetian, 2014)	Data analysis of social media can provide a wealth of information about the health of individual users, health across groups, and even access to healthy food choices in neighborhoods. The purpose of the analysis includes individually targeted healthcare personalization; determining health disparities, discover health access limitations, advertising, and public health monitoring.

Table 2.3. Applications of Sentiment Analysis (Cont.)

	Paper Title	**Applications**
Health Care	Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care (Altug Akay, Andrei Dragomir, Bj¨orn-Erik Erlandsson, 2014)	Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care (Altug Akay, Andrei Dragomir, Bj¨orn-Erik Erlandsson, 2014)
	Extracting Sentiment from Healthcare Survey Data: an Evaluation of Sentiment Analysis Tools (Despo Georgiou, Andrew MacFarlane, Tony Russell-Rose, 2015)	Extracting Sentiment from Healthcare Survey Data: an Evaluation of Sentiment Analysis Tools (Despo Georgiou, Andrew MacFarlane, Tony Russell-Rose, 2015)
Energy	Crude Oil- a Quick Market Sentiment Analysis (favresse, 2015)	This blog presents data plots from crude oil and oil price sentiment out of the millions of articles from news websites and social media.
	Production Estimation for Shale Wells with Sentiment-based Features from Geology Reports (Bin Tong, Hiroaki Ozaki, Makoto Iwayama, Yoshiyuki Kobayashi, Sahu Anshuman, Vennelakanti Ravigopal)	In this paper, to obtain data that describe the subsurface more exactly, information, including phrases that indicate possible bearing oil or gas and rock colors, is extracted from geology reports. Sentiments of the phrases are identified by sentiment analysis.

Table 2.3. Applications of Sentiment Analysis (Cont.)

	Paper Title	Applications
Energy	Analysis of Unstructured Data: Applications of text analytics and Sentiment Mining (Chakraborty)	This paper gives us ideas about how to extract meaningful customer intelligence to develop business operations and performance.
Education	SA-E: Sentiment Analysis for Education (Nabeela Altrabsheh, Mohamed Medhat Gaber, Mihaela Cocea, 2013)	Educational data mining (EDM) is becoming a hot topic right now. It mains to improve education levels through detecting students performance and how is students' study in real time. Students' feedback can be gained from some student response systems such as clickers and SMS, and social media.
	Potential Applications of Sentiment Analysis in Educational Research and Practice – Is SITE the Friendliest Conference? (Matthew Koehler, Spencer Greenhalgh, Andrea Zellner, 2015)	To be honest, SA in education is an underdeveloped area. In this paper, researchers explored some potential uses for SA in education. And there is a sample study that is using SA to compare the "friendliness" of two educational technology conferences and use these data to answer "Is SITE the friendliest conference?
Politics	Politics Sentiment (Politics Sentiment, 2012)	It is just a project in 2012. Collecting data about 2012 US presidential election from twitter and do SA are the main tasks in this project. The purpose is to predict the results of that election.

3. METHODOLOGY

In this research, design science approach is used – i.e., design and evaluation.

3.1. IDENTIFY THE PROBLEM

This research aims to study the impact of coupling a general lexicon with a specialized lexicon. Researchers focus on the petroleum industry in this research and developed a petrolexicon.

3.2. SOLUTIONS

There are three main steps for the construction of petrolexicon.

3.2.1. Original Data Extraction. Raw data comes mainly from two resources, Amazon engine oil product reviews and Onepetro database article. Nowadays, sentiment analysis in the petroleum industry has two main applications, analyzing user satisfaction for petroleum products and analyzing author's opinion in an article. Therefore, selecting those two data resources may make a contribution to improving lexicon's efficiency in the petroleum industry.

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 these traditional search engines to complete crawler's work (BAIKE, 2010): 1) many non-relative or less-relative webpages searched by traditional ones come out when different users may have different search goals and needs. 2) traditional search engines cannot afford some structured data. 3) traditional ones 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 can extract webpages from the Internet automatically. In this working process, web crawler needs to filter URLs which have no relations to the research according to specific web analysis algorithms, and extract and put useful URLs into a waiting list. Then it 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 aspect that is constructed.

3.2.2. LDA Model and NLP. Applying LDA-based topic modeling method is to extract aspects. For LDA-based topic modeling, each document d ϵ D of an unlabeled training corpus D is determined by a multinomial distribution θ. Given the topic z, a term t is calculated according to the multinomial distribution ϕ, determined by another hyper-parameter, a Dirichlet priori, β (Raymond Y.K. Lau, Stephen S.Y. Liao, Chunping Li, 2014).

Applying tf-idf measure is to select the topz most informative topics to represent product aspects. For the experiments reported in this paper, topz = 15 is adopted.

Since aspects have been selected, applying NLP parser is to extract opinion words. The combinations of aspects and sentiments are needed.

3.2.3. The Calculation of Polarity Scores. An amount of consumer reviews is used to establish the relations between sentiments and aspects through learning process. Combining the adjectives（opinion words）with the product aspects is a good step to establish pairs. The calculation is to give the pairs suitable polarity scores to present how good it is and how bad it is. The polarity score of a sentiment-aspect pair sa is defined as follows (Raymond Y.K. Lau, Stephen S.Y. Liao, Chunping Li, 2014):

$$WD(sa) = \tanh\left[\begin{array}{l} \dfrac{df(sa)}{\omega_{pos}} \times \Pr(pos|sa) \times \log_2 \dfrac{\Pr(pos|sa)}{\Pr(pos)} - \\ \dfrac{df(sa)}{\omega_{neg}} \times \Pr(neg|sa) \times \log_2 \dfrac{\Pr(neg|sa)}{\Pr(neg)} \end{array}\right] \tag{1}$$

$$polarity_{Ont}(sa) = \begin{cases} \dfrac{WD(sa) - \omega_{\omega d}}{1 - \omega_{\omega d}} & if\ WD(sa) > \omega_{\omega d} \\ -\left(\dfrac{|WD(sa)| - \omega_{\omega d}}{1 - \omega_{\omega d}}\right) & if\ WD(sa) < -\omega_{\omega d} \\ 0 & otherwise \end{cases} \tag{2}$$

4. EVALUATION AND COMPARISON

4.1. METHOD

As mentioned above, 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 the Biolexicon was used. A SocialSent lexicon was also used for the social network domain. The Petrolexicon and the Biolexicon are regarded as a specialized domain. 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 4.1 illustrates the analysis process.

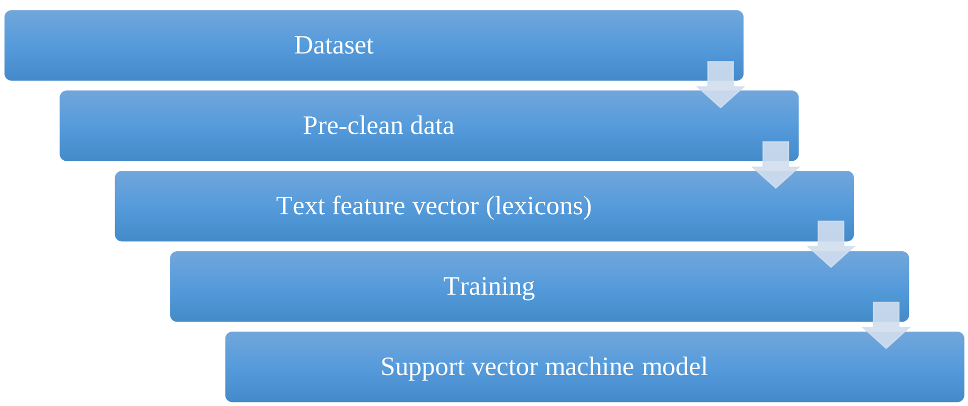

Figure 4.1. Analysis Procedure

4.2. PETROLEXICON, BIOLEXICON AND SOCIALSENT LEXICON

Petrolexicon is constructed using a fuzzy logic method. The items in this lexicon are pairs (aspects + opinion words). There have been 18,000 pairs in petrolexicon. Right now petrolexicon is only a small-scale domain lexicon. In the future, more items would be added to the lexicon. However, petrolexicon in this scale right now can already satisfy researchers' or companies' needs.

Biolexicon are relatively well developed since biostatistics has many well-developed techniques. Biolexicon includes over 2.2 M lexical entries and over 1.8 M terminology variants, as well as over 3.3 M semantic relations, including over 2 M synonym relations.

SocialSent is a set of code and datasets for better domain sentiment analysis. Items in this lexicon are mainly oral communication words from online communities.

4.3. RESULTS

The results are shown below for the three lexicons (Tables 4.1, 4.2, and 4.3). For example, for the Petrolexicon, compare the SentiWordNet with the Petrolexicon, and also compared the combination of SentiWordNet + Petrolexicon with SentiWordNet and Petrolexicon.

The results show that specialized lexicons (i.e., Petrolexicon and Biolexicon) seem to be performing better than SentiwordNet. Also, the combination of the central lexicon (i.e., in this case, SentiWordNet) and specialized lexicon seems to produce better results for Petrolexicon. For Biolexicon, the combination of the central lexicon and 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.

Table 4.1. Results for 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 4.2. Results for Biolexicon

Lexicon	Product Reviews	Petroleum News, Reports, and Blogs	Journal Articles
SentiWordNet	0.8518152	0.8364654	0.7615454
BioLexicon	0.9016564	0.9453122	0.9815457
SentiWordNet+Biolexicon	0.9015666	0.9423321	0.9815956

Table 4.3. Results for SocialSent

Lexicon	Product Reviews	Petroleum News, Reports, and Blogs	Journal Articles
SentiWordNet	0.7648151	0.8084144	0.8186455
SocialSent	0.7695952	0.8448518	0.8318656
SentiWordNet+SocialSent	0.7628494	0.8485265	0.8326451

5. DISCUSSIONS

The extension of the central lexicon with domain specific lexicons on demand is the goal of this research. Since petrolexicon is established and the practicability of the lexicon has been shown, petrolexicon can be a basic tool for sentiment analysis in the petroleum industry (by coupling it with a central lexicon such as SentiWordNet).

As discussed earlier, the combination of petrolexicon and SentiWordNet got a better result than petrolexicon itself. That is because petrolexicon only contains pairs of terminologies. Since the network of central lexicons and domain lexicons can be integrated into SA analysis, petrolexicon do not need to add general words.

Biolexicon contains general items and terminology variants. And also there is semantic structure in Biolexicon. Based on these features, biolexicon can be regarded as a well-developed domain lexicon. Petrolexicon can also be developed through this way, which may lead to a better sentiment analysis. Also, petrolexicon can add more terminology pairs to enlarge its scale.

6. CONTRIBUTIONS AND FUTURE RESEARCH

It is hypothesized that coupling a specialized lexicon to a general lexicon, such as SentiWordNet, will produce better results. The results suggest that this hypothesis is supported.

This study is expected to contribute to both academic researchers and practitioners. For academic research, a new stream of research is identifying and many more specialized lexicons can be created. Business or educational domain lexicon may be the next step. Research is also needed to investigate the best ways to couple the lexicons. For practitioners, this research suggests a new way to enhance the quality of sentiment analysis (i.e., coupling the central lexicon with specialized lexicon(s)).

BIBLIOGRAPHY

[1] Alan R. Hevner, Salvatore T. March, Jinsoo Park, Sudha Ram. (2004). Design Science in Information Systems Research. DIS Quarterly, 75-105.

[2] Altug Akay, Andrei Dragomir, Bj¨orn-Erik Erlandsson. (2013). Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care . IEEE Journal of Biomedical and Health Informatics , 210-218.

[3] Anatoliy Gruzd, Jenna Jacobson, Philip Mai, Barry Wellman. (2015). Emotions on Facebook: a content analysis of Mexico's Starbucks page. The 2015 International Conference on Social Media & Society. ACM.

[4] Anuj Sharma, Shubhamoy Dey. (2012). A comparative study of feature selection and machine learning techniques for sentiment analysis. The 2012 ACM Research in Applied Computation Symposium (pp. 1-7). ACM.

[5] BAIKE. (2010). Retrieved from Web Crawler: http://baike.baidu.com/link?url=vRXSRbTINNKhFO4ZlLMYMt1SYDfPCO9niSQU7U67As2sZGszEb_CDcovVSgHjuUp6U6ko4wji5258pwACRvtwh-J34quXfWXjwmN90TtoXX-PW5grbjNPlJCDkHzPBFw#ref_[1]_284853.

[6] Bin Tong, Hiroaki Ozaki, Makoto Iwayama, Yoshiyuki Kobayashi, Sahu Anshuman, Vennelakanti Ravigopal. (n.d.). Production Estimation for Shale Wells with. Retrieved from http://sentic.net/sentire/2015/tong.pdf.

[7] Chakraborty, G. (n.d.). Analysis of Unstructured Data: Applications of Text Analytics and. Retrieved from https://support.sas.com/resources/papers/proceedings14/1288-2014.pdf.

[8] 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), 2013 Fourth International Conference (pp. 1-6). Piraeus: IEEE .

[9] 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.

[10] Despo Georgiou, Andrew MacFarlane, Tony Russell-Rose. (2015). Extracting sentiment from healthcare survey data: An evaluation of sentiment analysis tools . Science and Information Conference (SAI) (pp. 352-361). London : IEEE.

[11] Favresse, j. (2015, 5 20). Crude Oil – A quick market sentiment analysis. Retrieved from https://amareos.com/blog/2015/05/20/crude-oil-a-quick-market-sentiment-analysis/.

[12] Geetika Gautam, Divakar yadav. (2014). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. Contemporary Computing (IC3), 2014 Seventh International Conference (pp. 437-442). Noida : IEEE.

[13] Hidenari IWAI, Yoshinori HIJIKATA, Kaori IKEDA, Shogo NISHIDA. (2014). Sentence-based Plot Classification for Online Review Comments. Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences (pp. 245-253). Warsaw : IEEE .

[14] Kyomin Jung, Byoung-Tak Zhang, Prasenjit Mitra. (2015). Deep Learning for the Web. the 24th International Conference on World Wide Web (pp. 1525-1526). International World Wide Web Conferences Steering Committee.

[15] Larissa A. de Freitas, Aline A. Vanin, Denise N. Hogetop, Marco N. Bochernitsan, Renata Vieira. (2014). Pathways for irony detection in tweets. The 29th Annual ACM Symposium on Applied Computing (pp. 628-633). SIGAPP.

[16] Li Bing, Keith C. C. Chan . (2014). A Fuzzy Logic Approach for Opinion Mining on Large Scale Twitter Data. The 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing (pp. 652-657). IEEE.

[17] Luis Trindade, Hui Wang, William Blackburn, Philip S. Taylor. (2014). Enhanced Factored Sequence Kernel for Sentiment Classification. Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences (pp. 519-525). Warsaw: IEEE.

[18] Maharani, W. (2013). Microblogging sentiment analysis with lexical based and machine learning approaches. Information and Communication Technology (ICoICT), 2013 International Conference (pp. 439-443). Bandung: IEEE.

[19] 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.

[20] Matthew Koehler, Spencer Greenhalgh, Andrea Zellner. (2015). Potential Applications of Sentiment Analysis in Educational Research and Practice – Is SITE the Friendliest Conference? In G. M. D. Slykhuis (Ed.), Proceedings of Society for Information Technology & Teacher Education International Conference (pp. 1348-1354). Las Vegas: Association for the Advancement of Computing in Education (AACE).

[21] Meng, Y. (2012, 4). Dissertations and Theses from the College of Business Administration. Retrieved from University of Nebraska–Lincoln : http://digitalcommons.unl.edu/businessdiss/28/.

[22] Nabeela Altrabsheh, Mohamed Medhat Gaber, Mihaela Cocea. (2013). SA-E: Sentiment Analysis for Education. In R. Neves-Silva, & R. N.-S. al. (Ed.), Intelligent Decision Technologies (pp. 353-361). Hampshire, UK: IOS Press.

[23] Neethu M S, Rajasree R, . (2013). Sentiment Analysis in Twitter using Machine. Computing, Communications and Networking Technologies (ICCCNT),2013 Fourth International Conference (pp. 1-5). Tiruchengode : IEEE.

[24] Onur Kucuktunc, B. Barla Cambazoglu, Ingmar Weber, Hakan Ferhatosmanoglu. (2012). A large-scale sentiment analysis for Yahoo! answers. the fifth ACm international conference on web search and data mining (pp. 633-642). ACM.

[25] Onur Kucuktunc, B. Barla Cambazoglu, Ingmar Weber, Hakan Ferhatosmanoglu. (2012). A large-scale sentiment analysis for Yahoo! answers. The fifth ACM International Conference on Web Search and Data Mining (pp. 633-642). ACM.

[26] Palace, B. (1996). Data Mining: What is Data Mining? Retrieved from Data Mining: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm.

[27] Petroleum Sentiment Analysis. (2015, 4 23). Retrieved from Sentiment Analysis: http://wenku.google.com/link?url=HkwN82RaHJnLeyig7d7-s3Q6QsbW3JPPse9DaMUqroRUsZ8-JwGB1RaGfacVzHhynew5G1GkGhNWoa-ohlFOG-3rg3OF7q_KNj9WHXv3kky.

[28] Politics Sentiment. (2012). Retrieved from USC Annenberg Innovation Lab: http://www.annenberglab.com/projects/politics-sentiment.

[29] Qihao Ji, Danyang Zhao. (2015). Tweeting live shows: a content analysis of live-tweets from three entertainment programs. The 2015 International Conference on Social Media & Society. ACM.

[30] Quanzeng You, Jiebo Luo. (2013). Towards social imagematics: sentiment analysis in social multimedia. The Thirteenth International Workshop on Multimedia Data Mining. ACM.

[31] Ramesh R, Divya G, Divya D, Merin K Kurian, Vishnuprabha V. (2015). Big Data Sentiment Analysis using Hadoop. IJIRST –International Journal for Innovative Research in Science & Technology.

[32] Ranjitha Kashyap, Ani Nahapetian. (2014). Tweet Analysis for User Health Monitoring. Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference (pp. 348-351). Athens : IEEE.

[33] Raymond Y.K. Lau, Stephen S.Y. Liao, Chunping Li. (2014, 4 24). Social Analytics: Learning Fuzzy Product Ontologies for Aspect-Oriented Sentiment Analysis. Decision Support Systems.

[34] Saeed Abdullah, Elizabeth L. Murnane, Jean M.R. Costa, Tanzeem Choudhury. (2015). Collective Smile: Measuring Societal Happiness from Geolocated Images. the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (pp. 361-374). ACM.

[35] Sentiment analysis. (2015, November 10). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Sentiment_analysis.

[36] Soujanya Poria, Alexander Gelbukh, Amir Hussain, Newton Howard, Dipankar Das, Sivaji Bandyopadhyay. (2013). Enhanced SenticNet with affective labels for concept-based opinion mining. IEEE Intelligent System , 31-38.

[37] Tan Li Im, Phang Wai San, Chin Kim On, Rayner, Patricia Anthony. (2013). Analysing Market Sentiment in Financial News Using Lexical Approach . Open Systems (ICOS), 2013 IEEE Conference (pp. 145-149). Kuching: IEEE.

[38] Text Mining. (2015, May 8). Retrieved from Statistics – Textbook: http://documents.software.dell.com/Statistics/Textbook/Text-Mining#index.

[39] Wanying Ding, Xiaoli Song, Lifan Guo, Zunyan Xiong, Xiaohua Hu. (2013). A Novel Hybrid HDP-LDA Model for Sentiment Analysis. Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences (pp. 329-336). Atlanta: IEEE.

[40] Wikipedia. (2016, 11 22). Retrieved from Web crawler: https://en.wikipedia.org/wiki/Web_crawler.

[41] Xiaojing Shi, Xun Liang. (2015). Resolving inconsistent ratings and reviews on commercial webs based on support vector machines . Service Systems and Service Management (ICSSSM), 12th International Conference (pp. 1-6). Guangzhou: IEEE.

[42] Xiaoxu Fei, Huizhen Wang, Jingbo Zhu. (2010). Sentiment Word Identification Using the Maximum Entropy Model. Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference (pp. 1-4). Beijing : IEEE.

[43] Yunqing Xia, Xiaoyu Li, Erik Cambria, Amir Hussain. (2014). A Localization Toolkit for SenticNet. Data Mining Workshop (ICDMW), 2014 IEEE International Conference (pp. 403-408). Shenzhen: IEEE.

[44] Zengcai Su, Hua Xu, Dongwen Zhang, Yunfeng Xu, . (2014). Chinese sentiment classification using a neural network tool — Word2vec . Multisensor Fusion and Information Integration for Intelligent Systems (MFI), 2014 International Conference (pp. 1-6). Beijing : IEEE.

VITA

Bo Yuan was born in Doongying, China. After finishing high school in 2009, she entered into China University of Petroleum (East China). She studied Geology and Geophysics degree, and Information Science and Technology degree at the Missouri University of Science and Technology between 2013 and 2017. She received a M.S. in Geology & Geophysics in 2015 and a M.S. in Information Science & Technology in May, 2017.