Sentiment Analysis at Sentence Level for Heterogeneous Datasets

Jawad Khan
Department of Computer Science & Engineering
Kyung Hee University
446-701, Republic of Korea
jkhanbk1@khu.ac.kr

Byeong Soo Jeong
Department of Computer Science & Engineering
Kyung Hee University
446-701, Republic of Korea
jeong@khu.ac.kr

Young-Koo Lee[1] and Aftab Alam[2]
Department of Computer Science & Engineering
Kyung Hee University
446-701, Republic of Korea
[1]yklee@khu.ac.kr,
[2]aftab@dke.khu.ac.kr

ABSTRACT
Sentiment analysis is a process used to automatically extract useful information from a collection of unstructured documents, in the form of customer reviews, blog comments or online social network sites data respectively. This useful information plays a significant role for decision making to know people's sentiments regarding online product reviews, services, events, entities. In this paper, rules based supervised machine learning method is proposed for sentiment analysis at sentence level for heterogeneous datasets. The text document in heterogeneous datasets is divided into two classes i.e., subjective and objective. We determine the semantic orientation and strength of subjective sentences based on the semantic score of its sentiment bearing words through SentiWordnet. Our experiment results show that the proposed method improves the performance of existing work for sentiment analysis in term of accuracy.

Categories and Subject Descriptors
[**Tools and Technologies**]: RapidMiner Studio 6.4.00, SentiWordnet 3.0.0, Penn Treebank annotation scheme, POS Tagger, Stanford Parsers.

General Terms
K-NN algorithm, Social Network Sites, Reviews, Classification

Keywords
Sentiment Analysis, Opinion Mining, Machine Learning, Linguistic Patterns

1. INTRODUCTION
Web 2.0 give birth to various social platforms including blogs, discussion forums, social seeking platforms and online social networks sites. These platforms enabled the folks to share their positive or negative opinions, experience, or sentiments about situations, events, persons or products. These expressions are always vital and beneficial for companies and has already been used for marketing strategies and decision making. In order to make beneficial such user generated data for business decision purpose, it must be analyzed first. But such data analysis is not feasible for human because of voluminous, unstructured in nature and involves cognitive overload. To overcome this problem, the field of sentimental analysis is used to analyze such data automatically. Sentiment analysis or opinion mining is the computational study that determines people's sentiments, opinions, emotions and perceptions toward entities, events and their attributes [1]. The main task of sentiment analysis is to determine the polarity of opinion sentences and classify it into positive and negative classes.

There are three different levels for the task of sentiment analysis: Document level, Sentence level, and Phrase level. Most of the existing work deals with document classification in which the whole document is classified into either positive or negative class based on the opinion expressed in them [2, 3]. Document level task is considered as a classification problem. In sentence level sentiment analysis first, it is determining to know whether the sentence is subjective or objective and if the sentence is subjective then the orientation of each subjective sentence is determined (whether the sentence is positive or negative) [4, 5]. Where the Phrase level sentiment analysis determine opinion expression and a phrase level classification is done [6, 7].

In this paper, we have proposed domain independent sentence based sentiment analysis method. Review documents and blog comments are comprised of both subjective and objective sentences [7, 8]. Subjective sentences are relevant and present user's attitude, view, or belief, while objective sentences are irrelevant and present information which is factual [7, 8]. Based on this idea, we filter out sentences that are objective and keep sentences that are subjective in nature. Then we extract sentiment phrases using linguistic patterns motivated by [2, 9], and determine the polarity and strength of subjective sentences based on the score of sentiment words.

In order to find the orientation of subjective sentences, we only consider Adjective, Adverb, Verb and their suitable combination which we refer to as linguistic patterns. In literature four types of words i.e., adjectives, adverbs, verbs and nouns or their combination [2, 8, 9] have been used for sentiment analysis because only these four types of words show the sentiment. Nouns are usually used for product features extraction and only some subjective sentences use nouns to express the sentiment.

The remaining paper is structured as follows: Section 2 describe related work, Section 3 present the detailed architectures and proposed methodology of the proposed system, Section 4 show

EDB, October 17-19, 2016, Jeju Island, Republic of Korea
ACM 978-1-4503-4754-9/16/10.
http://dx.doi.org/10.1145/3007818.3007848

experimental results and section 5 conclude our proposed method along with future direction.

2. Related Work

The research community is already active in the field of sentimental analysis and has been developed different technology, methods, strategies, and algorithms. Subjectivity classification is one of the renowned and initial classification scheme. Hatzivassiloglou et al. [10] tried to analyze different adjective features for subjectivity prediction and successfully distinguished opinion-based sentences from factual information. Similarly, further research studies on sentiment analysis have exploited different methods to distinguish subjective sentences from objective sentences and to determine the positivity or negativity of subjective sentences. Sentiment analysis techniques can be categorized as document-level, sentence-level, and phrase level. In document-level, the entire document is determined that whether it is positive or negative [2, 3]. Turney [2] presented an unsupervised learning technique based on the mutual information between document phrases and the words "excellent" and poor, where the mutual information is computed using statistical gathered by a research engine. Bo Pang and Lee [3, 11] proposed standard machine learning methods for the movie reviews classification. Sack [12] and Hearst [13] worked on sentiment level classification of whole documents using models inspired by cognitive linguistics. The sentence-level sentiment analysis method is based on determining opinion expressions or phrases. Sentence level sentiment classification include [5, 10, 14, 15]. Yu, Hong, and Hatzivassiloglou [16] proposed various unsupervised techniques for detecting opinion at the sentence level and used the results with a Bayesian classifier to determine whether a document is subjective or not. Wilson et al. [6] done Phrase level sentiment analysis and tried to enable automatically detect the contextual polarity for a bulky subset of sentiment expressions. Towards this end, they initially determined that whether an expression is neutral or polar and then disambiguated the polarity of the polar expressions. In [17], the authors proposed an unsupervised multi-view-based approach that uses a seed SA system together with domain-specific external information in order to mine a large corpus of messages for domain specific sentiment expression. In [18] presented machine learning algorithm for classification of the sentiment of twitter messages using distant supervision. There are two types of approaches: Corpus-based and dictionary based approach. Corpus-based approach finds opinion words with context specific orientation. Dictionary based approach depends on finding a set of seed opinion words, and searches the dictionary (WordNet) of their synonyms and antonyms. In [19, 20] lexicon based method has suggested for sentiment analysis.

Our work is different from the existing work in some aspects. We distinguished subjective sentences from objective sentences and then determined the semantic orientation and strength of subjective sentences based on its sentiment words through SentiWordnet. Our experiment on heterogeneous datasets show that our proposed method has improved accuracy as compared to the existing work.

3. Proposed system architecture

Figure 1. show the architecture of our proposed system for sentiment analysis. The input to the system are heterogeneous datasets (Training and Testing Datasets) from product reviews, Twitter tweets and Facebook comments on the web. The output is the sentiment results which show the orientation of subjective sentences.

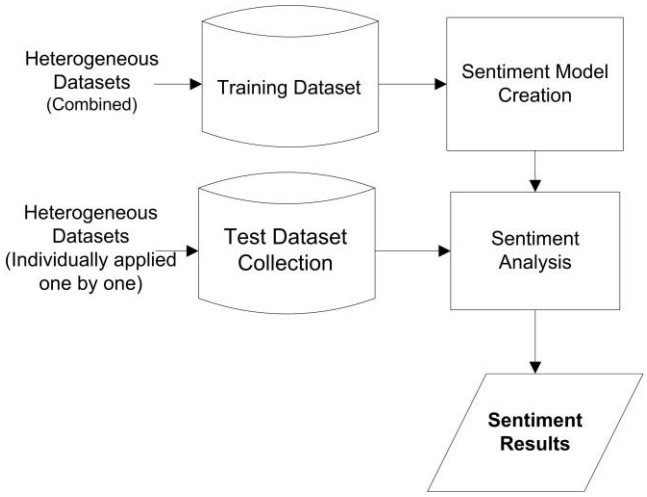

FIGURE 1. System Architecture

3.1 Heterogeneous Datasets

In order to prepare datasets for training and testing process, we apply pre-processing operation and convert the text into numeric vector for subjective sentences extraction. In the first step of preprocessing, we eliminate needless content, like dates, user names, tags and hash sign. We applied tokenization to represent the text documents as a bag of words, and to reduce the dimensionally of the documents words, we used special methods such as filtering stop words and stemming [21]. POS tagging is also used for linguistic patterns generation. POS tagging is the task that assign part of speech labels (or tags) to every word in a sentence. We used POS speech tags from Penn Treebank annotation scheme [22] and applied in the form of regular expression in Rapidminer for subjective sentences tagging. The purpose of regular expression is to identify sentiment words in linguistic patterns.

4. Sentiment Analysis Model Creation

In this section we trained our classifier on the annotated datasets of different subjective and objective sentences from heterogeneous datasets. We used different datasets which are freely available for research work [11, 18, 23] and made combined trained datasets from it. The training datasets is used to predict the probability of each word present in test datasets for subjectivity and objectivity purpose. we used the K-Nearest Neighbor algorithm (K-NN), [24] of Rapidminer [25] which classify test datasets into subjective and objective classes. Two other classifiers, Navie Bayes of Rapidminer and W-J48, from Weka were also used but their performance was not good. The usage of K-Nearest Neighbor algorithm (K-NN), is due to its best performance and following features.

- KNN is amongst the simplest of all machine learning algorithms
- It can be used even with few examples (small datasets).
- It works very well in low dimensions of datasets with respect to classes.

Due to the above features, we selected the K-NN as the learner for classification model.

5. Sentiment Analysis

In this section we determine the strength and semantic orientation (positive or negative) of subjective sentences. After the K-NN model creation we fed the individual test datasets from Movies reviews, product reviews, Facebook comments and tweets [11, 18, 23] to the trained model for the classification of text documents into subjective and objective class. We keep sentences that are subjective and filter out sentences that are objective. Subjective sentences are useful which show users sentiment, view or belief while objective sentences are not useful and present factual information.

5.1 Linguistic Patterns

We utilized syntactic rules for the suitable combination of sentiment words identification i.e., linguistic pattern given in table 1. In our previous work [26], we have used various patterns for sentiment analysis. In this work some patterns (pattern 6, 10, 11) are added without including noun words (usually noun is used for product feature extraction, very less sentences use nouns to express the sentiment). In existing work different patterns have been used by many researchers [2, 9]. We derived and designed linguistic patterns given in table 1 motivated by [2, 9].

TABLE 1. Linguistic patterns for the sentiment bearing words extraction.

P#	First Word	Second word	Third Word	Example
1	VB or VBD or VBN	JJ or JJR or JJS	-	look/VB excellent/JJ
2	RB or RBR or RBS	JJ or JJR or JJS	-	very/RB good/JJ
3	RB or RBR or RBS	RB or RBR or RBS	-	so/RB cool/RB
4	RB or RBR or RBS	VB or VBD or VBN	-	honestly/RB recommend/VB
5	VB or VBD or VBN	RB or RBR or RBS	-	work/VB perfectly/RB
6	JJ or JJR or JJS	RB or RBR or RBS	-	Joy/JJ Much/RB
7	JJ or JJR or JJS	VB or VBD or VBN	-	nice/JJ looking/VB
8	JJ or JJR or JJS	JJ or JJR or JJS	-	bloody/JJ awful/JJ
9	RB or RBR or RBS	RB or RBR or RBS	JJ or JJR or JJS	really/RB very/RB expensive/JJ
10	VB or VBD or VBN	RB or RBR or RBS	RB or RBR or RBS	exhausted/VB very/RB quickly/RB
11	VB or VBD or VBN	RB or RBR or RBS	JJ or JJR or JJS	Looking/VB very/RB nice/JJ

5.2 SentiWordnet for Semantic Orientation

In this section, we determine the strength and semantic orientation of subjective sentences through Sentiwordnet. SentiWordNet [27], a lexical resource used for sentiment analysis. SentiWrodnet is made up of synset from Wordnet. SentiWordnet automatically annotates polarity (Positive, negative or neutral) and sentiment score to all the synset of Wordnet. SentiWordnet assigns a numeric score to all the synset of Wordnet in the interval [0.0, 1.0]. The sum for each synset is 1.0. Using SentiWordnet dictionary in Rapidminer Studio the sentiment value is in the range [-1.0, 1.0] where -1.0 means very negative and 1.0 means very positive. From subjective sentences, we extract the sentiment-bearing words and determine their semantic scores and strength through Sentiwordnet. The final semantic score and strength of every subjective sentence are calculated based on the high numerical score of sentiment words. The average numerical score of sentiment words (which is either positive or negative) is assigned to the subjective sentences that contain these sentiment words. If there are two different sentiment words (positive and negative words) in subjective sentence, then in this case the word having high score value is assigned to subjective sentence.

For example:

"The camera is easy to use"
"The camera is very easy to use"

In the above sentences, we extract sentiment-bearing words (easy and very easy) then determine their semantic score and assign it to the individual sentence.

In the first sentence Sentiwordnet assign sentiment score 0.147 to sentiment-bearing word "easy" and in the second sentence it assigns sentiment score to sentiment-bearing words "very easy" 0.209 (easy = 0.147, very = 0.271, Sum = 0.147+0.271 = 0.418, Average = 0.418/2 = 0.209). It is clear that very easy is more positive than only easy. Negation and other words like "but" effect on the polarity of subjective sentences. In order to handle this issue we applied compositional semantic rules which are mentioned in [14].

6. Experiment results

We conducted our experiments using different datasets (Training datasets and Testing datasets) from 4 different heterogeneous datasets (Movie and product reviews, Facebook comments, and Twitter tweets) which are mentioned in below sub-section in details. For sentiment classification, we train the K-NN classifier using combined heterogeneous datasets [11, 18, 23]. We used SentiWordnet for the semantic orientation of subjective sentences. We also tested two other classifiers Naive Bayes and W-J48 from Weka but their performance was not good as compared to K-NN classifier.

6.1 Datasets

We used freely available movie review[1] datasets total 10,000 review sentences consists of 5000 subjective sentences and 5000 objective sentences. These datasets were used by [11] to classify review sentences into subjective and objective class. 500 reviews were taken from product reviews[2] collected by Hu and Liu [23].

[1] http://www.cs.cornell.edu/People/pabo/movie-review-data/
[2] www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

498 tweets were collected from twitter datasets [3] publically available for research work [18]]. 300 comments were extracted from Facebook. The sentiment results for the positive and negative sentences is given in the below tables.

Table 1. Accuracy of the sentiment orientation of positive and negative sentences for movie reviews

		Total	Positive	Negative	Accuracy
Subjective Sentences	Positive	167	150	17	0.898
	Negative	89	14	75	0.842

Table 2. Accuracy of the sentiment orientation of positive and negative sentences for product reviews

		Total	Positive	Negative	Accuracy
Subjective Sentences	Positive	200	180	25	0.900
	Negative	118	18	100	0.847

Table 3. Accuracy of the sentiment orientation of positive and negative sentences for Twitter tweets

		Total	Positive	Negative	Accuracy
Subjective Sentences	Positive	100	85	15	0.850
	Negative	125	17	108	0.864

Table 4. Accuracy of the sentiment orientation of positive and negative sentences for Facebook Comments

		Total	Positive	Negative	Accuracy
Subjective Sentences	Positive	195	168	27	0.861
	Negative	65	13	52	0.838

We compared the results of our proposed method for sentiment orientation at sentence level with the existing methods [18] and [23] on same datasets given in table 5. From the results it is clear that our proposed method performed best than existing methods in term of accuracy. Our proposed method achieved around 0.84-0.90 accuracy.

Table 5. Comparison of our proposed method with other methods using twitter tweets and product reviews datasets

Sentiment orientation at sentence level (Accuracy)			
Hu and Liu's approach (2004)	0.842	Our approach	0.873
Go, Bhayani, Huang approach (2009)	0.800	Our approach	0.857

[3] http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip

7. Conclusion

In this paper, we presented domain independent supervised machine learning method for sentiment analysis at sentence level utilizing syntactic rules. We extracted sentiment-bearing words from subjective sentences and determined its sentiment score and strength. We calculated the semantic orientation and strength of subjective sentences based on the high numerical score of sentiment words. From the results, we found that our proposed method obtained average 86 % accuracy at the sentence level. In the future, we will more improve the accuracy of our system by including some other features.

8. ACKNOWLEDGMENTS

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the Global IT Talent support program (IITP-2016-R01341610330001002) supervised by the IITP(Institute for Information and Communication Technology Promotion)

9. REFERENCES

[1] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, pp. 1-167, 2012.

[2] P. D. Turney, "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*, 2002, pp. 417-424.

[3] B. Pang, *et al.*, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 2002, pp. 79-86.

[4] T. Wilson, *et al.*, "Just how mad are you? Finding strong and weak opinion clauses," in *aaai*, 2004, pp. 761-769.

[5] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," in *Proceedings of the 20th international conference on Computational Linguistics*, 2004, p. 1367.

[6] T. Wilson, *et al.*, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the conference on human language technology and empirical methods in natural language processing*, 2005, pp. 347-354.

[7] J. Wiebe, "Learning subjective adjectives from corpora," in *AAAI/IAAI*, 2000, pp. 735-740.

[8] A. Kamal, "Subjectivity Classification using Machine Learning Techniques for Mining Feature-Opinion Pairs from Web Opinion Sources," *arXiv preprint arXiv:1312.6962*, 2013.

[9] W. Jin, *et al.*, "OpinionMiner: a novel machine learning system for web opinion mining and extraction," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 1195-1204.

[10] V. Hatzivassiloglou and J. M. Wiebe, "Effects of adjective orientation and gradability on sentence subjectivity," in *Proceedings of the 18th conference on Computational linguistics-Volume 1*, 2000, pp. 299-305.

[11] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, 2004, p. 271.

[12] W. Sack, "On the computation of point of view," in *AAAI*, 1994, p. 1488.

[13] M. Hearst, "Direction-based text interpretation as an information access refinement," *Text-based intelligent systems: Current research and practice in information extraction and retrieval*, pp. 257-274, 1992.

[14] K. Zhang, *et al.*, "SES: Sentiment elicitation system for social media data," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, 2011, pp. 129-136.

[15] J. M. Wiebe, *et al.*, "Development and use of a gold-standard data set for subjectivity classifications," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999, pp. 246-253.

[16] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003, pp. 129-136.

[17] Z. Ben-Ami, *et al.*, "Using Multi-View Learning to Improve Detection of Investor Sentiments on Twitter," *Computación y Sistemas*, vol. 18, pp. 477-490, 2014.

[18] A. Go, *et al.*, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, vol. 1, p. 12, 2009.

[19] X. Ding, *et al.*, "A holistic lexicon-based approach to opinion mining," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 2008, pp. 231-240.

[20] H. Li, *et al.*, "Extracting Verb Expressions Implying Negative Opinions," in *AAAI*, 2015, pp. 2411-2417.

[21] M. N. P. Katariya, *et al.*, "Text preprocessing for text mining using side information," 2015.

[22] A. AbeillÃ©, *Treebanks: Building and using parsed corpora* vol. 20: Springer Science & Business Media, 2003.

[23] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 168-177.

[24] D. T. Larose, "K-Nearest Neighbor Algorithm," *Discovering Knowledge in Data: An Introduction to Data Mining*, pp. 90-106, 2005.

[25] M. Hofmann and R. Klinkenberg, *RapidMiner: Data mining use cases and business analytics applications*: CRC Press, 2013.

[26] J. Khan and B. S. Jeong, "Summarizing Customer Review based on Product feature and opinion," presented at the ICMLC and ICWAPR, Jeju Island, South Korea, 2016.

[27] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of LREC*, 2006, pp. 417-422.