AN EXPERIMENTAL STUDY OF FEATURE EXTRACTION TECHNIQUES IN OPINION MINING

J. Ashok Kumar[1], S. Abirami[2]
Research Scholar[1], Assistant Professor[2]

[1, 2]Department of Information Science and Technology, Anna University, Chennai, India

ABSTRACT

The feature selection or extraction is the most important task in Opinion mining and Sentimental Analysis (OSMA) for calculating the polarity score. These scores are used to determine the positive, negative, and neutral polarity about the product, user reviews, user comments, and etc., in social media for the purpose of decision making and Business Intelligence to individuals or organizations. In this paper, we have performed an experimental study for different feature extraction or selection techniques available for opinion mining task. This experimental study is carried out in four stages. First, the data collection process has been done from readily available sources. Second, the pre-processing techniques are applied automatically using the tools to extract the terms, POS (Parts-of-Speech). Third, different feature selection or extraction techniques are applied over the content. Finally, the empirical study is carried out for analyzing the sentiment polarity with different features.

KEYWORDS

Sentiment Analysis, Opinion Mining, Feature Extraction, Polarity Classification, and Sentiment Polarity

1. INTRODUCTION

In this emerging trend, OMSA plays a vital role in social media contents, and it is used to determine the polarities of the contents into positive, negative and neutral for the product, user reviews, user comments, and etc. The sentiments are usually studied at the document level, sentence level, entity and feature or aspect level. The feature selection or extraction is one of the most important tasks in OMSA. An entity is the hierarchical representation of components and subcomponents. Each component is associated with set of attributes, whereas the large amount of documents is processed for sentiment with different features such as n-grams, part-of-speech, location based features, lexicon based features, syntactic features, structural or discourse features, and etc., [10].

In this paper, the experimental study is carried out with a freely available dataset for different feature selection techniques. This work is presented as a framework into the data collection process, pre-processing, feature selection, and experimental study for the performance evaluation. In section 2, the OMSA related works are presented. In section 3, the OMSA framework is described for data collection process, pre-processing technique and feature selection methods. In section 4, the experimental study is carried out. In section 5, the conclusion is presented with future challenges and developments.

DOI :10.5121/ijscai.2015.4102

2. RELATED WORKS

In OMSA, many feature selection or extraction techniques available. But only few related works are presented in this paper as follows. Jose M. Chenlo et al [10] demonstrated wide range of features such as n-grams, Part-of-speech, Location based features, Lexicon based features, Syntactic features, Structural or discourse features, and etc., for sentiment classification. In [9], the OMSA approach is presented with different frameworks and algorithms as a review and their results were compared and analyzed for readily available datasets. Farhan Hassan Khan et al. [4] proposed a new TOM framework to predict the polarity of words into positive or negative feelings in tweets, and to improve the accuracy level of this classification by using the noun and adjective features. Malhar Anjaria et al. [13] introduced a model to predict the election result by applying the user influence factor (re-tweets and each party garners) and extracting opinions using direct and indirect feature on the basis of the supervised algorithms such as simple probabilistic model, Uniform classification model, achieves maximum margin hyper plane, feed forward network, and dimensionality reduction by using the unigram, bigram and a unigram + bigram features. Tapia-Rosero A et al. [18] employed a method to detect similarity shaped membership functions in group decision making process by applying the get shape-string and feature-string algorithms. Jun ma et al. [11] stated a method to reduce the chance of applying inappropriate decisions in the multi-criteria group decision making (MCGDM) in three levels. Whereas the term set will be divided into several semantic-equal groups for the criterion, identifies an appropriate criterion, and each individual criterion to observe similarity of two opinions.

Xiaolin Zheng et al. [20] presented an unsupervised dependency analysis-based approach to extract appraisal expression pattern such as domain, aspect word, sentiment word, background word, and review. Vinodhini G et al. [19] introduced two frameworks by the combination of classifiers with principal component analysis (PCA) to reduce the dimension of feature set. By extracting the features from opinions expressed by users, and providing the positive, negative and neutral values of nouns, adjectives and verbs, Isidro Penalver-Martinez et al. [8] presented an innovative method called ontology based opinion mining to improve the feature-based opinion mining by employing the ontologies in selection of features and to provide a new vector analysis-based method for sentiment analysis. Alvaro Ortigosa et al. [1] introduced a new method is called sentiment extraction and change detection for extracting sentiments from texts. Using random independent, weighted versions, and random subspaces of the feature space respectively, Gang Wang et al. [5] conducted the comparative assessment to measure the performance of three ensemble methods (Bagging, Boosting, and Random Subspace) with five learners. Daekook Kang et al. [3] presented a new framework by combining the VIKOR ranking method and sentiment analysis for measurement of customer satisfaction in mobile services using the dictionary of attributes and dictionary of sentiment words which are expressed in verb phrases, adjective phrases and adverbial phrases. Sheng Huang et al. [17] proposed an automatic construction strategy of domain specific sentiment lexicon based on constrained label propagation by using sentiment term extraction nouns and noun phrases, adjectives, adverbs and its phrases. Arturo Montejo-Raez et al. [2] employed a method for sentiment classification by using weights of Word Net graph.

3. OMSA FRAMEWORK

The OMSA framework consists of the data collection process, pre-processing techniques, feature selection or extraction, and evaluation. This process is shown in Fig. 1, and described below in detail.

Figure 1. OMSA Framework

3.1 Data Collection Process and Pre-processing

The data collection process is the first stage in OMSA approach. In this stage, a freely available dataset is used for preprocessing the data. This readily available dataset is accessed for the purpose of experimental study. The collected dataset serves as input to the pre-processing stage and further the feature selection or extraction method has been applied over it to classify the polarity into positive, negative, and neutral. In this stage, the large amount of data is processed using the tools Gate Tool [7] and Semantria API [21].These tools are processing the data very quickly. Further, the process is analyzed for various feature extraction or selection method as discussed in section 3.3. The processing time is also compared in the above mentioned tools.

3.2 Feature Extraction methods

In this stage, all the documents available in the corpus are represented as Bag of words (BOW), and it is easy and very efficient method in text classification [3]. For this BOW, the sophisticated feature method needs to be applied to understand the documents in sentiment classification task. In this work, the POS, entity, phrases, weighting schemes and document features are considered for sentiment classification task.

3.2.1 Parts of Speech (POS)

In Parts of Speech (POS), the entire document content is represented as unigrams and N-grams, and which are divided into three groups. Group 1 consists of single words is called unigrams, and Group 2 consists of multiword is also called as N-grams. In this feature sets, the most relevant features are considered for sentiment classification.

3.2.2. Document Level

The document level features are considered to classify the textual reviews on a single topic into positive, negative, and neutral. In general, the document features determines the overall sentiment polarity.

3.2.3. Phrase Level

The phrase level features are used to determine whether an expression is positive, negative or neutral. Fourth, the entity features are used to extract name, location, address, and etc.

3.2.4. Weighting Scheme

The weighting scheme feature (tf-idf) for single word and multiword are considered for the sentiment classification in the document. The tf-idf value is calculated based on the below mentioned formula.

$$\text{IDF (Inverse Document Frequency)} = \text{Log}_2(^{N}/_{df})$$

$$\text{Weight}(t,d) = \text{tf}(t,d) \times \text{IDF}(t)$$

Where N is the total number of documents, df is represented as document frequency and tf is represented as term frequency, and t represents terms and d represents documents. An experimental study of the above methods has been carried out in this paper and discussed in the next section.

4. EXPERIMENTAL STUDY

An experimental procedure has been carried out with an extension of the OMSA approach [9]. In this approach, the Polarity Classification Algorithm (PCA) and evaluation procedure is applied to verify the accuracy. The evaluation procedure is tested with four different datasets namely Apple, Google, Microsoft, and Twitter. The contents or texts in the dataset are focused only on the topic of the companies as named above. Each datasets contained the tweet sentiment (positive, negative, neutral, and irrelevant) of the count 1313, 1381, 1415, and 1404 respectively. These datasets attained the accuracy of 96.73%, 96.89%, 96.96%, and 96.93% with the average accuracy of 96.88%. Also, the obtained average precision, recall, and F-measure are compared [9].

By using the above work as a model finding, the semantria trip advisor dataset is used for the sentiment polarity classification, and which contains 200 review documents. These documents are processed in GATE tool [7]. In this tool, unigrams, N-grams, and weighting scheme (tf-idf) are extracted using the plug-ins called TermRaider and PMI Score. Then, the extracted features are processed in Semantria API for the sentiment polarity [21]. The large amount of data is processed in less time for polarity classification. In this experiment, POS based features, entity, phrases, document features and weighting schemes are only considered as features. The dataset is classified into positive, negative, and neutral for the above mentioned features. The polarity scores are calculated as 1 for positive, -1 for negative, and 0 for neutral. The classification performance is evaluated and analyzed by using the confusion matrices, precision, recall, F-measure, and accuracy across the various features and the results are tabulated in Table 1 and Table 3.

Table 1. Types of features with count

Features	Polarity count
Single word	1273
Multiword	1237
Document level	636
Phrase feature	725
Tf-idf single word	1585
Tf-idf multiword	1492

Table 2. Confusion matrix

	X	Y	Z
X	tpX	eXY	eXZ
Y	eYX	tpY	eYZ
Z	eZX	eZY	tpZ

The class X, Y, and Z are represented as positive, negative, and neutral respectively in confusion matrices as shown in Table 2. The diagonal elements tpX, tpY, tpZ indicates that the correctly classified data for each class and the remaining elements are incorrectly classified data.

$$\text{Precision X} = \frac{\text{tpX}}{\text{tpX} + \text{eYX} + \text{eZX}}$$

Where tpX is the number of true positive predictions for the class X and eYX, eZX are false positives.

$$\text{Recall } X = \frac{tpX}{tpX + eXY + eXZ}$$

Where tpX is the number of true positive predictions for the class X, and eXY, eXZ are false negatives.

$$\text{F - measure} = \frac{2 \text{ x (Precision x Recall)}}{\text{Precision + Recall}}$$

$$\text{Accuracy} = \frac{2 \text{ x (TruePositive + TrueNegative + TrueNeutrals)}}{\text{TruePositive + FalsePositive + TrueNegative + FalseNegative + TrueNeutrals + FalseNeutrals}}$$

Table 3. The results obtained by using the confusion matrices

Feature Sets		Positive	Negative	Neutral	Precision (%)	Recall (%)	F-measure (%)	Accuracy (%)
Single Word	Positive	64	3	6	90.14	87.67	88.89	97.96
	Negative	7	7	5	46.67	36.84	41.18	
	Neutral	0	5	1176	99.07	99.58	99.32	
Multiword	Positive	329	3	6	97.92	97.34	97.63	97.90
	Negative	7	31	5	79.49	72.09	75.61	
	Neutral	0	5	851	98.72	99.42	99.07	
Document level	Positive	540	3	6	98.72	98.36	98.54	95.91
	Negative	7	47	5	85.45	79.66	82.46	
	Neutral	0	5	23	67.65	82.14	74.19	
Phrase level	Positive	576	3	6	98.80	98.46	98.63	96.41
	Negative	7	107	5	93.04	89.92	91.45	
	Neutral	0	5	16	59.26	76.19	66.67	
Tf-idf Single Word	Positive	78	3	6	91.76	89.66	90.70	98.36
	Negative	7	12	5	60.00	50.00	54.55	
	Neutral	0	5	1469	99.26	99.66	99.46	
Tf-idf Multiword	Positive	401	3	6	98.28	97.80	98.04	98.26
	Negative	7	56	5	87.50	82.35	84.85	
	Neutral	0	5	1009	98.92	99.51	99.21	

The OMSA approach is evaluated by using a single dataset is with six different feature levels to verify the predictive results. The polarity score is counted for all six features as shown in Table 1 and then the evaluated the polarity scores with 26 trained polarity scores. In this approach, the three polarity scores (positive, negative, and neutral) are considered for classification. The irrelevant score is not considered additionally in this OMSA approach. The obtained accuracy level is shown Table 3 and Fig.3 for the six different features (97.96% single word, 97.90% multiword, 95.91% document level, 96.41% phrase level, 98.36% tf-idf single word, 98.26% tf-idf multiword) in a single dataset. The precision, recall, and F-measure values will vary for different trained polarity scores and the accuracy level is considered as same all the respective feature sets.

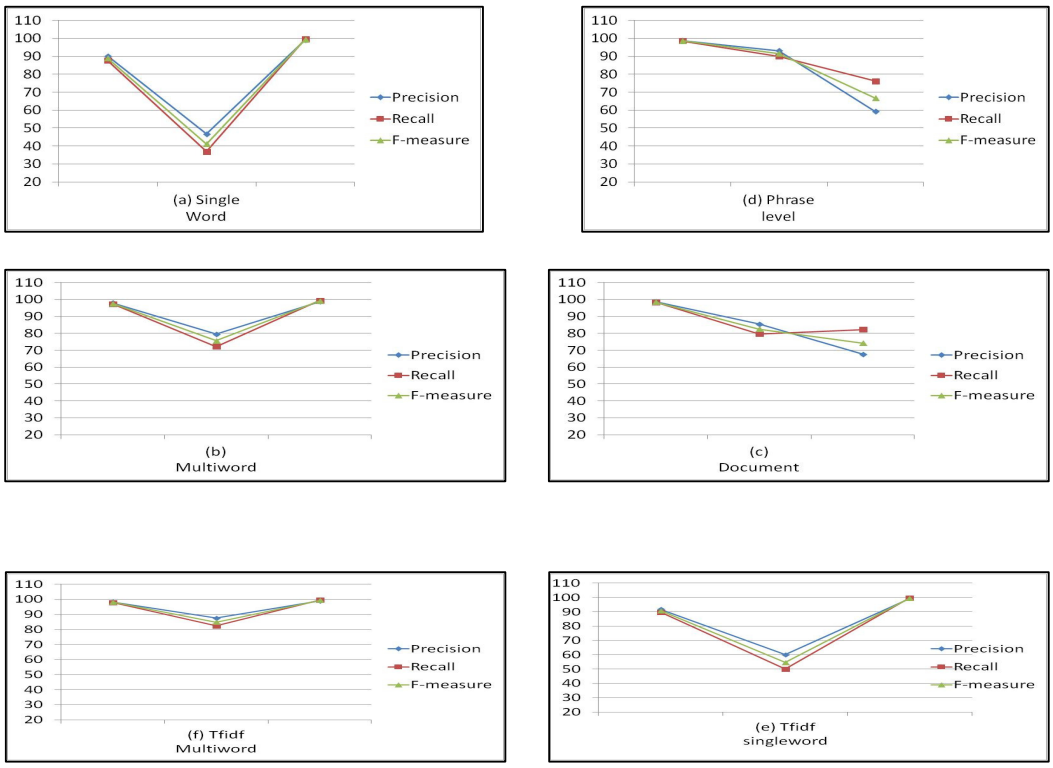

Figure 2. The graphical representation of Precision, Recall, and F-measures (a - f)

5. CONCLUSION AND FUTURE DIRECTIONS

The feature selection or extraction is one of the most important tasks in Opinion mining and Sentimental Analysis (OSMA) for calculating the polarity score. In this paper, we implemented the OMSA approach and analyzed the results by using a single dataset for different feature extraction or selection techniques namely single word, Multiword, Document Level, Phrase Level, Tf-idf single word and Tf-idf Multiword. The results seems to be different for above mentioned features. There are many challenges and future developments possible in OMSA approach like short length and irregular structure of the content such as named entity recognition, anaphora resolution, parsing, sarcasm, sparsity, abbreviations, poor spellings, punctuation and grammar, incomplete sentences, and the applications in [18] strategic planning, suitability analysis, and applications like fuzzy control, fuzzy time series to find the similarities, [11] missing value and unclear answers, [20] clause level and into aspect-based review summarization, sentiment classification, and personalized recommendation systems, [12] corresponding guidance and interference, [6] ontology, [12] weight of the edges, [17] constraints knowledge between sentiment terms and distinguishing the aspect-specific polarities.

REFERENCES

[1] Alvaro Ortigosa, Jose M. Martin, Rosa M. Carro.: Sentiment analysis in Facebook and its application to e-learning. Computers in Human Behavior. 31, 527-541 (2014)

[2] Arturo Montejo-Raez, Eugenio Martinez-Camara, M. Teresa Martin-Valdivia, L. Alfonso Urena-Lopez.: Ranked WordNet graph for Sentiment Polarity Classification in Twitter. Computer Speech and Language. 28, 93-107 (2014)

[3] Daekook Kang, Yongtae Park.: Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. Expert Systems with Applications. 41, 1041-1050 (2014)

[4] Farhan Hassan Khan, Saba Bashir, Usman Qamar.: TOM: Twitter opinion mining framework using hybrid classification scheme. Decision Support Systems. 57, 245-257 (2014)

[5] Gang Wang, Jianshan Sun, Jian Ma, Kaiquan Xu, Jibao Gu.: Sentiment Classification: The contribution of ensemble learning. Decision support systems. 57, 77-93 (2014)

[6] Gowsikhaa D, Abirami S, Baskaran R.: Construction of image ontology using low level features for image retrieval. Proceedings of the International Conference on Computer Communication and Informatics. 129-134 (2012)

[7] H. Cunningham, A. Hanbury, and S. Rüger. Scaling up high-value retrieval to medium-volume data. In H. Cunningham, A. Hanbury, and S. Rüger, editors, Advances in Multidisciplinary Retrieval (the 1st Information Retrieval Facility Conference). LNCS volume number: 6107, Lecture Notes in Computer Science, Vienna, Austria, May 2010. Sprin.

[8] Isidro Penalver-Martinez, Francisco Garcia-Sanchez, Rafael Valencia-Garcia, Miguel Angel Rodriguez-Garcia, Valentin Moreno, Anabel Fraga, Jose Luis Sanchez-Cervantes.: Feature-based opinion mining through ontologies. Expert Systems with Applications. 41, 5995-6008 (2014)

[9] J. Ashok Kumar, S. Abirami, S. Murugappan.: Performance analysis of the recent role of OMSA approaches in Online Social Networks. SAI-2014, 21-32, (2014).

[10] Jose M. Chenlo, David E. Losada.: An empirical study of sentence features for subjectivity and polarity classification. Information Sciences. 280, 275-288 (2014)

[11] Jun Ma, Jie Lu, Guangquan Zhang.: A three-level-similarity measuring method of participant opinions in multiple-criteria group decision supports. Decision Support Systems. 59, 74-83 (2014)

[12] Kyoungok Kim, Jaewook Lee.: Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction. Pattern Recognition. 47, 758-768 (2014)

[13] Malhar Anjaria & Ram Mohana Reddy Guddeti.: Influence factor based opinion mining of twitter data using supervised learning. Sixth IEEE International conference on communication systems and networks (COMSNETS). ISSN: 1409-5982, (2014)

[14] Ning Ma, Yijun Liu.: SuperedgeRank algorithm and its application in identifying opinion leader of online public opinion supernetwork. Expert Systems with Applications. 41, 1357-1368 (2014)

[15] Rui Xia, Chengqing Zong, Shoushan Li.: Ensemble of feature sets and classification algorithms for sentiment classification. Information Scinces. 181, 1138-1152 (2011).

[16] R.V. Vidhu Bhala, S. Abirami.: Trends in word sense disambiguation. Artificial Intelligence Review: An International Science and Engineering Journal. DOI 10.1007/s10462-012-9331-5, Springer, (2012)

[17] Sheng Huang, Zhendong Niu, Chongyang Shi.: Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. Knowledge-Based Systems. 56, 191-200 (2014)

[18] Tapia-Rosero A, A. Bronselaer, G. De Tre.: A method based on shape-similarity for detecting similar opinions in group decision-making. Information Sciences. 258, 291-311 (2014)

[19] Vinodhini G, Chandrasekaran RM.: Measuring quality of hybrid opinion mining model for e-commerce application. Measurement. 55, 101-109 (2014)

[20] Xiaolin Zheng, Zhen Lin, Xiaowei Wang, Kwei-Jay Lin, Meina Song.: Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification. Knowledge-Based Systems. 61, 29-47 (2014)

[21] https://semantria.com/