Postprint: SVM-Based Multi-Feature Fusion for Hierarchical Weibo Sentiment Classification

Authors: Yang Shuang, Chen Fen

Date: 2017-11-08T00:00:00+00:00

Abstract

[Objective] To more accurately identify netizens' attitudes and monitor online public opinion, a 5-level sentiment classification method based on SVM multi-feature fusion is proposed. [Method] From four aspects including part-of-speech features, sentiment features, sentence pattern features, and semantic features, 14 features such as verbs, nouns, sentiment words, and negation words are extracted, and the SVM method is used to perform 5-level classification of Weibo sentiment. [Results] Experimental results show that the proposed method achieves an accuracy of 82.40%, a recall of 81.91%, and an F-score of 82.10% for 5-level sentiment classification. [Limitations] The scale of the training corpus needs to be further expanded. [Conclusion] The method achieves favorable results in 5-level sentiment classification.

Full Text

Analyzing Sentiments of Micro-blog Posts Based on Support Vector Machine with Multi-Feature Fusion

Yang Shuang, Chen Fen
(School of Economics and Management, Nanjing University of Science & Technology, Nanjing 210094, China)

Abstract

[Objective] To more accurately identify netizen attitudes and monitor online public opinion, this paper proposes a five-level sentiment classification method based on Support Vector Machine (SVM) with multi-feature fusion. **[Methods]** We extracted fourteen features from four dimensions: part-of-speech features, sentiment features, syntactic patterns, and semantic features, including verbs, nouns, sentiment words, negation words, etc., and applied SVM to classify micro-blog sentiments into five levels. **[Results]** Experimental results show that the

proposed method achieves an accuracy of 82.40%, recall of 81.91%, and F-value of 82.10% for five-level sentiment classification. [**Limitations**] The scale of the training corpus needs to be further expanded. [**Conclusions**] The method demonstrates good performance in five-level sentiment classification.

Keywords: Micro-blog; Sentiment Orientation; Support Vector Machine; Parsing
Classification Number: G35; TP391

1. Introduction

Micro-blog has become China' s largest internet information dissemination platform in terms of user base, containing a wealth of subjective sentiment information. Sentiment classification of micro-blog posts enables rapid and accurate understanding of public demands and provides reliable support for online public opinion analysis.

Current research on micro-blog sentiment classification primarily employs semantic-based or machine learning-based methods, categorizing sentiments as positive/negative or positive/neutral/negative. However, this coarse-grained classification cannot precisely reflect netizens' emotional stances. In online public opinion scenarios, some netizens express absolute positions that are difficult to influence, while others hold unstable positions temporarily swayed by certain remarks. Therefore, three-level classification is overly absolute; a five-level approach (very positive, positive, neutral, negative, very negative) is more appropriate. While existing studies have focused on five-level classification of medium-to-long texts such as product reviews, research on five-level classification for short micro-blog texts remains limited.

This paper employs the SVM (Support Vector Machine) model and considers four feature dimensions—part-of-speech, sentiment, syntactic patterns, and semantics—to extract multiple sentiment resource features including word categories, sentiment words, sentiment intensity, sentiment scores, and semantic relationships for five-level micro-blog sentiment classification.

2. Related Work

Text sentiment classification techniques fall into two main categories: sentiment dictionary-based methods and machine learning-based methods. Dictionary-based approaches construct sentiment lexicons and calculate sentiment orientation values through specific algorithmic models to analyze text polarity. Kamps et al. [**?**] utilized WordNet' s synonym structure to compute semantic distances between new words and seed words for sentiment orientation calculation. Shen et al. [**?**] constructed dictionaries for negation words, degree adverbs, interjections, and sentiment words, achieving 80.6% accuracy in micro-blog sentiment orientation through rule-based calculation. Zheng et al. [**?**] similarly built sentiment dictionaries, incorporating semantic rules among sentiment words, nega-

tion words, and degree adverbs to compute sentiment polarity values.

Machine learning approaches treat sentiment classification as a special text classification task, training models on annotated datasets to determine text orientation. Pang et al. [**?**] pioneered the application of machine learning to sentiment classification, finding that unigram features with SVM achieved the best results with approximately 80% accuracy. Barbosa et al. [**?**] trained standard SVM classifiers on data from three Twitter sentiment analysis websites, reaching 81.3% precision. Davidov et al. [**?**] used hashtags and emoticons as features to train a KNN-like classifier for binary sentiment classification, achieving up to 86% accuracy. Xia et al. [**?**] employed syntactic parsing and CRFs to extract candidate evaluation objects for SVM-based micro-blog sentiment classification, attaining 91.4% accuracy.

Existing research predominantly focuses on three-level classification with relatively high accuracy. However, three-level classification inadequately addresses practical requirements, particularly for product reviews, prompting scholars to investigate five-level classification. Ding et al. [**?**] improved Conditional Random Fields (CRFs) through a two-layer approach: the first layer determined polarity, and the second layer provided five-level intensity classification. Wei et al. [**?**] conducted multi-level sentiment analysis for e-commerce product reviews, classifying them into five intensity levels (strongly negative, generally negative, neutral, generally positive, strongly positive) using complex sentence patterns and dictionary-based algorithms, though their work focused on document-level rather than sentence-level classification. Liao et al. [**?**] proposed a method based on bag-of-opinions and linguistic rules, calculating sentiment polarity values of collocation quadruples for five-level classification of automobile reviews, but required existing domain ontology features that couldn' t cover all documents. These methods primarily target product reviews; micro-blog texts are shorter and more informal, leaving a research gap in multi-level sentiment classification for short texts.

Building upon existing research, this paper uses Word2Vec to discover new online sentiment words, incorporates semantic features, employs syntactic dependency parsing to obtain semantic relationships with sentiment words, and proposes a method that fuses multiple sentiment resource features for five-level micro-blog sentiment classification using SVM.

3. Methodology

3.1 Dictionary Construction

We constructed three dictionaries for sentiment analysis: a sentiment dictionary, a negation dictionary, and a degree adverb dictionary. Based on HowNet' s sentiment lexicon, we expanded it using Word2Vec [**?**] to discover new online sentiment words. Word2Vec transforms words into vector representations based on semantic relationships, automatically identifying new sentiment words through semantic distance between word vectors. The principle involves using

a Huffman tree to build a statistical language model, employing shallow neural network backpropagation to transmit error losses and update model parameters and word vectors through iterative training, as shown in Equation (1):

$$\arg\max_{\theta} \log \prod_{w \in C} p(w|Context(w))$$

where θ represents neural network parameters, and C denotes the matrix vector $V \times K$ formed by all vocabulary in the corpus (V is vocabulary size, K is vector dimension). Using Huffman tree structure, $p(w|Context(w))$ in Equation (1) is defined as in Equation (2):

$$p(w|Context(w)) = \prod_{j=1}^{l_w - 1} \{\sigma([\![y_{w,j} = 1]\!] \cdot \theta_{j-1}^T x_w)\}^{1-y_{w,j}} \cdot \{1 - \sigma(\theta_{j-1}^T x_w)\}^{y_{w,j}}$$

where l_w represents the number of non-leaf nodes from root to leaf node corresponding to word w, with corresponding Huffman codes $y_{w,j}$, θ_{j-1} are neural network weight parameters, x_w is the word vector of w, and $\sigma(x)$ is the sigmoid function defined in Equation (3):

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

After iterative learning over the corpus with sliding windows, the model generates statistical language model parameters θ and word vector matrix C for all vocabulary.

Using Word2Vec to expand sentiment words followed by manual screening and adjustment, the final sentiment dictionary contains 4,566 positive and 4,371 negative sentiment words. The negation dictionary was built upon negation words from *Modern Chinese Grammar* and expanded to 28 words. The degree adverb dictionary started from HowNet' s degree adverb list and was manually expanded to 256 words, with weights assigned from 0.5 to 2 based on intensity. Partial degree adverbs and their weights are shown in .

3.2 Feature Selection

Different sentiment levels exhibit distinct semantic and syntactic characteristics. Feature selection is crucial for SVM classification, as accuracy, recall, and efficiency depend on appropriate feature selection. Through literature review and observation of real micro-blog corpora, we extracted 13 features across four dimensions: part-of-speech, sentiment, syntactic patterns, and semantics, as detailed in .

(1) Part-of-Speech Features
Micro-blog language is characterized by brevity and conciseness. Users often express ideas with single words or phrases lacking complete structure. Incorporating part-of-speech features helps parse sentence structure and assists sentiment judgment. Based on literature [**?**] and corpus observation, we selected verbs, adjectives, and adverbs as classification features.

(2) Sentiment Features
Sentiment words most directly reflect the poster' s emotional state, categorized as positive or negative. For five-level classification, sentiment intensity is crucial, manifested through degree adverb weights preceding sentiment words. For example, in "She looks very beautiful," the positive word "beautiful" is preceded by "very" (weight=2), increasing its intensity from 1 to 2. For multiple degree adverbs, the highest weight is taken as the intensity feature. Sentiment score is also included as a feature, with higher scores indicating clearer orientation, calculated as in Equation (4):

$$Score = \sum_{i=1}^{n} rawscore_i \times Intense_i$$

where n is the number of sentences in a micro-blog post, $rawscore_i$ is the base score of sentiment words in sentence i (+1, -1, or 0), and $Intense_i$ is the weight of modifying degree adverbs or negation words.

(3) Syntactic Pattern Features
Negation words can reverse sentiment orientation. For instance, "Had an unfun time today!" would be misclassified as positive without considering the negation "un." Thus, negation words are essential features. Question marks and exclamation marks indicate emphasis, with different frequencies expressing varying emotional intensities. Their occurrence counts serve as auxiliary features for sentiment discrimination.

(4) Semantic Features
Syntactic parsing analyzes whether word sequences conform to given grammars and extracts syntactic structures [**?**]. Dependency parsing reveals internal structure and relationships, providing comprehensive sentiment representation. We used the Stanford Parser [**?**] and extracted three relationship types based on corpus observation and literature [**?**]:

- **advmod (Adverbial Modifier):** Modifies adverb intensity. For "She looks very beautiful," the extraction yields advmod(beautiful, very), indicating "very" modifies the adjective "beautiful."
- **amod (Adjectival Modifier):** An adjective modifying a noun phrase. For "This is truly a godly reply," the result is amod(reply, godly), showing "godly" modifies "reply."
- **nsubj (Nominal Subject):** Modifies nominal subjects. For "A different anti-Japanese war drama, good!" the extraction is nsubj(good, drama),

indicating "good" modifies the nominal subject "drama."

3.3 Sentiment Classification Model

Micro-blog corpora contain noise such as #topics#, URLs, and @mentions that lack user opinions and may affect segmentation and POS tagging. We first filter these elements before processing. The NLPIR2016 segmentation tool from the Institute of Computing Technology, Chinese Academy of Sciences [**?**], performs word segmentation and POS tagging on filtered corpora. We selected SVM as the classification model, representing training and test sets with extracted features. The training set trains the SVM model with optimized parameters, and the test set is classified by the trained model. The classification framework is shown in [Figure 1: see original paper].

4. Experiments

4.1 Experimental Data

We used partial COAE2014 micro-blog evaluation corpora, manually annotating 5,000 posts across five levels: "very positive," "positive," "neutral," "negative," and "very negative." Annotation was performed by research team members, with distribution shown in .

Annotation primarily relied on degree adverb levels and punctuation. Statements containing high-intensity degree adverbs like "very" express stronger emotions than those with low-intensity or no degree adverbs. Similarly, multiple exclamation or question marks indicate stronger emotions than punctuation-free statements. For example: - Example (1): "This jade looks nice!" - Example (2): "This jade is really very very beautiful!!!"

Example (2) expresses stronger sentiment than (1), so (1) is labeled +1 while (2) is labeled +2.

4.2 Feature Extraction Results

After annotation, we performed segmentation, POS tagging, and feature extraction as described in Section 3.2. All programs were written in Java on the Eclipse platform under Windows 7 64-bit with 4GB RAM. Partial feature extraction results are shown in .

4.3 Model and Evaluation Metrics

We used LibSVM for SVM training and classification [**?**], splitting each sentiment category 4:1 into training and test sets. Features were normalized before training to improve speed. Default LibSVM parameters were used: SVM type C_SVC with RBF kernel. Accuracy, recall, and F1-score served as evaluation metrics.

5. Experimental Results and Analysis

5.1 Impact of Feature Combinations

We verified the effect of different feature combinations on classification using accuracy as the metric, with results shown in .

The results demonstrate that using all features yields the highest accuracy of 82.40%. Sentiment word features contribute most significantly, improving accuracy by 23.33%. Degree adverb weights provide the second-largest improvement (0.83%), while other features offer slight enhancements.

5.2 Comparative Evaluation

We compared our method with the cascaded CRFs approach [**?**], which is commonly used for five-level classification by transforming it into a coarse-to-fine process. The method first performs three-level classification, then incorporates POS, evaluation word, conjunction, and polarity features for five-level classification, achieving 83.75% accuracy on COAE2008 tasks. Applying this method to our corpus yielded the comparison results in .

Our SVM-based method achieves 82.40% accuracy, substantially higher than cascaded CRFs (75.31%). Recall is also significantly improved at 81.91% versus 73.30%. The F1-score of 82.10% represents a 7.80% improvement over cascaded CRFs (74.30%). The cascaded CRFs features are designed for medium-to-long texts and perform poorly on short micro-blog texts. Our method expands the sentiment dictionary using Word2Vec and selects features across multiple dimensions, achieving high accuracy for five-level micro-blog sentiment classification.

6. Conclusion

This paper proposes an SVM-based method for five-level sentiment classification of micro-blog posts. By leveraging part-of-speech, sentiment, syntactic pattern, and semantic features, the method achieves favorable results compared to existing five-level classification approaches.

The primary limitation is the relatively small training corpus, particularly for "very positive" and "very negative" categories. Generally, larger training corpora yield more accurate models. Future work will expand the training data to further improve model accuracy.

References

[1] Wang Xuemeng, Wang Yuping. Research of Emergency Network Public Sentiment Warning Based on the Analysis of Emotional Tendency [J]. Journal of Southwest University of Science and Technology: Philosophy and Social Science Edition, 2016, 33(1): 63-66.

[2] Kamps J, Marx M, Mokken R J, et al. Using WordNet to Measure Semantic Orientations of Adjectives [C]// Proceedings of the 4th International Conference on Language Resources and Evaluation. 2004.

[3] Shen Y, Li S, Zheng L, et al. Emotion Mining Research on Micro-blog [C]// Proceedings of the 1st IEEE Symposium on Web Society. 2009.

[4] Zheng Cheng, Yang Xi, Zhang Jigeng. Micro-blog Sentiment Analysis of Combined Sentiment Dictionary and Rules [J]. Computer Knowledge and Technology, 2014, 10(13): 3111-3113.

[5] Zhang Yang, Liu Xiaoxia, Sun Kailong, et al. Research on Text Orientation Identification Based on Emotional Description [J]. Computer Engineering and Applications, 2015, 51(4): 158-161, 195.

[6] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment Classification Using Machine Learning Techniques [C]// Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. 2002.

[7] Borbosa L, Feng J. Robust Sentiment Detection on Twitter from Biased and Noisy Data [C]// Proceedings of the 23rd International Conference on Computational Linguistics. Beijing: Tsinghua University Press. 2010.

[8] Davidov D, Tsur O, Rappoport A. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys [C]// Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010: 241-249.

[9] Xia Mengnan, Du Yongping, Zuo Benxin. Micro-blog Opinion Analysis Based on Syntactic Dependency and Feature Combination [J]. Journal of Shandong University: Natural Science, 2014, 49(11): 22-30.

[10] Ding S, Jiang T, Wen N. Research on Sentiment Orientation of Product Reviews in Chinese Based on Cascaded CRFs Models [C]// Proceeding of the 2012 International Conference on Machine Learning and Cybernetics (ICMLC 2012). IEEE, 2012.

[11] Wei Jingjing, Wu Xiaoyin. Research on Multi-level Sentiment Analysis System of E-Commerce Product Review and Implementation [J]. Software, 2013, 34(9): 65-67, 94.

[12] Liao Jian, Wang Suge, Li Deyu, et al. The Bag-of-Opinions Method for Car Review Sentiment Polarity Classification [J]. Journal of Chinese Information Processing, 2015, 29(3): 113-120.

[13] Word2Vec [EB/OL]. [2015-01-12]. http://word2vec.googlecode.com/svn/trunk/.

[14] Liu Z, Yu W, Chen W, et al. Short Text Feature Selection for Micro-blog Mining [C]// Proceedings of the International Conference on Computational Intelligence and Software Engineering. IEEE, 2010.

[15] Wu Mingfen, Chen Tao. Sentences Tendency Judgement by POS and Dependency Based on SVM [J]. Journal of Wuyi University: Natural Science Edi-

tion, 2012, 26(4): 66-71.

[16] Liu Haitao. Dependency Grammar: From Theory to Practice [M]. Beijing: Science Press, 2009.

[17] Stanford Parser [EB/OL]. [2015-06-16]. http://nlp.stanford.edu/software/lex-parser.shtml.

[18] Peng Yue. Internet Opinion Leader Detection Based on Text Sentiment Analysis [D]. Nanjing: Nanjing University of Science and Technology, 2014.

[19] NLPIR/ICTCLAS [EB/OL]. [2015-12-02]. http://ictclas.nlpir.org/.

[20] LibSVM [EB/OL]. [2015-07-12]. https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Author Contributions

Yang Shuang: Dictionary construction, program design, manuscript drafting.
Chen Fen: Research conception, study design, final manuscript revision.

Conflict of Interest

All authors declare no conflict of interest.

Supporting Data

Supporting data [1-3] are available in the journal' s online version at http://www.infotech.ac.cn; supporting data [4] is self-archived by the authors at E-mail: doubleyou1001@163.com.

[1] Yang Shuang, Chen Fen. senti_dic.rar. Positive and negative sentiment dictionaries expanded using Word2Vec.
[2] Yang Shuang, Chen Fen. weight_dic.txt. Degree adverbs with manually assigned weights.
[3] Yang Shuang, Chen Fen. TestData.rar. Manually annotated test data.
[4] Yang Shuang. Senti_analysis.rar. Feature selection program.

Received: 2016-08-29
Revised: 2016-10-26

Note: Figure translations are in progress. See original paper for figures.