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

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—part-of-speech features, sentiment features, sentence pattern features, and semantic features—14 features including 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 method achieves an accuracy of 82.40%, a recall rate of 81.91%, and an F-value of 82.10% for 5-level sentiment classification. [Limitations] The scale of the training corpus needs to be further expanded. Conclusion The method achieves good results in 5-level sentiment classification.

Full Text

Abstract

This paper proposes a five-level sentiment classification method based on Support Vector Machine (SVM) with multi-feature fusion to more accurately identify netizen attitudes and monitor online public opinion. The method extracts fourteen features from four dimensions: part-of-speech characteristics, sentiment features, sentence pattern features, and semantic features, including verbs, nouns, sentiment words, and negation words. Using SVM, we classify micro-blog sentiments into five levels. Experimental results demonstrate that the proposed method achieves 82.40% precision, 81.91% recall, and 82.10% F-value for five-level sentiment classification. The primary limitation is that the scale of the training corpus needs further expansion. The method demonstrates effective performance for five-level sentiment classification of micro-blog posts.

Keywords: Micro-blog, Sentiment Analysis, Support Vector Machine, Parsing

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 opinion, providing reliable foundations for online public opinion analysis. Current research on micro-blog sentiment classification primarily employs semantic-based or machine learning methods, categorizing sentiments as positive/negative or positive/neutral/negative. However, this approach cannot precisely reflect netizens' emotional stances [1]. In online public opinion contexts, some netizens express absolute positions on events that are difficult to influence, while others hold unstable positions merely temporarily swayed by certain remarks. Therefore, three-level classification is overly absolute; a five-level approach comprising very positive, positive, neutral, negative, and very negative is more appropriate. Existing sentiment classification research predominantly focuses on five-level classification of medium-to-long texts such as product reviews, with limited investigation into five-level classification of short micro-blog texts.

This paper employs the Support Vector Machine (SVM) model for classification, extracting multiple sentiment resource features—including part-of-speech, sentiment words, sentiment intensity, sentiment scores, and semantic relationships—from four dimensions: part-of-speech characteristics, sentiment features, sentence pattern features, and semantic features to achieve five-level sentiment classification of micro-blog posts.

Related Work

Text sentiment classification techniques fall into two main categories: sentiment dictionary-based methods and machine learning methods. Sentiment dictionary-based approaches construct sentiment lexicons and calculate sentiment orientation values through specific algorithmic models for polarity analysis. Kamps et al. [2] utilized WordNet' s synonym structure to compute semantic distances between new words and seed words for sentiment orientation calculation. Shen et al. [3] constructed dictionaries for negation words, degree adverbs, interjections, and sentiment words, establishing rules to calculate micro-blog sentiment orientation with 80.6% accuracy. Zheng et al. [4] incorporated semantic rules among sentiment words, negation words, and degree adverbs into micro-blog sentiment calculation, combining sentiment dictionaries with rules to compute sentiment polarity values for classification.

Machine learning methods treat sentiment classification as a special text classification task, training models on annotated datasets to determine text orientation [5]. Pang et al. [6] applied machine learning to sentiment classification, finding that Unigram features with SVM achieved the best results with approximately 80% accuracy. Barbosa et al. [7] trained a standard SVM classifier using data from three Twitter sentiment analysis websites, achieving 81.3% precision. Davi-

dov et al. [8] used hashtags and emoticons as features to train a KNN-like classifier for binary sentiment classification, reaching 86% accuracy. Xia et al. [9] employed syntactic analysis and CRFs to extract candidate evaluation objects for SVM-based micro-blog sentiment classification, attaining 91.4% accuracy.

Current research primarily 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. [10] improved Conditional Random Fields (CRFs) through a two-layer approach: the first layer determined polarity, and the second provided five-level intensity classification, achieving favorable results. Wei et al. [11] 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. However, this method focused on document-level rather than sentence-level five-level classification. Liao et al. [12] proposed a multi-level sentiment classification method based on the Bag-of-Opinions model and linguistic rules, calculating collocation quadruple polarity values for vector representation and using cosine similarity for five-level classification of automobile reviews. This approach required existing domain ontology features and could not cover all documents with extracted sentiment collocations. These methods primarily target product reviews; micro-blog posts are shorter and more informal, leaving five-level classification of short micro-blog texts underexplored.

Building upon existing research, this paper employs Word2Vec to discover novel sentiment terms, incorporates semantic features, and utilizes syntactic dependency parsing to obtain semantic relationships with sentiment words, proposing an SVM-based method that fuses multiple sentiment resource features for five-level micro-blog sentiment classification.

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 the HowNet sentiment dictionary, we used Word2Vec [13] to discover novel online sentiment terms. Word2Vec transforms words into vectors based on semantic relationships, automatically identifying new sentiment terms through semantic distances between vectors. The principle involves a probabilistic model in statistical language modeling using Huffman trees. For training corpora, back propagation (BP) in shallow neural networks transmits error loss while simultaneously updating model parameters and word vectors. After multiple iterations, the statistical language model is generated along with word vectors for all vocabulary, as shown in Equation (1).

$$\arg\max \log($$

where θ represents neural network parameters, and C denotes the matrix vector $V \times K$ for all corpus vocabulary (V is vocabulary size, K is vector dimension). Using Huffman tree data structures, the term in Equation (1) is defined in Equation (2).

y_{wl} represents the number of non-leaf nodes from the root to leaf node w_y, with corresponding Huffman codes y_w. For neural network weight parameters $i\theta-$, x_{wC} denotes the word vector for w_x, and $\sigma(x)$ is the sigmoid function defined in Equation (3).

Through sliding window iterations over the corpus, the model obtains statistical language model parameters θ and the word vector matrix C.

Using Word2Vec to expand sentiment terms with manual screening and adjustment, the final sentiment dictionary contains 4,566 positive and 4,371 negative sentiment words. The negation dictionary, based on negation words from *Modern Chinese Grammar* with extensions, contains 28 negation words. The degree adverb dictionary, built upon HowNet' s degree adverbs with manual collection, contains 256 degree adverbs assigned weights from 0.5 to 2 based on intensity. Table 1 shows sample degree adverbs and their weights.

Table 1 Degree Adverb Examples - absolutely, extremely, very, super, overly ···(Weight: 2.0) - very, how, even more, extremely···(Weight: 1.5) - relatively, comparatively, more or less···(Weight: 1.0) - slightly, somewhat, not very, not overly···(Weight: 0.5)

3.2 Feature Selection

Different sentiment levels exhibit distinct semantic and syntactic characteristics. Feature selection is crucial for SVM classification [14], as precision, recall, and system efficiency depend on appropriate feature selection. Through literature review and examination of real micro-blog corpora, we extracted thirteen features across four dimensions: part-of-speech, sentiment, sentence pattern, and semantic features, as detailed in Table 2 .

Table 2 Feature Types and Definitions - F1: Number of verbs in the micro-blog post - F2: Number of adjectives - F3: Number of adverbs - F4: Number of positive sentiment words - F5: Number of negative sentiment words - F6: Maximum weight of degree adverbs - F7: Sentiment score of the post - F8: Number of negation words - F9: Number of exclamation marks - F10: Number of question marks - F11: Adverbial modifiers related to sentiment words - F12: Adjectival modifiers related to sentiment words - F13: Nominal subjects related to sentiment words

(1) Part-of-Speech Features

Micro-blog language is characterized by brevity and conciseness. Users often express ideas using single words or phrases without complete structures. Incorporating part-of-speech features helps parse sentence structure and assists sentiment judgment. Following literature [15] 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. Positive words convey optimistic attitudes, while negative words express pessimism. Both serve as classification features.

For five-level classification, sentiment intensity is crucial. In this study, intensity is represented through degree adverb weights preceding sentiment words. For example, in "她长得非常好看" (She looks very beautiful), the positive word "好看" (beautiful) is preceded by "非常" (very) with weight 2, increasing the intensity from 1 to 2. For multiple degree adverbs, the maximum weight is used as the intensity feature. Sentiment score is also included as a feature, with higher scores indicating clearer sentiment orientation, calculated using Equation (4).

$$\text{Score} = \text{rawscore} \times \text{Intense}$$

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

(3) Sentence Pattern Features

Negation words can reverse sentiment orientation. For instance, "今天玩的不开心!" (Not happy today!) would be misclassified as positive without considering the negation, when the true sentiment is negative. Thus, negation words are essential features.

Question and exclamation marks indicate emotional emphasis, with frequency correlating with sentiment intensity. Including these punctuation counts as features aids sentiment discrimination.

(4) Semantic Features

Syntactic parsing analyzes whether word sequences conform to grammatical rules and extracts valid syntactic structures [16]. Dependency parsing reveals internal structure and relationships, comprehensively representing sentiment orientation. We used the Stanford Parser [17] for syntactic analysis. Based on corpus observation and literature [18], we extracted three relationship types:

1. **advmod (adverbial modifier)**: Modifies adverb intensity. For example, "她长得非常漂亮" (She looks very beautiful) yields advmod(漂亮, 非常), indicating "非常" modifies the adjective "漂亮".

2. **amod (adjectival modifier)**: An adjective modifying a noun phrase. For example, "这可真是神回复啊" (This is truly a godly reply) yields amod(回复, 神), showing "神" modifies "回复".
3. **nsubj (nominal subject)**: Modifies nominal subjects. For example, "不一样的抗日神剧, 好看!" (A different anti-Japanese war drama, good!) yields nsubj(好看, 剧), indicating "好看" modifies the nominal subject "剧".

3.3 Sentiment Classification Model

Micro-blog corpora contain noise such as #topics#, URLs, and @user mentions that lack subjective opinions and may affect segmentation and POS tagging. We first filter these elements before processing. The NLPIR2016 tool from the Institute of Computing Technology, Chinese Academy of Sciences [19], performs segmentation and POS tagging on filtered corpora. We selected SVM as the classification model, representing training and test sets using the extracted features. The training set trains the SVM model with parameter optimization, and the trained model classifies the test set. The classification model is illustrated in Figure 1 [Figure 1: see original paper].

Experiments

4.1 Dataset

We used a subset of the COAE2014 micro-blog evaluation corpus, manually annotating 5,000 posts as very positive, positive, neutral, negative, or very negative. Annotation was performed by team members, with distribution shown in Table 3 .

Annotation primarily relied on degree adverb levels and punctuation. Statements containing high-intensity degree adverbs like "非常"(very) express stronger sentiment than those with low-intensity or no adverbs. Similarly, multiple exclamation or question marks indicate stronger emotion than punctuation-free statements.

Example 1: 这个翡翠挺好看! (This jade looks nice!)
Example 2: 这个翡翠真的非常非常好看!!! (This jade looks really, really nice!!!)

Example 2 expresses stronger sentiment than Example 1, receiving a +2 label versus +1 for Example 1.

4.2 Feature Extraction

After annotation, we segmented and POS-tagged the corpus, extracting features as described in Section 3.2.

Programs were written in Java on the Eclipse platform under Windows 7 64-bit with 4GB RAM. Table 4 shows partial feature extraction results.

4.3 Model and Evaluation Metrics

We used LibSVM for SVM training and classification [20], splitting each sentiment category 4:1 into training and test sets. Features were normalized before training to improve speed. Default LibSVM parameters were adopted: C_{SVC} SVM type with RBF kernel. Evaluation metrics included precision, recall, and F1-value.

Results and Analysis

5.1 Impact of Feature Combinations

We evaluated different feature combinations using precision as the metric, with results shown in Table 5 .

Table 5 Results for Different Feature Combinations - Part-of-speech only: 57.60% - + Sentiment words: 80.93% - + Degree adverb weights: 81.76% - + Sentiment scores: 81.95% - + Negation words + punctuation: 82.14% - + Semantic features: 82.22% - All features: **82.40%**

Using all features yields the highest precision (82.40%). Sentiment words contribute most significantly, improving precision by 23.33%. Degree adverb weights provide a 0.83% improvement, while other features offer slight enhancements.

5.2 Comparative Evaluation

We compared our method with the cascaded CRFs approach from literature [10]. Cascaded Conditional Random Fields (CCRFs) stack multiple CRF layers, converting five-level classification into a coarse-to-fine process where lower layers provide preliminary results for higher-layer decision support. Literature [10] used cascaded CRFs for three-level classification first, then combined POS, evaluation word, conjunction, and polarity features for five-level classification, achieving 83.75% on COAE2008 tasks. Applying this method to our corpus yielded the comparison results in Table 6 .

Table 6 Comparative Experimental Results | Method | Precision | Recall | F1-value | |—–|———-|—–|———-| | Our method | 82.40% | 81.91% | 82.10% | | Cascaded CRFs | 75.31% | 73.30% | 74.30% |

Our method achieves 82.40% precision, significantly higher than cascaded CRFs (75.31%). Recall is also substantially improved (81.91% vs. 73.30%). The F1-value of 82.10% represents a 7.80% improvement over cascaded CRFs (74.30%). Literature [10]' s features target medium-to-long texts and are less suitable for short micro-blog posts, reducing accuracy. Our Word2Vec-expanded sentiment dictionary and multi-dimensional feature selection across POS, sentiment, sentence pattern, and semantic dimensions achieve higher accuracy for five-level micro-blog sentiment classification.

Conclusion

This paper proposes an SVM-based method for five-level sentiment classification of micro-blog posts, leveraging POS, sentiment, sentence pattern, and semantic features. Compared with existing five-level classification methods, our approach demonstrates superior performance.

The primary limitation is the relatively small training corpus, particularly for very positive and very negative categories. Future work will expand the training corpus to further improve model accuracy.

References

[1] Wang X, Wang Y. 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 C, Yang X, Zhang J. Micro-blog Sentiment Analysis of Combined Sentiment Dictionary and Rules [J]. *Computer Knowledge and Technology*, 2014, 10(13): 3111-3113.

[5] Zhang Y, Liu X, Sun K, 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] Barbosa 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 M, Du Y, Zuo B. 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 J, Wu X. Research on Multi-level Sentiment Analysis System of E-Commerce Product Review and Implementation [J]. *Software*, 2013, 34(9): 65-67, 94.

[12] Liao J, Wang S, Li D, 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 International Conference on Computational Intelligence and Software Engineering*. IEEE, 2010.

[15] Wu M, Chen T. Sentences Tendency Judgement by POS and Dependency Based on SVM [J]. *Journal of Wuyi University: Natural Science Edition*, 2012, 26(4): 66-71.

[16] Liu H. *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 Y. 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 concept, study design, final manuscript revision.

Conflict of Interest

All authors declare no conflict of interest.

Support Data

Supporting data [1-3] are available at the journal's website: http://www.infotech.ac.cn. Supporting data [4] are stored by the authors and available via E-mail: doubleyou1001@163.com.

[1] Yang S, Chen F. senti_{dic}.rar. Positive and negative sentiment dictionaries expanded using Word2Vec.

[2] Yang S, Chen F. weight_{dic}.txt. Degree adverbs with manually assigned weights.
[3] Yang S, Chen F. TestData.rar. Manually annotated test dataset.
[4] Yang S. 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.