Constructing a Sensitive Lexicon for Online Rumors: A Case Study of Sina Weibo Rumors (Postprint)

Authors: Xia Song, Lin Rongrong, Liu Kan

Date: 2023-10-08T00:00:00+00:00

Abstract

[Purpose/Significance] Online rumors severely impact the dissemination of normal information on the internet, making the identification of online rumors critically important in practice. This study constructs a Weibo-based sensitive word lexicon for online rumors to improve the detection accuracy of rumor identification. [Method/Process] Addressing the characteristics of short texts on social media platforms like Weibo, this study first abandons traditional word segmentation algorithms and designs an LBCP word extraction algorithm, which combines positional information and improved TF-IDF weighting to extract seed word sets for the sensitive lexicon. Subsequently, clustering algorithms are employed to supplement near-synonyms of seed words into the lexicon, followed by the addition of commonly used alternative words, thereby obtaining the final sensitive word lexicon. [Results/Conclusion] By utilizing sensitive word features for rumor judgment, and building upon the extraction of Weibo's content features, user features, propagation features, and sentiment analysis features, the addition of sensitive word features leads to a significant improvement in rumor detection rate, achieving favorable identification performance.

Full Text

Construction of a Sensitive Lexicon for Online Rumors: A Case Study of Sina Weibo Rumors

Xia Song, Lin Rongrong, Liu Kan
School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430074

Abstract

[**Purpose/Significance**] Online rumors severely disrupt the dissemination of legitimate information on the internet, making rumor identification a task of critical practical importance. This study constructs a rumor-sensitive lexicon specifically tailored for microblogging platforms to enhance the accuracy of online rumor detection. [**Method/Process**] Addressing the characteristics of short-text content on social media platforms, we first abandon traditional word segmentation algorithms and design the LBCP (Location-Based Cohesion and Polymerization) word extraction algorithm. This algorithm incorporates positional information and an improved TF-IDF weighting scheme to extract a seed word set for the sensitive lexicon. We then supplement the seed words with their synonyms through clustering algorithms and further incorporate common alternative expressions, ultimately yielding a comprehensive sensitive lexicon. [**Result/Conclusion**] Experimental results demonstrate that incorporating sensitive word features significantly improves rumor identification performance. When added to content features, user features, propagation features, and sentiment analysis features, the sensitive word features produce markedly enhanced detection rates and achieve favorable identification outcomes.

Keywords: Sensitive Lexicon; Word Embedding; Feature Space; Online Rumors

---

1. Introduction

In-depth analysis of online rumors facilitates timely discrimination between authentic and false information, thereby fostering a healthy online environment. Current rumor identification approaches predominantly focus on user characteristics and propagation patterns. However, rumor-sensitive words constitute a crucial feature for rumor detection, and analyzing these terms can significantly improve discrimination accuracy and curb rumor proliferation. A lexicon represents a collection of vocabulary, typically comprising both general and specialized word repositories. Widely used specialized lexicons include trending word databases, professional ontological lexicons, sensitive word libraries, and sentiment lexicons. Existing sensitive lexicons primarily target reactionary content, violent terrorism, pornography, and spam advertisements, finding broad application in forums, bulletin boards, and spam email detection. Nevertheless, no comprehensive lexicon currently exists specifically for online rumors.

The proposed rumor-sensitive lexicon is designed for microblogging and social networking platforms, dedicated exclusively to rumor identification. It encompasses three categories of content: (1) Fabricated events, such as false reports of earthquakes or riots; (2) Exaggerated facts, including excessive or false product promotion by manufacturers and defamation of competitors; and (3) Repurposed outdated information and fraud, such as reposting missing child alerts with altered times, locations, or phone numbers to induce calls to fraudulent

numbers. These rumors can trigger widespread public concern and even become focal points of societal attention within certain periods. If not addressed promptly, their potential security threats are incalculable. Traditional lexicons inadequately identify the sensitive words involved in such rumors. Consequently, the lexicon constructed in this study, based on microblog rumors, offers substantial practical value and provides both speed and quality assurance for rumor identification on social platforms.

2. Related Work

Sensitive lexicon construction primarily involves sensitive information identification, extraction, and expansion. Current sensitive information extraction methods predominantly rely on manual annotation and selection or traditional weighting calculations [1] to evaluate and select terms, followed by iterative identification of sensitive information based on reference word forests and subsequent expansion through relevant algorithms [2]. For instance, Liu Geng et al. [3] employed a generalized Jaccard coefficient method to calculate associated terms for sensitive words.

Numerous studies on sensitive events and trending topics have achieved promising results by focusing on sensitive lexicons and hot word sets. Lexicon construction resembles keyword extraction, typically building upon existing professional vocabulary and employing feature weighting methods. Xu Linhong [4] established a sentiment ontology by first determining the classification system based on the current state of sentiment classification and then integrating various sentiment vocabulary resources through a combination of manual classification and automatic acquisition. Hou Li [5] utilized N-Gram and various filtering rules for term recognition in public log data, effectively identifying health-related word sets. C. Quan et al. [6] constructed an emotion lexicon by identifying emotional seed words based on emotion category symbols, intensity levels, sentiment words, degree words, negation words, conjunctions, and rhetorical devices. F. Peng et al. [7] employed linear-chain Conditional Random Fields (CRFs) for Chinese word segmentation and new word detection based on character, word, and multi-word domain integration, using a probabilistic new word detection method. K. J. Chen et al. [9] implemented an online new word extraction system that identifies unknown words primarily through statistical information and grammatical-semantic context. Peng Yun et al. [10] extracted product sentiment words by embedding semantic relationships obtained from word sense comprehension and syntactic analysis into a topic model, proposing the SRC-LDA model for topic word extraction.

Constructing a comprehensive lexicon requires more than establishing a basic word set; expansion is necessary to achieve completeness. Vocabulary expansion resembles keyword expansion, proceeding through semantic or meaning-based approximation. H. Chen et al. [11] extracted words with similar semantic in-

formation from dictionaries for expansion. S. Yu et al. [12] utilized the VIPS (VIsion-based Page Segmentation) algorithm for query expansion, which obtains DOM structure and visual information (derived from HTML elements and attributes) by calling an analyzer embedded in web browsers. J. M. Ponte and W. B. Croft [13] proposed integrating statistical language models with information retrieval, ranking word information using word frequency and document frequency. T. Pedersen and A. Kulkarni [14] identified similar words through clustering for semantic expansion. P. D. Turney and M. L. Littman [15] identified lexical semantic orientation by calculating similarity between target words and benchmark words. A. Neviarouskaya et al. [16] expanded sentiment lexicons through synonym and antonym relationships, contextual semantic relations, derivational relations, and compounding with known lexical units.

However, the aforementioned sensitive lexicon construction methods are not entirely suitable for online rumor corpus development. First, no reference word forest currently exists for rumors. Moreover, rumor dissemination exhibits diverse transformation forms and propagation patterns. Certain words appear frequently in both rumor and normal microblogs, making them unsuitable as sole indicators for rumor determination. Given these characteristics of rumor-sensitive words, we designed a word extraction algorithm that extracts sensitive words and performs multi-level expansion, aiming to establish a practical online rumor-sensitive lexicon.

---

3. Methodology

3.1 Challenges in Rumor-Sensitive Lexicon Construction Constructing a microblog rumor-sensitive lexicon—a highly specialized and domain-biased resource—requires extensive microblog rumor corpora and faces several challenges during development:

(1) **Artificial Interference.** Rumor publishers frequently employ various methods to evade keyword matching and filtering, such as inserting meaningless numbers and symbols between sensitive compound words (e.g., "抵@制!共&$产&0党" for "boycott the Communist Party"). These complex and variable forms do not impede normal human reading but cannot be resolved through direct sensitive lexicon matching.

(2) **Accuracy.** Sensitive words appearing in rumor microblogs often also appear in normal microblogs, causing significant bias in text sensitivity scoring. Most words only exhibit rumor characteristics within specific contexts.

(3) **Segmentation Issues.** Increasingly informal internet language, emerging neologisms, and the timeliness of rumors render traditional segmentation tools unsuitable for such texts, thereby affecting rumor identification.

For artificially interfered rumor texts containing symbols within sensitive words,

we address detection through stop-word expansion and preprocessing methods that remove stop words after segmentation. For accuracy issues, we introduce position weighting and sensitivity weighting to extract sensitive words, using the word frequency ratio between rumor and normal microblogs and position weighting (whether the word appears in the title) as evaluation factors, while expanding the seed word set with similar and associated words. For segmentation problems, we propose the L-CPBL word extraction algorithm based on sensitivity heat, which abandons traditional segmentation tools and extracts text fragments based on internal and external cohesion degrees, making it more suitable for online social text.

3.2 Overall Design The fundamental approach for constructing the online rumor-sensitive lexicon in this study involves: first collecting online rumor corpora, then utilizing a word extraction algorithm to construct a seed word set, and subsequently expanding the seed word set to obtain a comprehensive rumor-sensitive lexicon. The overall process is illustrated in Figure 1 [Figure 1: see original paper].

Since current segmentation software is predominantly general-purpose and ineffective for discovering domain-specific words, sensitive words, and neologisms in rumor texts, seed word collection does not employ direct segmentation. Instead, we designed the LBCP algorithm for word extraction, calculating internal and external cohesion degrees combined with word weighting and position weighting to obtain the seed word set. We then expand the seed word set from three aspects—approximate words, associated words, and substitute words—ultimately merging them into a rumor-sensitive lexicon.

3.3 LBCP Word Extraction Algorithm The LBCP (Location-Based Cohesion and Polymerization) word extraction algorithm considers word position and contextual information. The process for extracting rumor seed words is shown in Figure 2 [Figure 2: see original paper]. The algorithm first sets a sliding window to extract candidate words, calculating their internal cohesion (representing word aggregation degree) and external cohesion (describing word association with context). It then considers word position in the text (weight of 2 for titles, 1 for body text) and uses improved TF-IDF weighting to calculate comprehensive scores for candidate words, extracting the top-ranked candidates as the seed word set.

Microblog rumor texts are short and informal, containing numerous neologisms and trending vocabulary. Relying on traditional lexicons introduces significant errors. Our method effectively circumvents the problem of excessive dependence on lexicons inherent in traditional segmentation methods. The improved TF-IDF weighting calculation better reflects the information integrity and domain relevance of rumor vocabulary, while position weighting appropriately captures the importance of word location (whether in titles).

3.3.1 Internal Cohesion Internal cohesion primarily serves for word segmentation, typically using a fixed-length sliding window to find segmentation points. For example, in Figure 3 [Figure 3: see original paper], with a window length of 6 (i.e., each word not exceeding 5 characters), the sliding window contains "央视已经报道" (six characters). Calculations show that "央视" has the highest internal cohesion, so "央视" is extracted. The window then slides right to extract "已经报道此事" (six characters) to obtain "已经", and this process continues sequentially.

Specifically, internal cohesion calculation first divides the text at each character within the sliding window into left and right parts, computing the product of their occurrence probabilities in the corpus (i.e., p(left) ×p(right)). The maximum product determines the segmentation point. For "央视已经报道", we sequentially calculate: p(央) ×p(视已经报道), p(央视) ×p(已经报道), p(央视已) ×p(经报道), …, p(央视已经报) ×p(道). The product p(央视) ×p(已经报道) is maximal, so "央视" is extracted, and this product serves as its internal cohesion.

Let text X in the sliding window consist of n Chinese characters C_{1}C_{2}…C (Figure 3 [Figure 3: see original paper]). The internal cohesion h(x) calculation first identifies the candidate word boundary position i using formula (1) to segment candidate word x = C_{1}C_{2}…C .

$$\arg\max\{(p(c_1)\times p(c_2...c_n)),...,(p(c_1...c_i)\times p(c_{i+1}...c_n)),...,(p(c_1...c_{n-1})\times p(c_n))\} \quad (1)$$

Formula (2) then records word x and its internal cohesion p (x) within the window. Finally, formula (3) calculates the sum of internal cohesion across all windows as the final internal cohesion h(x), where k represents the frequency of word x in the entire document.

$$p_i(x) = p(c_1...c_i) \times p(c_{i+1}...c_n) \quad (2)$$

$$h(x) = \sum_{i=1}^{k} p_i(x) \quad (3)$$

3.3.2 External Cohesion Considering only internal cohesion may incorrectly treat combinations like "的 …" as independent words. Therefore, we also consider contextual relationships using external cohesion. If a word can be considered independent, it should co-occur with various words in different linguistic contexts, possessing rich "left sets" and "right sets". External cohesion is measured using left and right information entropy. Assuming word x forms phrases x x with left neighbors and xx with right neighbors, the external cohesion g(x) is calculated as:

$$g(x) = \min\{-\sum p(x_{lx})\log(p(x_{lx})), -\sum p(xx_r)\log(p(xx_r))\} \quad (4)$$

Since entropy represents uncertainty, higher entropy indicates greater uncertainty and richer collocations. By calculating left and right information entropy for a text fragment, if the entropy is relatively high, the fragment can be treated as an independent, high-frequency rumor word.

3.3.3 Improved TF-IDF Weighting After obtaining numerous new and existing words from rumor corpora through internal and external cohesion, we can filter them using weighting. Our weighting builds upon TF-IDF with two improvements: (1) Since old news is often reposted or the same rumor is merely modified with different names, locations, or phone numbers, the requirement for document frequency is high while inverse document frequency is less critical. We assign different weights to TF and IDF in the formula (adding in formula (5) to make TF weight greater than IDF weight). (2) We eliminate the influence of varying document lengths on word weighting (adding denominator in formula (5) for cosine normalization) while using logarithmic term frequency to mitigate the impact of frequency magnitude differences. The improved TF-IDF weight for word x is:

$$w_1(x) = \frac{\lambda \log(tf + 1.0) \times \log(idf + 1.0)}{\sqrt{\sum_{i=1}^{n} (\lambda \log(tf_i + 1.0) \times \log(idf_i + 1.0))^2}} \quad (5)$$

where N is the total number of microblogs, is the TF weight, tf and idf represent the TF and IDF values of word x, and the denominator performs cosine normalization using TF values tf and IDF values idf of all words appearing in rumor microblogs.

3.3.4 Position Weighting For microblog rumors, words in titles or topics are more representative than content words, better reflecting the microblog's theme and thus its rumor sensitivity. The individual position weight value for word x in microblog text is defined as:

$$L_i = \begin{cases} 2 & \text{if word x is in title or topic} \\ 1 & \text{otherwise} \end{cases} \quad (6)$$

Scanning microblog content, if " 【】 " or "[]" or "#" is detected, the enclosed text is treated as the title or topic with individual position weight value of 2; otherwise, it is 1. The position weight for word x is calculated as:

$$w_2(x) = \frac{\sum_{i=1}^{D} L_i}{D} \quad (7)$$

where w_{2}(x) is the position weight of the vocabulary, L represents the position weight value of word x in microblog i, and D is the total number of microblogs containing word x.

3.3.5 Word Extraction Algorithm Process The LBCP word extraction algorithm proceeds as follows:

Step 1: Use formulas (1), (2), and (3) for segmentation (word length not exceeding threshold t) and calculate internal cohesion h(x) for each segmented word.

Step 2: Use formula (4) to calculate external cohesion g(x) for each segmented word, sum internal and external cohesion, and filter words with sum greater than threshold .

Step 3: For filtered words, calculate improved TF-IDF weight w_{1}(x) and position weight w_{2}(x), then compute comprehensive score using formula (8):

$$\mathrm{LBCP}(x) = w_1(x) \times w_2(x) \quad (8)$$

Step 4: Sort by comprehensive score LBCP(x) and select top M vocabulary as the seed word set.

3.4 Extended Word Set To achieve effective expansion of rumor-sensitive words, we extend the seed word set from three aspects: approximate words, associated words, and substitute words.

3.4.1 Approximate Word Set Word2Vec-generated word embeddings reflect word context and semantic relationships. Approximate word set expansion primarily uses Word2Vec calculations followed by clustering to find similar words to seed words, obtaining context- and semantics-based approximate words. The process is shown in Figure 4 [Figure 4: see original paper].

3.4.2 Associated Word Set A single sensitive seed word may appear in both rumors and normal microblogs. For example, "free" can appear in both fraudulent vendor rumors and legitimate business promotions. However, when "free" co-occurs with "forward", it is highly likely to be a rumor. Therefore, we calculate high-frequency co-occurring words for each seed word—words with strong relevance that help improve rumor identification rates.

We use mutual information to find seed word associations. The construction process for seed word association sets is shown in Figure 5 [Figure 5: see original paper]. However, such word pairs in rumors often have high mutual information but low frequency, contributing little to rumor identification. Therefore, we incorporate frequency information into mutual information calculation:

$$\mathrm{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)} \quad (9)$$

where p(x, y) represents the co-occurrence frequency of words x and y.

3.4.3 Substitute Word Set Rumor publishers often employ various methods to evade sensitive word matching and filtering, such as converting sensitive words to pinyin, English, or abbreviations. Therefore, we also identify substitute word sets for seed words. This processing volume is manageable and completed manually in this study, including: (1) **Pinyin**: Replacing characters with pinyin, e.g., "abduct" →"guai walk"; (2) **English**: English equivalents of seed words; (3) **Abbreviations**: Common abbreviated forms, e.g., "Shenzhou VI" →"Shen 6".

Following these three expansion methods, the seed word set and various extended word sets are merged to construct the online rumor-sensitive lexicon.

---

4. Experiments

4.1 Dataset We crawled 30,034 rumors published on the Sina Weibo Community Management Center, "Rumor Crusher," and various regional rumor-refuting platforms. Simultaneously, we collected over 35,000 normal microblogs from accounts such as China News Service and CCTV News as positive samples. All data underwent preprocessing including noise removal and stop-word elimination. Noise removal primarily deleted microblogs shorter than 5 characters, as these carry minimal information; their deletion improves processing efficiency.

4.2 Seed Word Set Extraction Following the seed word extraction approach in Section 3.2, we treated all text segments not exceeding threshold t (set to 9 in this study) as potential words. Through sample data experiments, we determined thresholds for internal and external cohesion, ultimately extracting 43,363 candidate words not exceeding threshold t. The internal cohesion, external cohesion, and word frequencies in rumor and normal microblogs for candidate words are shown in Figure 6 [Figure 6: see original paper].

Based on these results and incorporating position weighting factors, we sorted by LBCP comprehensive values and selected the top 300 words as the rumor seed word set. Table 1 shows partial seed words extracted by the LBCP algorithm.

Table 1 Rumor Seed Word Set
child, abduct, spread, reward, urgent, serious, informed, girl, help, forward, plea, search, contact, death, explosion, deceased, share, money, injury, black, leukemia, rescue, passed away, pesticide, lost, cancer, steal, rescue, expert, sell, breed, food, cause, truth, died, arrested, infection, must, remind, kill, banned, missing, emergency rescue, confirmed, crime, beaten to death……

4.3 Seed Word Set Expansion (1) Approximate Word Set Expansion. Using Word2Vec, we computed word vectors for the 300 seed words, calculated dimension-wise mean vectors, and applied KNN clustering to identify the 300

words most similar to seed words. The experiment yielded 1,785 words belonging to the same clusters as seed words as extended approximate words, shown partially in Table 2 .

Table 2 Partial Approximate Word Set
friend, surveillance, nearby, extremely, hold, watch out, brother, man, report, sister, contact, dual, care, infant, reach, Hengtian, professor, main, escape, occasion, casualties, casualties, TV, slander, protect, record, woman, devil, health care, hospital, Japan, contain, decisive, brush, toxin, bureau, discover, heaven and earth, blood pressure, this time, remaining fragrance, drink, trigger, plug, truly, orchid, version, sea area, seven, police, relay, only, light, number of people, news, firefighting, Tianjin, sacrifice, Hangzhou……

(2) Associated Word Set and Substitute Word Set Expansion. We calculated mutual information between seed words and other words in the corpus, sorted in descending order to obtain a final associated word set of 175 words. For example, extended words for "abduct" include: found, rumor-mongering, passed away, unlucky, recently, curse, fund, truth, torment, etc.

The seed word substitute word set contains pinyin, English, and abbreviated forms of the 300 seed words, such as: help →Help, bangmang, bm; reward →Remuneration, fee, pay, choujin, cj, etc.

Upon completion of these three expansion methods, the entire rumor-sensitive lexicon contains 300 seed words, 1,785 approximate words, 175 associated words, and 300 substitute words, totaling 2,260 entries.

4.4 Microblog Rumor Identification We additionally crawled 5,000 rumors from Sina Weibo "Rumor Crusher" and regional rumor-refuting platforms published between January and March 2018, along with 5,000 normal microblogs from verified accounts including China News Service, Toutiao, and CCTV News. These 10,000 microblogs served as test data to verify the lexicon's effectiveness.

From the mixed 10,000 microblogs, we extracted traditional features and sensitive word features as input. Traditional features include user information (follower count, following count, registration age, published microblog count, verification status), structural features (repost count, microblog length, presence of "@", hashtags, URLs, emojis, punctuation usage, first-person pronouns), and average word vector sums. Sensitive word features include sensitive word count and total sensitive word scores.

Using these features, we constructed microblog rumor classifiers with Random Forest, SVM, GBRT, CNN, BiLSTM, and TextCNN. Since rumor identification is prioritized, we aimed to maximize recall rate (proportion of actual rumors correctly identified) while maintaining high precision. Experiments employed 10-fold cross-validation. Table 3 compares accuracy and recall before and after adding sensitive lexicon features.

Table 3 Effect of Sensitive Lexicon Features on Rumor Discrimination

Method	Traditional Features	Traditional Features + Sensitive Word Features
	Accuracy	Recall
Random Forest	79.82%	62.98%
SVM	81.44%	65.65%
GBRT	80.38%	66.09%
CNN	82.68%	72.66%
BiLSTM	81.25%	77.12%
TextCNN	85.10%	86.71%

Experiments demonstrate that integrating sensitive word features with traditional features substantially improves both accuracy and recall across all classification methods. Notably, BiLSTM achieves over 95% accuracy with nearly 90% recall. These results confirm that the constructed rumor-sensitive lexicon effectively enhances microblog rumor identification rates.

5. Conclusion

An online rumor-sensitive lexicon constitutes a critical foundation for rumor identification. This study aimed to construct such a lexicon and demonstrate its effectiveness through auxiliary experiments. Leveraging large-scale corpora, we built a rumor-sensitive lexicon through the L-CPBL word extraction algorithm and expansions of similar and extended words. The first step involves seed word set extraction. The L-CPBL algorithm is a fast, dictionary-independent word extraction method that combines improved LTC weighting and position weighting factors for more accurate seed word extraction. Subsequently, we expanded the seed word set based on word vector space optimization and clustering algorithms, yielding a comprehensive lexicon applicable to rumors.

The constructed lexicon is suitable for short-text social media platforms like microblogs, and its construction process does not rely on manual expert identification or selection. It can be updated synchronously based on corpora, saving time and costs while improving efficiency. However, the rumor-sensitive lexicon has timeliness constraints, requiring continuous collection of large-scale rumor corpora that depend on officially published rumor information for annotation, consuming considerable time and resources. Future research could investigate lexicon updates using temporal algorithms or propagation-based approaches to better address timeliness issues.

References

[1] Xu Jianmin, Wang Jinhua, Ma Weiyu. Improved TF-IDF Feature Word Extraction Method Using Ontological Association[J]. Information Science, 2011,

29(2): 279-283.

[2] Zhou Xiao. Research on Internet-Based Sentiment Lexicon Expansion and Optimization[D]. Shenyang: Northeastern University, 2011.

[3] Liu Geng, Fang Yong, Liu Jiayong. Sensitive Word Library Design Based on Associated Words and Expansion Rules[J]. Journal of Sichuan University (Natural Science Edition), 2009, 46(3): 669-673.

[4] Xu Linhong, Lin Hongfei, Pan Yu, et al. Construction of Sentiment Lexicon Ontology[J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(2): 180-185.

[5] Hou Li, Li Jiao, Hou Zhen, et al. Research on New Word Recognition Method Based on Hybrid Strategy in Public Health Domain[J]. Library and Information Service, 2015, 59(23): 115-122.

[6] Quan C, Ren F. Construction of a Blog Emotion Corpus for Chinese Emotional Expression Analysis[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2009: 1446-1454.

[7] Peng F, Feng F, McCallum A. Chinese Segmentation and New Word Detection Using Conditional Random Fields[C]//Proceedings of International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2004: 562-569.

[8] Zhou Qiang. Research on Automatic Construction of Chinese Combinatory Categorial Grammar Lexicon[J]. Journal of Chinese Information Processing, 2016, 30(3): 196-203.

[9] Chen K J, Ma W Y. Unknown Word Extraction for Chinese Documents[C]//Proceedings of International Conference on DBLP. Taipei: Morgan Kaufmann Publishers, 2002: 169-175.

[10] Peng Yun, Wan Changxuan, Jiang Tengjiao, et al. Product Feature and Sentiment Word Extraction Based on Semantic-Constrained LDA[J]. Journal of Software, 2017, 28(3): 676-687.

[11] Chen H, Lynch K, Basu K, et al. Generating, Integrating and Activating Thesauri for Concept-Based Document Retrieval[J]. IEEE Intelligent Systems and Their Applications, 1993, 8(2): 25-34.

[12] Yu S, Cai D, Wen J, et al. Improving Pseudo-Relevance Feedback in Web Information Retrieval Using Web Page Segmentation[C]//Proceedings of the 12th International Conference on World Wide Web. New York: ACM, 2003: 11-18.

[13] Ponte J M, Croft W B. A Language Modeling Approach to Information Retrieval[C]//Proceeding of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1998: 275-281.

[14] Pedersen T, Kulkarni A. Identifying Similar Words and Contexts in Natural Language with Sense Clusters[C]//Proceedings of the 20th National Conference on Artificial Intelligence. Pittsburgh: AAAI Press, 2010: 1694-1695.

[15] Turney P D, Littman M L. Measuring Praise and Criticism: Inference of Semantic Orientation from Association[J]. ACM Transactions on Information Systems, 2003, 21(4): 315-346.

[16] Neviarouskaya A, Prendinger H, Ishizuka M. SentiFul: A Lexicon for Sentiment Analysis[J]. IEEE Transactions on Affective Computing, 2011, 2(1): 22-36.

Author Contributions

Xia Song: Designed the model, completed experiments, revised the paper.
Lin Rongrong: Collected data, conducted experiments, wrote the initial draft.
Liu Kan: Proposed research ideas, designed research framework, revised and finalized the paper.

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