International Journal of Data Warehousing and Mining
Volume 18 • Issue 1

Semi-Supervised Sentiment Classification on E-Commerce Reviews Using Tripartite Graph and Clustering

Xin Lu, School of Electronics and Information Engineering, South China Normal University, China
Donghong Gu, School of Electronics and Information Engineering, South China Normal University, China
Haolan Zhang, Ningbo Institute of Technology, Zhejiang University, China
Zhengxin Song, School of Electronics and Information Engineering, South China Normal University, China
Qianhua Cai, School of Electronics and Information Engineering, South China Normal University, China*
Hongya Zhao, Shenzhen Polytechnic, China
Haiming Wu, School of Electronics and Information Engineering, South China Normal University, China

ABSTRACT

Sentiment classification constitutes an important topic in the field of natural language processing, whose main purpose is to extract the sentiment polarity from unstructured texts. The label propagation algorithm, as a semi-supervised learning method, has been widely used in sentiment classification due to its describing sample relation in a graph-based pattern whereas current graph developing strategies fail to use the global distribution and cannot handle the issues of polysemy and synonymy properly. In this paper, a semi-supervised learning methodology, integrating the tripartite graph and the clustering, is proposed for graph construction. Experiments on e-commerce reviews demonstrate the proposed method outperform baseline methods on the whole, which enables precise sentiment classification with few labeled samples.

KEYWORDS

Clustering, Label Propagation, Semi-Supervised Learning, Sentiment Classification

INTRODUCTION

The past two decades have witnessed the flourishing of electronic commerce (e-commerce) in a variety of fields (Huang et al., 2018). The sizable volume of e-commerce is growing at a rapid, steady pace (Yu et al., 2013). E-commerce provides people with daily opportunities to purchase products and services in online marketplaces (Hajli et al., 2017). Along with these shopping activities, consumer reviews reflect users' experiences and feelings (Zhang & Zhong, 2019). Consumer engagement always delivers specific sentiments; therefore, these reviews facilitate the purchase decision of other customers and benefits business sales. As such, a deep understanding of sentiment information serves

DOI: 10.4018/IJDWM.307904 *Corresponding Author

as the foundation of opinion mining and processing, which aims to outline individuals' true intentions through their words (Bhargava et al., 2016).

In the field of natural language processing, sentiment analysis refers to the identification of language that carries an evaluative or affective attitude (Esuli & Sebastiani, 2005). Opinions are retrieved through unstructured texts. Then, the sentiment is classified into positive, negative, and neutral categories (Fu et al., 2018).

More recently, both supervised and unsupervised machine learning models have been applied to the sentiment analysis tasks. The former results in high costs and time to generate training samples. The latter lacks accuracy and processing reliability (Gao et al., 2013).

Semi-supervised sentiment classification is proven to be a flexible alternative for analyzing efficiency (Chapelle et al., 2006). Semi-supervised learning falls between unsupervised learning and supervised learning, which includes a small amount of labeled data and a large amount of unlabeled data (Li & Ye, 2018). Compared with the reliance on labeled samples of supervised learning and the low accuracy of unsupervised learning, semi-supervised learning uses as little cost as possible to obtain the classification accuracy close to supervised learning. This is acceptable in most practical scenarios.

Among these methods, the label propagation algorithm, as a graph-based semi-supervised learning approach, holds great promise in sentiment classification (Li et al., 2016). In general, the label propagation algorithm is used due to its intuitive, interpretable processing and easy resolve (Yang & Shafiq, 2018). Notably, label propagation is carried out by the graph. Once the graph is built, every instance is mapped into a node. The edge weight between two nodes represents the similarity of the two instances (Krishnakumari & Akshaya, 2019). Thus, the problem is formulated as a form of propagation on a graph where a node's label propagates to neighboring nodes due to their proximity (Zhu et al., 2005). The labeled data act like sources that push labels through an unlabeled label (Xiaojin & Zoubin, 2002). In this way, the development of the label propagating graph is of great significance as it identifies the relation among samples. Before the deployment of a semi-supervised learning model, the graph must be established to reflect prior knowledge of the domain.

In line with the graph-developing principle, traditional strategies like word-document bipartite graph, K-nearest neighbor (KNN) graph, and Exp-weighted are applied to convey the relation within the texts (Rossi et al., 2016). Notwithstanding, the construction of graphs in a label propagation algorithm remains limited, primarily because the colloquial expressions of words in the document usually result in polysemy and synonymy issues. In a polysemy issue, the same sentiment word may express different degrees or completely opposite sentiment tendencies in different contexts. In a synonymy issue, the same sentiment may be expressed by different sentiment words (Potts, 2016). On the other hand, traditional graph-based methods pay more attention to the local distribution of the sample instead of the global information within the dataset (Yao et al., 2019). For this reason, the traditional graph-based methods are taken as a secondary choice unless a specific word with clear information can be recognized.

The objective of this research is to devise a sentiment classification framework for e-commerce reviews. The research focuses on the semi-supervised learning model integrated with other algorithms. A tripartite graph for defining words-documents-phrases is designed to better describe the sentiment information (Zhu et al., 2014). A seeds-based semi-supervised hierarchical clustering algorithm (S3HC algorithm) is developed to figure out the global distribution information. By combing the tripartite graph and clustering algorithm, a high-quality relation graph of samples can be obtained, and the label propagation strategy can be deployed (Lu & Xu, 2019).

The contributions of this article include:

1. The tripartite graph is constructed to compute the similarities among review texts, which can eliminate the impacts of polysemy and synonymy.
2. The S3HC algorithm, via the clustering strategy, aims at mining the hierarchical distribution information at a global level.

3. The integrating of the tripartite graph and S3HC algorithm can deliver a graph that reflects the exact relationship of the texts and further facilitates the sentiment classification tasks.
4. In the case of few labeled samples, the proposed method can still achieve a higher classification accuracy on all three datasets.

This article is organized as follows. The next section introduces the present research of semi-supervised learning and background theories employed in this article. This is followed by a model construction for sentiment analysis, developed by using tripartite graph and clustering. Then, the article explores experimental results with the analysis. Finally, it provides concluding remarks.

PREREQUISITES

Literature Review

Previous research on sentiment classification is dominated by three basic approaches: (1) unsupervised learning methods; (2) supervised learning methods; and (3) semi-supervised learning methods. Unlike the other two, semi-supervised learning is the machine learning paradigm concerned with utilizing little labeled data to build better classifiers and regressors (Goldberg & Zhu, 2010). Research outcomes have shown that the application of semi-supervised learning methods improves the accuracy on low resource sentiment polarities classification (Gupta et al., 2018). Indeed, the semi-supervised learning approaches are already employed in the sentiment analysis of microblogs, Twitter, and online product comments to achieve a decent working performance (Karan et al., 2018; Sun et al., 2016; Yu et al., 2015).

Typically, semi-supervised learning approaches can be further divided into the generative method (Nigam et al., 2000), semi-supervised support vector machine (S3VM, Joachims, 1999), disagreement-based method (Blum & Mitchell, 1998), and graph-based method. Early work assumes that all the data (labeled and unlabeled) is generated by one implicit model. In this way, the generative method must be carried out on the foundation of reliable specialized knowledge and accurate prediction to ensure the model effectiveness (Nigam et al., 2000). Furthermore, S3VM can be regarded as a derivation of SVM in semi-supervised learning, which is able to find the hyperplane for classifying the two categories. The basis of S3VM is the assumption of low-density separation. Hence, the objective function of S3VM has to be the nonconvex function with more than one low-density separation. In addition, the disagreement-based methods take advantage of the disagreements among the classifiers to improve the properties. Notably, this requires several classifiers that can generate significant differences. In turn, for the issues of a few labeled samples, especially the data without multiple views, the model must be dedicatedly designed.

Compared to the previous approaches, the graph-based method provides a better resolving strategy. Until now, graph-based semi-supervised learning methods have played a large role in sentiment classification. Typically, the use of label propagation for the graph-based semi-supervised learning model has been deeply explored in applications in the natural language processing (NLP) area (Liu et al., 2018). The application of domain knowledge, as well as the extrinsic information, plays an important role in the field of sentiment classification. Li et al. (2018) proposed a two-view label propagation algorithm for sentiment classification based on the analysis of the task and corpus. Zhuang et al. (2017) exploited label information in both the graph learning and label propagation stages. They formulated the label information of the samples into any self-representation methods, keeping the same computational cost. Graph-based deep learning methods, or graph neural networks (GNN), are creative for sentiment analysis according to the ongoing research (Kipf & Welling, 2017; Yang et al., 2016). In comparison to the label propagation method, GNN still needs validation sets for model training and selecting, which results in the large cost of sample making tasks.

Label Propagation Algorithm

Label propagation intends to propagate label information from labeled samples to nearby samples through weighted edges until a global stable stage is obtained (Xiaojin & Zoubin, 2002). Suppose we have n sample points $X = \left\{x_1, x_2, \ldots x_n\right\}$ in which the first l samples are labeled by $\left\{y_1, y_2, \ldots y_l\right\}$ and the remaining samples are unlabeled. The purpose is to predict the labels of the unlabeled samples. We concentrate on an undirected graph:

$$T_{ij} = P\left(i \rightarrow j\right) = \frac{W_{ij}}{\sum_{k=1}^{n} W_{ik}}, W \in \mathbb{R}^{n \times n} \tag{1}$$

and define T as the probabilistic transition matrix. W_{ij} is the edge strength between x_i and x_j.

At this stage, the authors prefer to introduce a matrix $Y_L \in R^{l \times c}$, representing the labels attached to samples. Correspondingly, the probability distribution matrix, delivered as $F \in R^{n \times c}$, consists of F_L and F_U, where F_L is initialized as Y_L and F_U is randomly initialized. The propagating step is carried out until the matrix F reaches a converge. Due to the label propagation definition, we have $F^{t+1} = \overline{TF^t}$. $\bar{T}$ represents the row-normalized matrix of T. Distinctively, each F_L^t tL is reset from Y_L, which is consistent with labeled samples. In this way, we can model the geometric relationships of all samples in the form of a graph.

Formally, this algorithm predicts each sample $y_m (l < m \leq n)$ with a $\hat{y}_m$ as:

$$\hat{y}_m = \arg max\left(F_{mj}\right) \tag{2}$$

For the label propagation algorithm, the key procedure is the construction of a graph, especially when the training data is scarce. Generally, there are two types of graphs, namely fully connected graphs and sparse graphs, which are widely used to measure the similarities between nodes. As for the sparse graph, the KNN graph, with the weight value of $w_{ij} = 1 / k$ where node i falls into the first k^{th} nearest neighbor of node j, is commonly employed.

On the other hand, the fully connected graph (such as Exp-weighted graph) takes the radial basis function (RBF) to represent the weight, which is $w_{ij} = \exp\left(-\frac{x_i - x_j^{\ 2}}{\sigma^2}\right)$ where $x_i - x_j^{\ 2}$ indicates the distance between the two nodes. In sentiment analysis, we employ the term frequency-inverse document frequency (TF-IDF) principle for instances of vectorization and Euclidean metric to characterize the distances.

Word-Document Bipartite Graph

The word-document bipartite graph was initially developed to capture the relationship in the corpus (Sindhwani & Melville, 2008). Unlike the fully connected graph and sparse graph, the bipartite graph describes the connection between the word and document instead of using feature words as the input vectors. More effectively, more lexical knowledge related to target words can be introduced. Regarding parameter selection, both the fully connected graph and sparse graph contain hyperparameters (i.e., k in the sparse graph and σ in the fully connected graph), which are difficult to determine. It can be derived for the bipartite graph; however, no hyperparameter is involved.

According to Figure 1, the left-side nodes represent words and the right-side nodes represent documents. An edge is added between words if they occur within a document. Each of the edge weights can be set in a way to suit a specific text classification task. In this way, the transition probabilities of document-to-word and word-to-document can be calculated.

There is a connection between the two sides once an edge is constructed. This indicates that the word belongs to the document. Then, we define the transition probability from one document to another:

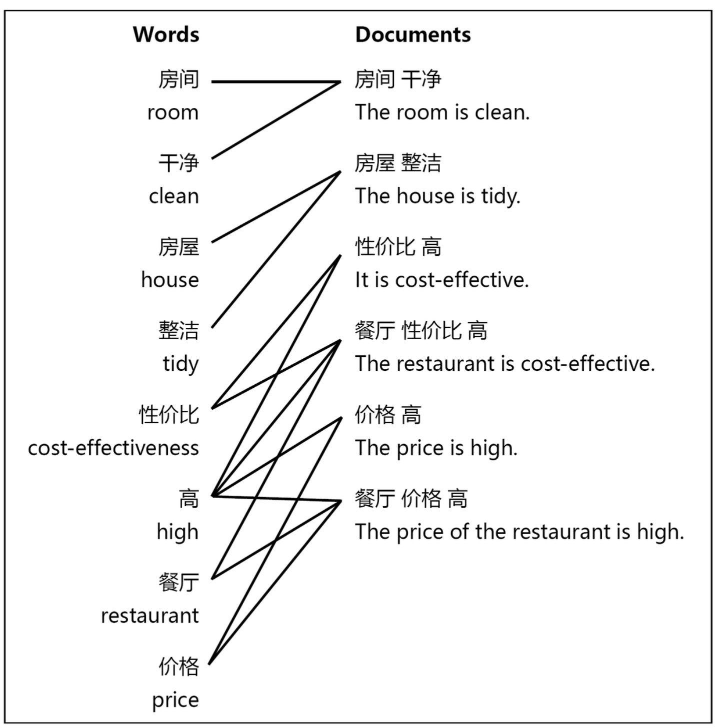
$$T_{ij} = P\left(d_i \to d_j\right) = \sum_{k=1}^{m} \frac{tf_{ik}}{\sum_{k'=1}^{m} tf_{ik'}} \cdot \frac{tf_{jk}}{\sum_{j=1}^{n} tf_{jk}} \tag{3}$$

Together with the document-to-word transition probability and the word-to-document transition probability as:

$$T_{ik} = P\left(d_i \to w_k\right) = \frac{tf_{ik}}{\sum_{k'=1}^{m} tf_{ik'}} \tag{4}$$

$$T_{kj} = P\left(w_k \to d_j\right) = \frac{tf_{jk}}{\sum_{j=1}^{n} tf_{ik}} \tag{5}$$

Figure 1. Example of a bipartite graph representing the word-document relationship



where m is the number of feature words, n is number of documents, and tf_{ik} is the term frequency of the word w_k in the document d_i. In practical use, the transition probability between two documents can be converted to:

$$T = T_{dw} \cdot T_{wd} \tag{6}$$

where T_{dw} stands for the probabilistic transition matrix from all the documents to all the words while T_{wd} stands for that from all the words to all the documents.

METHODOLOGY

This article proposes a semi-supervised learning model to identify the sentiment information from e-commerce reviews. To measure the information with global consistency, the word document-phrase tripartite graph is designed, which originates from the application of the word-document bipartite graph. Meanwhile, the semi-supervised clustering is introduced to optimize the working performance.

Tripartite Graph

As pointed out, the consumer review is usually of the colloquial expressing form. For this reason, the synonyms pattern and bigram phrases pattern are developed to construct the word document-phrase tripartite graph.

Synonyms Pattern

As shown in Figure 1, the sentences "房间干净 (The room is clean.)" and "房间整洁 (The house is tidy.)" deliver a similar meaning. Considering that the bipartite graph is based on the word cooccurrence principle, the words within the two sentences are so different that they cannot be identified via the bipartite graph. This establishes the synonyms pattern to address the word synonymy:

1. **Dataset Generating:** A vocabulary list that serves as a lexicon with all words from the corpus built via sample collection.
2. **Word Similarity Calculating:** Computing and recording of the relationship between any word to another in the lexicon.
3. **Synonymy Information Integrating:** Every word in each review is reinterpreted with its synonyms.

Hence, the review can be resolved based on the synonyms pattern. Specifically, word similarity is calculated on the foundation of the word2vec model, which is trained to measure similarity via cosine distance (Mikolov et al., 2013). The word2vec model originated from an unsupervised language model. We employ sentiment information and part of speech (POS) information for distinguishing the words of different sentiment polarity or different POS. Let v_i and v_j be the vectors of words w_i and w_j, respectively. The similarity between the two words are:

$$sim\left(w_i, w_j\right) = \left(\frac{v_i^T \cdot v_j}{v_i \cdot v_j}\right) \cdot Senti\left(w_i, w_j\right) \cdot POS\left(w_i, w_j\right) \tag{7}$$

where $Senti\left(w_i,w_j\right)$ and $POS\left(w_i,w_j\right)$ are indicator functions for sentiment information and POS information (Dong & Dong, 1999; Hu & Liu, 2004; Ku & Chen, 2007). The researchers use a publicly available sentiment lexicon to obtain the sentiment polarity of words. We concatenate each adjacent word in a sentence to extract the bigram; we remove the bigram that appears in low frequency. Only if w_i and w_j are of the same polarity will we get $Senti\left(w_i,w_j\right)=1$ and otherwise $Senti\left(w_i,w_j\right)=0$. The value of $POS\left(w_i,w_j\right)$ can be obtained in the same way.

Further, a similarity threshold parameter $\beta \in \left(0,1\right)$ is given to characterize the outcome. For each $sim\left(w_i,w_j\right)>\beta$, we identify w_i and w_j as synonyms and vice versa.

Bigram Phrases Pattern

A major property of the consumers' review is the nonstandard expression, which results in the polysemy of words. In line with Figure 1, both "high cost-effectiveness" and "high price" have the same word, "high." Still, the sentiments are opposite. To address this issue, this article proposes the definition of bigram phrases by integrating the neighboring words according to context. In this way, the phrases make more sense than single words in the reviews.

In addition to the synonyms pattern, the bigram phrases can be integrated into the relationship computation. This is given by:

$$T_{ij}=P\left(i\rightarrow j\right)=\sum_{k=1}^{m}\frac{tf_{ik}}{\sum_{k'=1}^{m+b}tf_{ik'}}\cdot\frac{tf_{jk}}{\sum_{j=1}^{n}tf_{jk}}+\sum_{q=1}^{b}\frac{tf_{iq}}{\sum_{q'=1}^{m+b}tf_{iq'}}\cdot\frac{tf_{jq}}{\sum_{j=1}^{n}tf_{jq}} \tag{8}$$

where m is the number of feature words, b is the number of phrases, and n is the number of documents. After integration, the feature word w_k has a frequency of tf_{ik} in the revised document $d_i^{'}$. The phrase p_q has that of tf_{iq}. Appropriately, both the feature words and the bigram phrases are taken into consideration for sentiment classification. As shown in Figure 2, the words in green refer to the synonyms pattern. Those in blue refer to the bigram phrases pattern. This results in a more accurate understanding of the sentiment information.

Semi-Supervised Clustering

Classical similarity computing approaches are developed largely relying on the samples' local distribution information. In this research, however, the global distribution property of samples within the entire dataset must be considered for similarity determination. The devising of semi-supervised clustering based on seed set reveals the deep distribution of cluster in the corpus. This can be employed in line with the tripartite graph.

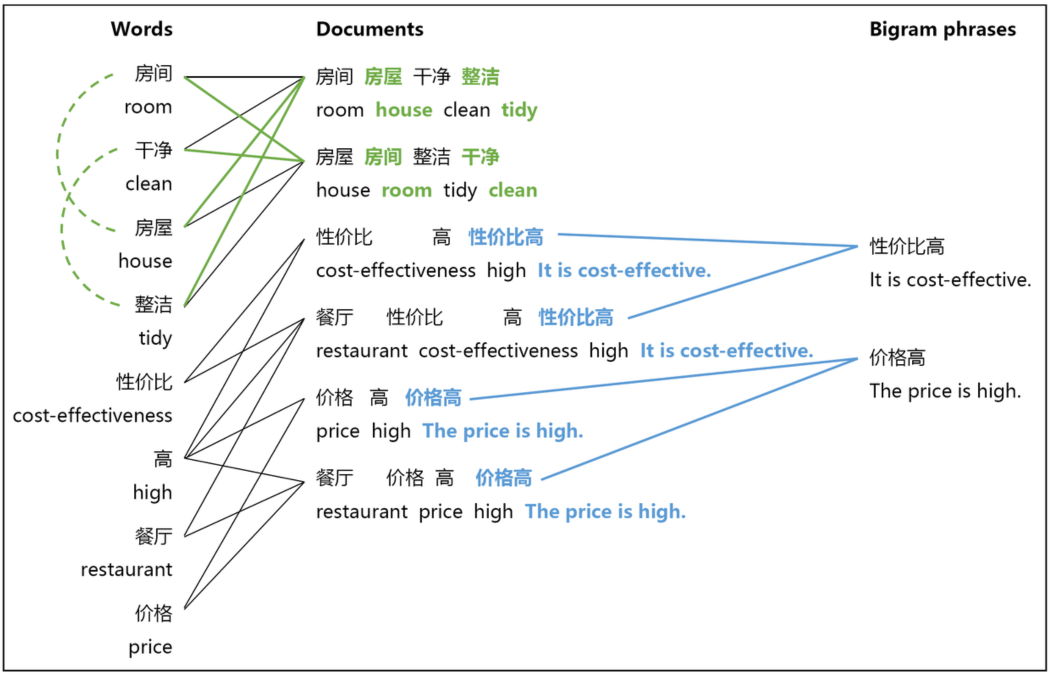
Seeds-Based Semi-Supervised K-Means Algorithm

In semi-supervised clustering, some labeled data is used with the unlabeled data to obtain a better clustering. Basu et al. (2002) proposed two kinds of semi-supervised K-Means clustering algorithms, namely the seeded-K-Means and constrained-K-Means. These guide the clustering process (Basu et al., 2002). For details, see Algorithm 1.

Seeds-Based Semi-Supervised Hierarchical Clustering Algorithm

According to the seeds-based semi-supervised K-Means algorithm, the number of clusters should be set as two for common sentiment binary classification. In practical use, the cluster number cannot be simply selected as two in which way the samples with strong relation may be neglected.

Figure 2. Example of a tripartite graph with synonyms pattern and bigram phrases pattern



Algorithm 1. Seeds-based Semi-supervised K-Means

Input: Set of data points $\chi = \{x_1, x_2, \ldots, x_n\}$, $x_n \in \mathbb{R}^d$, number of clusters K labeled samples $S = \bigcup_{l=1}^{K} S_l$ as initial seeds

Output: Clustering outcomes $C = \{c_1, c, \ldots, c_K\}$

1: **Initialization** of cluster centers $\mu_h^{(0)} \leftarrow \frac{1}{|S_h|} \sum_{x \in S_h} x$, for $h = 1, \ldots, K;\ t \leftarrow 0$

2: **Cluster assigning:** for Seed-KMeans take **2(a)**; for Constrained-KMeans take **2(b)**

2(a): Assign each data x to the cluster h^* (i.e., set $\chi_{h^*}^{(t+1)}$) where $h^* = argmin_h \| x - \mu_h^{(t)} \|^2$

2(b): For $x \in S$, if $x \in S_h$, assign ***x*** to the cluster h (i.e., set $\chi_h^{(t+1)}$); For $x \notin S$, assign x to the cluster h^* (i.e., set $\chi_h^{(t+1)}$) where $h^* = argmin_h \| x - \mu_h^{(t)} \|^2$

3: **Re-calculation** of cluster centers $\mu_h^{(t+1)} \leftarrow \frac{1}{|\chi_h^{(t+1)}|} \sum_{x \in \chi_h} x$ while $t \leftarrow (t+1)$

4: **Repeating the preceding steps until the convergence.**

Traditional K-Means clustering is, however, unable to find a proper cluster number. Note that the divisive hierarchical clustering theory can be applied to cluster number selection. Therefore, this article proposes a seeds-based semi-supervised hierarchical clustering algorithm (S3HC) for mining the distribution of the samples in the corpus.

Hierarchical clustering constructs a tree hierarchy by decomposing data sets. The divisive hierarchical clustering subdivides the dataset into increasingly smaller clusters until a termination condition is reached. Similarly, the S3HC algorithm integrates the seeded-K-Means with the divisive hierarchical clustering, setting the cluster splitting termination condition via labeled samples and deciding the cluster number.

The process of S3HC algorithm can be described as in Algorithm 2. According to Figure 3(a), the dataset contains six documents with two labeled samples (d_1 and d_4) of different categories. In line with the processing of S3HC, a hierarchical clustering tree is established. Each node on this hierarchical clustering tree represents a cluster (i.e., $C_1, C_2, \ldots, C_5$). Figure 3(b) aims to fully exploit

Algorithm 2. Seeds-Based Semi-supervised Hierarchical Clustering

Input: Set of data points $\chi = \left\{x_1, x_2, \ldots, x_n\right\}$, $x_n \in \mathbb{R}^d$, samples vectorization via TF-IDF algorithm, labeled samples $S = \bigcup_{l=1}^{K} S_l$ as initial seeds, K is the number of categories of labeled samples

Output: Clustering outcomes $C = \left\{c_1, c, \ldots, c_K\right\}$. . For each c_r, $r = \left\{1, 2, \ldots, k''\right\}$, k'' as the number of clusters

$C = \varnothing$, $C = C \cup \left\{\chi\right\}$

function SUBDIVISION (χ, K, S)

    if K $\neq 0$ and K $\neq 1$ **then** ▷ The condition of subdivision

        disjoint K partitioning $\left\{\chi\right\}_{l=1}^{K}$ of χ via Seeded-KMeans (χ, K, S)

        for $l = 1 \to K$ **do** ▷ Continue to subdivide the newly obtained K sub-cluster respectively

            $C = C \cup \left\{\chi_l\right\}$

            $S' = \bigcup_{l=1}^{K'} S'_l$ is labeled samples of χ_l

            K' is the number of categories of labeled samples in χ_l

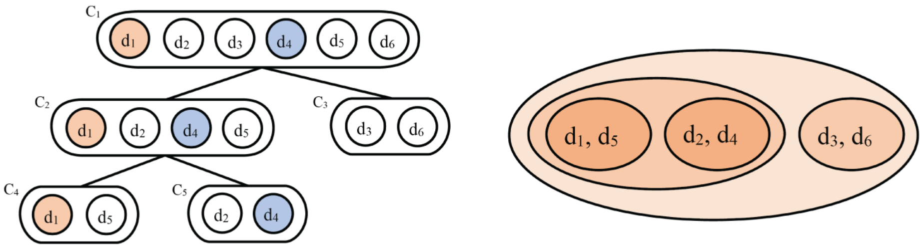
            SUBDIVISION $\left(\chi_l, K', S'\right)$

        end for

    end if

end function

Figure 3. Example of S3HC algorithm



(a) hierarchical clustering tree (b) cluster distribution

the hierarchical cluster structure information. The global distribution matrix is constructed by integrating the clusters from the hierarchical clustering tree.

Integration of Tripartite Graph and S3HC Algorithm

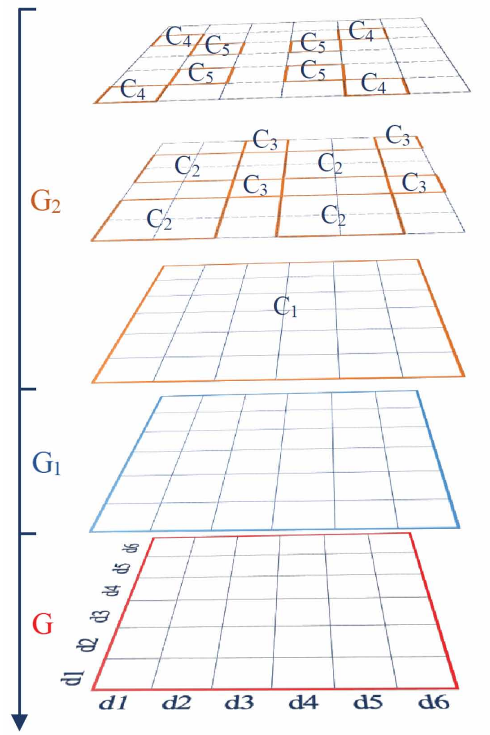
This article proposes a method to detect the sample distributing information using the tripartite graph and S3HC algorithm. As a result, a more accurate sample relation graph can be obtained from the review corpus. The primary steps of the proposed method are as follows:

1. With the application of the semi-supervised clustering algorithm, each cluster is acquired with a uniformly distributed subtransition matrix $T_{C_j} \in \mathbb{R}^{|C_j| \times |C_j|}, j \in \{1,2,\ldots,k''\}$, where $|C_j|$ is the number of samples in cluster C_j:
    a. For seeded-K-Means and constrained-K-Means, the value of each element in T_{C_j} is set as $1/|C_j|$ considering the cluster size. The fewer samples in the cluster, the closer relationship exists (and vice versa).
    b. As for the S3HC algorithm, the element value in the T_{C_j} is given as $1/\left(|C_j| \cdot depth_j\right)$, where $depth_j$ indicates the depth of cluster C_j in the cluster tree. The clusters with more depth tend to contain less samples than those with less depth; therefore, we utilize the depth information to balance the transition probability of samples within different depths.
2. We integrate all the k'' sub-transition matrices into one transition matrix $T_C \in \mathbb{R}^{n \times n}$ with the global distributing information of all samples. According to the working principle of S3HC, one sample can belong to different clusters of different depths. Hence, the specific element in the transition matrix comes from the combination of the corresponding elements from each subtransition matrix.
3. At this stage, two matrices are acquired: the transition matrix T_C from semisupervised clustering and sample relation matrix T from tripartite graph. An optimized matrix T_F for characterizing sample distribution is obtained via weighted fusion. This is given by:

$$T_F = (1-\alpha)T + \alpha T_C \tag{9}$$

As for the example given in Figure 3, the weighted fusion process can be illustrated as Figure 4. In line with the weighted fusion principle, G_2, which corresponds to matrix T_C, represents the sample relation via S3HC algorithm. G_1, which corresponds to matrix T, is the sample transition probability

Figure 4. Weighted fusion of graphs



by using tripartite graph. Characteristically, graph G, which corresponds to matrix T_F, stands for the weighted fusion outcome.

EXPERIMENT

The experimental setup and results are used to evaluate the effectiveness of the proposed sentiment classification framework.

Experimental Setup

Data

The comparison experiments are on both Chinese and English sentiment corpus. For Chinese corpus, we conduct the processing on ChnSentiCorp (Tan & Zhang, 2008) from a hotel domain (ChnSentiCorp-Htl-del-4000) and a laptop domain (ChnSentiCorp-NB-del-4000). Each dataset contains 2,000 positive and 2,000 negative documents for model training and testing, respectively. Correspondingly, for English corpus, the Amazon Electronic product review (Fang et al., 2014) and IMDB movie reviews (Hajli et al., 2017) are employed. The Amazon Electronic product review contains 1,000 positive and 1,000 negative documents for model training and testing. All the samples in the IMDB dataset, 25,000 positive and 20,000 negative documents, also consider the resolution of unbalance issue.

Preprocessing

Original samples are transformed to processable data for further analysis. For Chinese corpus, we employ the ICTCLAS. This is provided by the Chinese Academy of Science for Chinese word segmentation and POS tagging (Maas et al., 2011). The function words without sense and the neutral words without sentiment must be removed from the documents. Consequently, the Chinese POS tagging set of the Institute of Computing Technology (Zhang & Liu, 2002) and the Stop word list of the Harbin Institute of Technology (Liu et al., 2004) are utilized for stop-word removal. For English corpus, we use the Stanford CoreNLP for lemmatization and POS tagging (HIT-SCIR, 2013). Moreover, we remove the punctuation from reviews and leave the top 10,000 feature words with the highest word frequency for higher computational efficiency.

For Chinese word embeddings, the SogouT corpus, which contains over 130 million pages crawled in the Chinese Web, is employed for language model training in word2vec (Manning et al., 2014). We use pretrained word representations for English word embeddings. These are trained on Common Crawl and Wikipedia in advance (Liu et al., 2012). Typically, the issue of synonymy is appropriately addressed by word embedding. Afterward, considering that the domain knowledge will affect the sentiment delivery, we tend to make use of the specific domain knowledge in an e-commerce review for further processing.

Dataset Subdivision

In aiming to classify the unlabeled samples automatically, we subdivide the dataset into labeled sample sets (L) and unlabeled sample sets (U) via the method of randomly dividing. Size *L*, denoted as $\left|L\right|$, is from 10 to 100, with an interval of 10. The unlabeled samples are considered test samples to calculate model accuracy. In addition, each experiment is run 10 times to reduce the impact of randomness. The average outcome is taken as the experimental result.

Evaluation Protocol

The average accuracy is widely adopted to evaluate the classification methods. This is defined as:

$$Average\ accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

where TP , TN , FP , and FN represent the number of true positive samples, true negative samples, incorrectly identified positive samples, and incorrectly identified negative samples, respectively.

On the other hand, two hyperparameters must be determined in our framework. These are the weighted fusion of graphs threshold α and the word-similarity threshold β . The labeled samples are limited; therefore, we take the leave-one-out cross-validation (LOOCV) method to determine the hyperparameters. We take corpus ChnSentiCorp-Htl-del-4000 and $\left|L\right| = 100$ as an example. It can be seen from Figure 5 that when threshold $\alpha < 0.8$, the average result is continuously improved as the global distribution information between samples is gradually strengthened. The highest value occurs when $\alpha = 0.8$. As α increases, the average result decreases. When $\alpha = 1$, the model achieves the worst accuracy rate (77%). In this occasion, the model does not use the tripartite graph information, indicating that tripartite graph can improve model performance by alleviating the synonym and polysemy issue. As can be seen from Figure 6, the best performance is achieved when the word similarity threshold β is set to 0.5. Higher values may lead to the introduction of nonsynonym words. Lower values will lead to an insufficient number of synonyms to capture synonym patterns in documents.

Likewise, two other hyperparameters are worked out in baseline methods in the same manner. The first is the parameter k in KNN graph. The second is the parameter σ in Exp-weighted graph.

Figure 7 is a workflow chart of the experiment procedure. E-commerce consumer reviews are collected and preprocessed into normalized data. Hereafter, both the tripartite graph and the S3HC relation graph can be obtained. Via weighted fusion of the graphs, the sample relation graph is applied

Figure 5. Weighted fusion of graphs threshold α results

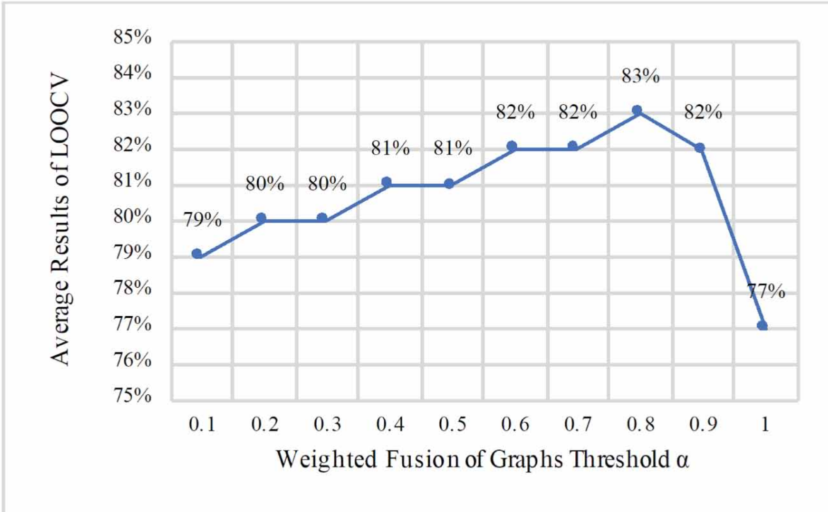
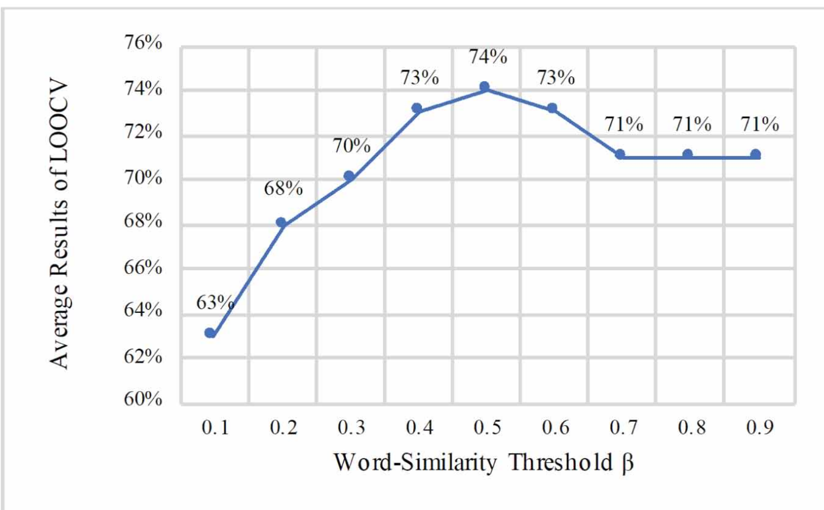


Figure 6. Word-similarity threshold β results

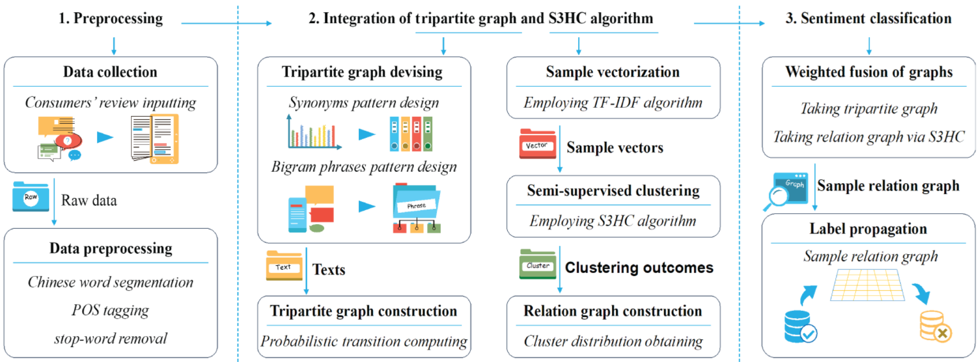


to the label propagation algorithm. This transmits the categories from labeled to unlabeled samples. In this way, the sentiment information can be identified.

RESULTS

The performance of the sentiment classification method proposed in this article is analyzed by comparing it to verified approaches. The KNN graph, Exp-weighted graph, and word-document bipartite graph were traditional methods used for comparison. Deep learning methods use multilayer perceptron (MLP), Planetoid, and graph convolutional networks (GCN, Kipf & Welling, 2017; Yang et al., 2016).

Figure 7. Flow chart of the experiment



Regarding the ablation study, the bipartite graph was taken in combination with the synonyms pattern and bigram phrases pattern. Further processes were taken to verify the capabilities of the tripartite graph and S3HC. Regarding the working performance evaluation, the tripartite graph is combined with the Seeded-KMeans clustering, Constrained-KMeans clustering, and the S3HC algorithm. This was abbreviated as Tripartite+SK, Tripartite+CK, and Tripartite+S3HC, respectively.

The classification accuracy of the methods in the experiment are obtained by using the labeled samples from the aforementioned datasets (see Tables 1-5). The performance of different methods is statistically compared. The classification accuracy of traditional methods (i.e., KNN graph, Exp-weighted graph, and Word-document bipartite graph) raises progressively in line with the increasing number of labeled samples. Comparing deep learning methods in all datasets, these approaches tend to obtain better outcomes than the three deep learning methods.

In the ChnSentiCorp dataset and Amazon Electronic product review, the tripartite graph can effectively remove the polysemy and synonymy; therefore, it has an even higher accuracy in both datasets. Quantitative results of the proposed methods provide evidence that combining the tripartite graph with different clustering algorithms can generate a higher classification accuracy. In this way, sample global information obtained from clustering can benefit the relation description.

Generally, the combination of the tripartite graph and S3HC algorithm presents a best working performance due to its subdividing of text and capturing of the hierarchical distribution. Despite the

Table 1. Accuracy on ChnSentiCorp-Htl-del-4000 (%)

	|*L*|	10	20	30	40	50	60	70	80	90	100
Baselines	KNN graphs	62.82	64.34	65.23	65.47	65.75	66.69	66.62	66.63	667.61	67.79
	Exp-weighted graph	69.22	72.54	74.68	75.79	75.72	76.10	76.32	77.06	76.96	77.34
	Bipartite graph	67.13	69.79	71.77	73.08	73.73	75.09	75.14	75.96	76.66	77.34
This paper	Bipartite+Phrase	66.84	71.00	73.34	74.52	75.60	76.87	77.06	78.06	78.59	79.34
	Bipartite+Synonym	71.94	73.39	75.91	76.70	77.43	77.93	78.13	78.74	79.14	79.29
	Tripartite	71.94	73.37	76.43	77.29	77.98	78.54	78.96	79.62	79.95	80.22
	S3HC	68.00	71.64	73.32	75.37	77.90	77.32	77.61	78.04	78.93	78.49
	Tripartite+SK	74.10	76.38	79.33	80.98	81.41	81.43	81.69	81.91	82.26	82.39
	Tripartite+CK	73.86	75.90	79.34	81.29	81.65	81.72	81.85	82.20	82.36	82.58
	Tripartite+S3HC	**75.37**	**77.97**	**79.97**	**81.67**	**82.13**	**82.23**	**82.23**	**82.39**	**82.81**	**83.53**

Table 2. Accuracy on ChnSentiCorp-NB-del-4000 (%)

	|L|	10	20	30	40	50	60	70	80	90	100
Baselines	KNN graph	65.16	65.78	66.50	67.15	68.20	68.39	68.79	68.77	69.34	69.07
	Exp-weighted graph	69.38	73.16	75.07	76.66	77.53	77.47	77.67	78.28	78.43	78.85
	Bipartite graph	65.60	70.00	71.61	73.53	75.43	74.87	77.53	77.78	78.54	79.15
This paper	Bipartite+Phrase	67.27	72.81	73.46	75.31	77.02	76.82	78.82	79.29	79.22	80.07
	Bipartite+Synonym	72.91	76.52	76.89	77.76	78.12	78.12	78.34	78.34	78.40	78.40
	Tripartite	73.43	77.10	77.41	78.28	78.83	79.06	79.13	79.25	79.52	79.71
	S3HC	73.57	79.33	80.94	81.77	80.66	81.35	81.56	81.92	82.12	82.07
	Tripartite+SK	79.05	82.99	83.73	83.92	83.93	83.86	83.82	83.89	83.97	84.00
	Tripartite+CK	78.87	**83.11**	**83.89**	**84.05**	**84.05**	83.88	84.03	83.99	84.08	84.06
	Tripartite+S3HC	**80.04**	82.38	82.84	83.43	83.60	**84.19**	**84.10**	**84.64**	**84.73**	**84.77**

Table 3. Accuracy on Amazon Electronic product review dataset (%)

	|L|	10	20	30	40	50	60	70	80	90	100
Baselines	KNN graph	50.16	50.32	50.34	50.81	50.63	50.94	50.87	51.50	51.07	51.43
	Exp-weighted graph	71.14	71.25	71.41	71.50	71.34	71.73	71.97	72.37	72.49	72.50
	Bipartite graph	70.33	70.79	71.02	71.21	71.47	71.70	71.61	71.72	71.75	72.01
This paper	Bipartite+Phrase	70.52	70.92	71.39	71.43	71.55	71.77	71.82	72.17	71.97	72.32
	Bipartite+Synonym	72.19	72.71	72.67	72.75	72.86	72.90	73.05	73.33	73.31	74.05
	Tripartite	73.29	73.33	73.77	73.79	74.12	74.33	74.63	75.02	75.82	75.93
	S3HC	75.12	75.37	75.65	75.79	76.02	76.21	76.83	77.07	77.10	77.67
	Tripartite+SK	76.78	77.17	78.39	79.62	79.73	78.12	78.67	78.94	78.89	79.49
	Tripartite+CK	76.64	77.23	78.47	**79.60**	**79.84**	78.02	78.69	78.87	78.93	79.67
	Tripartite+S3HC	**79.01**	**79.14**	**79.20**	79.36	79.39	**79.57**	**79.87**	**80.03**	**80.17**	**80.21**

Table 4. Accuracy on IMDB review dataset (%)

	|L|	10	20	30	40	50	60	70	80	90	100
Baselines	KNN graph	54.16	54.34	54.73	54.96	55.04	55.04	54.78	55.06	53.25	53.29
	Exp-weighted graph	65.14	67.43	69.34	69.12	70.87	72.09	73.17	74.25	73.95	74.24
	Bipartite graph	62.73	64.28	65.44	65.19	66.41	67.00	67.91	69.23	69.90	69.91
This paper	Bipartite+Phrase	64.51	65.67	67.04	66.39	67.43	68.03	68.90	69.92	70.31	70.31
	Bipartite+Synonym	78.69	79.21	**79.91**	**79.64**	**80.28**	**80.48**	**80.71**	**80.98**	**81.11**	81.05
	Tripartite	78.50	78.97	79.59	79.30	79.84	80.09	80.20	80.71	80.86	80.79
	S3HC	73.57	79.33	80.94	81.77	80.66	81.35	81.56	81.92	82.12	82.07
	Tripartite+SK	76.40	76.57	76.89	75.12	75.92	76.77	76.99	77.61	77.29	78.35
	Tripartite+CK	73.82	73.90	72.91	72.08	72.00	71.68	71.81	71.45	71.16	71.56
	Tripartite+S3HC	**79.36**	**79.57**	79.59	77.46	78.78	79.47	79.80	80.65	80.63	**81.30**

Table 5. Experimental results compared to deep learning algorithms (%)

	|*L*|	10	20	30	40	50	60	70	80	90	100
ChnSentiCorp-Htl-del-4000	MLP	71.50	67.41	63.53	64.55	68.94	70.25	75.57	74.21	75.60	75.00
	Planetoid	65.49	70.13	68.49	70.53	74.71	75.31	76.87	75.69	76.65	77.69
	GCN	**77.19**	**79.37**	78.59	77.37	79.01	79.34	79.21	78.73	79.51	78.64
	Tripartite+S3HC	75.37	77.97	**79.97**	**81.67**	**82.13**	**82.23**	**82.39**	**82.81**	**83.53**	**83.21**
ChnSentiCorp-NB-del-4000	MLP	57.34	71.58	70.91	73.21	74.03	74.26	75.45	75.82	75.50	76.41
	Planetoid	60.13	71.56	72.09	73.97	75.70	76.73	78.14	78.27	77.57	77.77
	GCN	77.89	77.61	80.15	80.00	80.25	77.46	78.63	80.41	77.57	78.80
	Tripartite+CK	78.87	**83.11**	**83.89**	**84.05**	**84.05**	83.88	84.03	83.99	84.08	84.06
	Tripartite+S3HC	**80.04**	82.38	82.84	83.43	83.60	**84.19**	**84.10**	**84.64**	**84.73**	**84.77**
Amazon Electronic product review	MLP	68.36	68.75	67.20	67.74	67.15	67.79	68.44	69.27	68.59	68.30
	Planetoid	67.63	66.23	67.11	67.20	68.05	69.19	70.37	71.26	71.07	70.19
	GCN	66.78	66.90	67.90	68.11	69.88	69.43	69.45	69.07	71.40	72.71
	Tripartite+S3HC	**79.01**	**79.14**	**79.20**	**79.36**	**79.39**	**79.57**	**79.87**	**80.03**	**80.17**	**80.21**
IMDB	MLP	60.27	58.53	56.21	56.74	57.15	57.69	58.40	59.08	58.61	58.04
	Planetoid	57.63	56.23	57.11	57.20	58.05	59.19	60.37	61.26	61.07	60.41
	GCN	56.90	56.93	57.68	58.82	59.14	59.14	59.07	59.07	61.24	62.07
	Bipartite+Synonym	78.69	79.21	**79.91**	**79.64**	**80.28**	**80.48**	**80.71**	**80.98**	**81.11**	81.05
	Tripartite+S3HC	**79.36**	**79.57**	79.59	77.46	78.78	79.47	79.80	80.65	80.63	**81.30**

sophisticated method and inputting samples, the algorithm cannot always improve results. Notably, when the numbers of the labeled sample are 20, 30, 40, and 50 in dataset ChnSentiCorp-NB-del-4000, the classification of tripartite+S3HC drops slightly. This suggests that the noise within the raw data can be amplified and transmitted during the cluster splitting.

In the IMDB dataset, the application of synonyms pattern sees great progress in classification accuracy. This is because documents generate more connections and the number of connected components decrease through the graph. Basically, large amounts of samples of colloquial expressing forms will lead to too many connected components. As the labeled samples are selected randomly, most of the connected components may not have labeled samples. Therefore, the samples in the connected components will be classified incorrectly. For this reason, the improvement caused by clustering is slight. Correspondingly, there is a significant gap between the deep learning models and synonyms pattern-based methods.

CONCLUSION

To better classify the sentiment polarity of the e-commerce review, an approach using tripartite graph and seeds-based semi-supervised hierarchical clustering is designed and presented. Experimental results reveal that the working performance can be evaluated by applying the proposed methods to consumer reviews. The methodology presented in this article improves the classification accuracy. It also as addresses the optimal description of the sentiment information.

Regarding future work, semi-supervised sentiment classification of complex conditions may impact accuracy and robustness. The current method is insufficient due to difficulties in obtaining high-quality labeled samples. In future work, the active learning method can be used to achieve a delicate selection of high-quality labeled samples.

The current method uses a label propagation algorithm (LPA) as the learning method of the sample category. A drawback of LPA is the instability of classification results. In addition, the results of multiple times propagation may vary. The effects of deep learning methods like graph neural networks in graph structure modeling continues to improve due to the development of deep learning. Future work can combine the deep learning method with end-to-end training to improve stability.

ACKNOWLEDGMENT

This work was supported by the National Statistical Science Research Project of China under Grant No. 2016LY98, the Science and Technology Department of Guangdong Province in China under Grant Nos. 2016A010101020, 2016A010101021 and 2016A010101022, the Characteristic Innovation Projects of Guangdong Colleges and Universities (No. 2018KTSCX049), the Science and Technology Plan Project of Guangzhou under Grant Nos. 202102080258 and 201903010013, the Humanity and Social Science Foundation of the Ministry of Education of China (21A13022003), the Zhejiang Provincial Natural Science Fund (LY19F030010), the Zhejiang Provincial Social Science Fund (20NDJC216YB), the Zhejiang Educational Science Fund (GH2021642).

REFERENCES

Basu, S., Banerjee, A., & Mooney, R. (2002). Semi-supervised clustering by seeding. *Proceedings of 19th International Conference on Machine Learning*.

Bhargava, K., Gujral, T., Chawla, M., & Gujral, T. (2016, March). Comment based seller trust model for e-commerce. In *2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT)* (pp. 387-391). IEEE. doi:10.1109/ICCTICT.2016.7514612

Blum, A., & Mitchell, T. (1998, July). Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory* (pp. 92-100). doi:10.1145/279943.279962

Chapelle, O., Scholkopf, B., & Zien, A. (2006). *Semi-supervised learning*. MIT Press. doi:10.7551/mitpress/9780262033589.001.0001

Dong, Z., & Dong, Q. (1999). *Hownet: Word sets for sentiment analysis (beta)*. http://www.keenage.com

Esuli, A., & Sebastiani, F. (2005, October). Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management* (pp. 617-624). doi:10.1145/1099554.1099713

Fang, F., Dutta, K., & Datta, A. (2014). Domain adaptation for sentiment classification in light of multiple sources. *INFORMS Journal on Computing*, *26*(3), 586–598. doi:10.1287/ijoc.2013.0585

Fu, X., Yang, J., Li, J., Fang, M., & Wang, H. (2018). Lexicon-enhanced LSTM with attention for general sentiment analysis. *IEEE Access: Practical Innovations, Open Solutions*, *6*, 71884–71891. doi:10.1109/ACCESS.2018.2878425

Gao, W., Wang, Z., & Li, S. (2013). Semi-supervised sentiment classification with an ensemble strategy. *Journal of Chinese Information Processing*, *27*(3), 120–126.

Goldberg, A. B., & Zhu, X. (2010). *New directions in semi-supervised learning* [Doctoral dissertation]. University of Wisconsin-Madison.

Gupta, R., Sahu, S., Espy-Wilson, C., & Narayanan, S. (2018, April). Semi-supervised and transfer learning approaches for low resource sentiment classification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5109-5113). IEEE. doi:10.1109/ICASSP.2018.8461414

Hajli, N., Wang, Y., Tajvidi, M., & Hajli, M. S. (2017). People, technologies, and organizations interactions in a social commerce era. *IEEE Transactions on Engineering Management*, *64*(4), 594–604. doi:10.1109/TEM.2017.2711042

HIT-SCIR. (2013). *Stop word list of hit*. http://ir.hit.edu.cn/

Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177). ACM.

Huang, Y., Chai, Y., Liu, Y., & Shen, J. (2018). Architecture of next-generation e-commerce platform. *Tsinghua Science and Technology*, *24*(1), 18–29. doi:10.26599/TST.2018.9010067

Joachims, T. (1999, June). Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning* (pp. 200-209). Academic Press.

Karan, R. S., Shirsat, K. K., Kasar, P. L., & Chaudhary, R. (2018, March). Sentiment analysis on Twitter data: A new approach. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)* (pp. 1-4). IEEE. doi:10.1109/ICCTCT.2018.8551169

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *The 5th International Conference on Learning Representations (ICLR)*.

Krishnakumari, K., & Akshaya, P. (2019). A survey on graph based approaches in sentiment analysis. *International Research Journal of Engineering and Technology*, *6*(7), 1322–1330.

Ku, L. W., & Chen, H. H. (2007). Mining opinions from the Web: Beyond relevance retrieval. *Journal of the American Society for Information Science and Technology*, *58*(12), 1838–1850. doi:10.1002/asi.20630

Li, S., Xu, J., Zhang, D., & Zhou, G. (2016, December). Two-view label propagation to semi-supervised reader emotion classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2647-2655). Academic Press.

Li, Y., & Ye, J. (2018, July). Learning adversarial networks for semi-supervised text classification via policy gradient. In *Proceedings of the 24th ACM Sigkdd International Conference on Knowledge Discovery & Data Mining* (pp. 1715-1723). doi:10.1145/3219819.3219956

Liu, H., Wang, J., Li, S., Li, J., & Zhou, G. (2018, August). Semi-supervised sentiment classification based on auxiliary task learning. In *CCF International Conference on Natural Language Processing and Chinese Computing* (pp. 372-382). Springer. doi:10.1007/978-3-319-99501-4_33

Liu, Q., Zhang, H., & Zhang, H. (2004). *Chinese POS tagging set of institutes of computing technology*. http://ictclas.nlpir.org/nlpir/html/readme.htm

Liu, Y., Chen, F., Kong, W., Yu, H., Zhang, M., Ma, S., & Ru, L. (2012). Identifying web spam with the wisdom of the crowds. *ACM Transactions on the Web*, *6*(1), 1–30. doi:10.1145/2109205.2109207

Lu, W., & Xu, J. (2019). Sentiment analysis for online user reviews based on tripartite network. *Data Analysis and Knowledge Discovery*, *3*(8), 10–20.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 142-150). Academic Press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55-60). doi:10.3115/v1/P14-5010

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, *39*(2), 103–134. doi:10.1023/A:1007692713085

Potts, C. (2007). The expressive dimension. *Theoretical Linguistics*, *33*(2), 165–198. doi:10.1515/TL.2007.011

Rossi, R. G., de Andrade Lopes, A., & Rezende, S. O. (2016). Optimization and label propagation in bipartite heterogeneous networks to improve transductive classification of texts. *Information Processing & Management*, *52*(2), 217–257. doi:10.1016/j.ipm.2015.07.004

Sindhwani, V., & Melville, P. (2008, December). Document-word co-regularization for semi-supervised sentiment analysis. In *2008 8th IEEE International Conference on Data Mining* (pp. 1025-1030). IEEE. doi:10.1109/ICDM.2008.113

Sun, Q., Niu, J., Yao, Z., & Qiu, D. (2016, December). Research on semantic orientation classification of Chinese online product reviews based on multi-aspect sentiment analysis. In *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies* (pp. 262-267). doi:10.1145/3006299.3006325

Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for Chinese documents. *Expert Systems with Applications*, *34*(4), 2622–2629. doi:10.1016/j.eswa.2007.05.028

Xiaojin, Z., & Zoubin, G. (2002). *Learning from labeled and unlabeled data with label propagation.* Tech. Rep., Technical Report CMU-CALD-02–107, Carnegie Mellon University.

Yu, W., Yan, C., Ding, Z., Jiang, C., & Zhou, M. (2013). Modeling and validating e-commerce business process based on Petri nets. *IEEE Transactions on Systems, Man, and Cybernetics. Systems*, *44*(3), 327–341. doi:10.1109/TSMC.2013.2248358

Yang, Y., & Shafiq, M. O. (2018, August). Large scale and parallel sentiment analysis based on Label Propagation in Twitter Data. In *2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering* (pp. 1791-1798). IEEE. doi:10.1109/TrustCom/BigDataSE.2018.00270

Yang, Z., Cohen, W., & Salakhudinov, R. (2016, June). Revisiting semi-supervised learning with graph embeddings. In *International Conference on Machine Learning* (pp. 40-48). PMLR.

Yao, L., Mao, C., & Luo, Y. (2019). Graph convolutional networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 7370–7377. doi:10.1609/aaai.v33i01.33017370

Yu, Z., Wong, R. K., Chi, C. H., & Chen, F. (2015, December). A semi-supervised learning approach for microblog sentiment classification. In *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)* (pp. 339-344). IEEE. doi:10.1109/SmartCity.2015.94

Zhang, H., & Liu, Q. (2002). *Ictclas*. http://ictclas.nlpir.org/

Zhang, S., & Zhong, H. (2019). Mining users trust from e-commerce reviews based on sentiment similarity analysis. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 13523–13535. doi:10.1109/ACCESS.2019.2893601

Zhu, J., Lafferty, R., & Rosenfeld. (2005). *Semi-supervised learning with graphs* [Doctoral dissertation]. Carnegie Mellon University.

Zhu, L., Galstyan, A., Cheng, J., & Lerman, K. (2014, June). Tripartite graph clustering for dynamic sentiment analysis on social media. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data* (pp. 1531-1542). doi:10.1145/2588555.2593682

Zhuang, L., Zhou, Z., Gao, S., Yin, J., Lin, Z., & Ma, Y. (2017). Label information guided graph construction for semi-supervised learning. *IEEE Transactions on Image Processing*, *26*(9), 4182–4192. doi:10.1109/TIP.2017.2703120 PMID:28541200

Donghong Gu received the B.S. degree from South China Normal University, China, in 2019, where he is currently pursuing the M.S. degree with the School of Physics and Telecommunication Engineering. His research interests include nature language processing, text classification, and sentiment analysis.

Hao Lan Zhang is currently with the Joint Center for SCDM, NIT, Zhejiang University. He is the DP of The Board of Jiangsu New Materials Co., the 151 Talented Scholar of Zhejiang Province and AP at Hebei Academy of Science, China. Prior to that, he was a Research Fellow at RMIT University, Australia. Dr. Zhang received the Ph.D. in Computer Science from Victoria University, Australia. His research interests include Intelligent Information Systems, Multiagent Systems, Health Informatics, Knowledge-based Systems, Data Mining, etc.

Zhengxin Song received the B.S. degree from Wuhan Polytechnic University, China, in 2020. He is currently pursuing the M.S. degree with the School of Physics and Telecommunication Engineering, South China Normal University. His research interests include nature language processing, text classification, and sentiment analysis.

Qianhua Cai received her master's degree in the school of computer science of South China Normal University. After graduation, she joined the School of Electronics and Information Engineering of South China Normal University. Research interests: data mining, text data analysis.

Hongya Zhao obtained her PhD in Hong Baptist University and worked in City University of Hong Kong and MD Anderson Cancer Center of Texas University as the Postdoc. Her research of interest includes statistics, pattern recognize and bioinformatics.