Computing and Informatics, Vol. 41, 2022, 788–812, doi: 10.31577/cai_2022_3_788

DANMAKU TEXT CLUSTERING ALGORITHM BASED ON FEATURE EXTENSION AND WORD-PAIR FILTERING OBTM

Di Wu, Zhuyun Huang

Department of Information and Electronic Engineering
Hebei University of Engineering
Handan, Hebei, China
e-mail: wudiwudi@hebeu.edu.cn

Abstract. The danmaku text clustering is a hot topic in online video reviews. Given the problem of unsatisfactory clustering accuracy caused by short text and many new words, the danmaku text clustering algorithm based on feature extension and word-pair filtering OBTM is proposed. First, a new-word discovery algorithm based on weight optimization is proposed to retain the features of new words in the danmaku text. Then, the internal information and external knowledge of new words are used to expand the features of the danmaku text for reduced feature sparsity. Furthermore, the OBTM topic model based on word-pair filtering is designed to eliminate noise features. Finally, the Single-Pass algorithm based on cluster center iteration is proposed to obtain the clustering results of topic feature words. Experimental results show that the algorithm proposed in this paper is 13.33 %, 8.52 %, 6.25 % higher than the OBTM, Word2vec + BTM, OurE.Drift* algorithm, respectively, in terms of clustering accuracy.

Keywords: Danmaku text, short text clustering, feature extension, OBTM, new word discovery

1 INTRODUCTION

Recently, sending danmaku has become a popular way to comment on videos. Most of these comments are short Chinese texts that are displayed directly on the video screen [1]. Danmaku is rich in topics and varied in content. Cluster analysis can integrate danmaku topics well. Compared with the comments in the comment area,

danmaku text is shorter and contains new words (Table 1). The danmaku text features are sparse and affect the clustering accuracy. Therefore, addressing the sparse features of danmaku text is the key to improve the clustering accuracy.

	Number of Texts (Entry)	Average Text Length (Characters)	Proportion of New Word (%)
Danmaku	5 942	13.60	16.28
Comments	3 242	45.44	1.26

Table 1. Statistics Related to The Danmaku and Comments in The Comment Area (video source: `https://www.bilibili.com/video/BV1JK4y1r7yY`)

The danmaku text analysis is generally divided into supervised and unsupervised methods. Supervised methods usually improve the accuracy of danmaku feature extraction through the text annotation method [2, 3]. However, the models of the supervised methods have high complexity, which take a long time to train and rely on the tagging of the danmaku word position or other information (e.g., user information and video topic.) tag addition. Therefore, the unsupervised methods of danmaku text analysis research have attracted attention. Li and Li [4] used the special emotional dictionary for danmaku to extract danmaku feature words and retained the context information of these feature words through the word vector model. Qing et al. [5] used the custom dictionary of net new words obtained the emotional feature value, and then used K-means to cluster the emotional feature value. The unsupervised method no longer relies on the bullet screen text labels. At the same time, the unsupervised method improves the accuracy of clustering and sentiment analyses by preserving the features of new words in the danmaku.

In the text preprocessing stage, net new words in the danmaku are usually deleted, resulting in sparse danmaku text features. Therefore, the danmaku text analysis methods retain the text features through new-word recognition. New-word recognition can be divided into rule [6] and statistical methods [7]. The statistic-based method includes word frequency statistics [8], mutual information [9], and left and right information entropy [10]. The rule-based method has high accuracy in identifying new words [11], but the work of summarizing rules is complicated. The statistic-based method is simple but has low accuracy of new-word recognition [12]. An algorithm combining rule-based and statistic-based methods [13] is proposed to solve the problem of large workload and low accuracy of new-word recognition. However, the grammatical rules of danmaku text are difficult to define. Compared with the above methods, the method of multiple statistical combinations proposed in this paper has higher accuracy in the recognition of new words in the danmaku.

Short comment texts similar to the danmaku text are mostly unlabeled. This type of data is suitable for clustering by using topic modeling methods. The topic model is an unsupervised method. The Biterm topic model (BTM) is a modeling method for short text. This method is proposed by Cheng et al. [14] and proves that BTM is more efficient than Latent Dirichlet Allocation (LDA) [15]. However, the text length is too short, and BTM may lose many potential coherent and remarkable

word co-occurrence patterns that cannot be observed in the corpus [16]. Therefore, the methods that extract specific terms based on the relational model [15] and expanding the text length based on the external knowledge [17, 18] are proposed to improve BTM. In addition, methods that enhance the text semantics based on the word embedding algorithm [19] and the secondary clustering based on the classic clustering algorithm [20] can improve the clustering accuracy of BTM. Given the problem of sparse text features caused by the short length of the danmaku text, this paper proposes the use of external knowledge to extend text features.

The danmaku text has the characteristics of coherent topic and is a type of short texts streams. When processing streaming data, the accuracy loss of the above model is relatively large. Therefore, twitter-LDA [21] and Online Biterm Topic Model (OBTM) [14] showed advantages, but the sparseness of short text features should be reduced. Hu et al. proposed the online BTM based short-text stream classification by using short-text expansion and concept drifting detection (OurE.Drift) [22]. On the basis of OBTM, this method uses external knowledge to extend short texts, further alleviating the sparsity of text space. However, the OurE.Drift is suitable for texts with standardized grammar. Danmaku is colloquial, and its external knowledge extraction depends on the feature words of the danmaku. The feature words of danmaku are mostly new words. Therefore, the external knowledge of danmaku new words is used to expand the text features in this paper. The danmaku is colloquial to produce feature noise, leading to the unsatisfactory result of topic extraction. Therefore, in this paper, low-frequency word pairs are deleted to reduce the feature noise in OBTM modeling. Through the two above improvements, the clustering accuracy of danmaku text is improved in a targeted manner.

The danmaku text clustering algorithm based on feature expansion and word-pair filtering OBTM (FEF-oBTM) is proposed to solve the problem of poor clustering accuracy caused by short text and many new words. The new-word recognition algorithm based on weight optimization is proposed, which uses weights to change the degree of influence of mutual information and information entropy on new word recognition. A feature expansion method based on danmaku new words, which uses word2vec to extract new word features in external knowledge and internal information of the text, is designed. An OBTM topic model based on word-pair filtering, which uses $tf - idf$ method to reduce noise word pairs, is designed. A Single-Pass clustering method based on cluster center iteration is proposed to improve the order sensitivity problem. These methods are used to reduce the feature sparsity of short text for improved clustering accuracy of the danmaku text.

The rest of this paper is organized as follows. In Section 2, we introduce the definition of new word, new word recognition algorithm and F-oBTM algorithm. Section 3 shows the details of the algorithm. In Section 4, experiments prove the effectiveness of the algorithm. Finally, we offer our conclusion in Section 5.

2 PROBLEM DEFINITION

2.1 Definition of New Word

New words have the characteristics of popularity and rapid change, and also known as unknown words. In computational linguistics, unknown words are words that have not appeared in dictionaries. For the lexicon-based model, new words refer to words that appear in corpus but not in the dictionary [23].

The jieba segmentation lexicon has advantages because it is widely used in China and can search the maximum probability path and most probable combination based on the word frequency [24]. Therefore, this paper uses jieba to segment the danmaku data set.

Definition 1. For the danmaku data set and preprocessing of this paper, a new word is defined as a word not included in the jieba segmentation lexicon.

2.2 New Word Discovery Based on Weight Optimization

In view of the internal compactness and external dispersion, a new-word recognition method based on pointwise mutual information and information entropy is selected. The short text length of danmaku and insufficient context information result in a small proportion of information entropy in new-word recognition. Therefore, the optimization of the weight of mutual information and information entropy is proposed.

The pointwise mutual information reflects the correlations between two adjacent words. A large PMI value of the pointwise mutual information results in a remarkable correlation between words x and y and vice versa. The pointwise mutual information formula is shown in Equation (1).

$$PMI\,(x,y) = \log_2 \frac{P\,(x,y)}{P\,(x)\,P\,(y)}, \tag{1}$$

$$P\,(x,y) = P\,(x|y)\,P\,(y)\,. \tag{2}$$

In Equation (1), $P\,(x)$ or $P\,(y)$ represents the total probability of occurrence of a single word. $P\,(x,y)$ represents the probability that words x and y appear simultaneously. According to Equation (2), $P\,(x,y)$ can be obtained. $P\,(x|y)$ represents the probability of x in the text containing y.

$$H_L\,(w_i) = -\sum_{\forall a \in A} P\,(aw_i|w_i) \times \log_2 P\,(aw_i|w_i)\,, \tag{3}$$

$$H_R\,(w_i) = -\sum_{\forall b \in B} P\,(w_i b|w_i) \times \log_2 P\,(w_i b|w_i)\,. \tag{4}$$

In Equation (3), $H_L\,(w_i)$ represents the left information entropy. $A\,(a \in A)$ is the set of words on the left of the preselected word w. $P\,(aw_i|w_i)$ indicates

the probability that on the left of w_i is a, when the preselected word is w_i. In Equation (4), $H_R(w_i)$ represents the right information entropy. $B(b \in B)$ is the set of words on the right of the preselected word w. $P(w_i b|w_i)$ indicates the probability that on the right of w_i is b, when the preselected word is w_i.

Let the pointwise mutual information weight be λ_1 and the information entropy weight be λ_2. According to the above formulas, the weighted optimized preselected word score is calculated as in Equation (5).

Definition 2. (preselected word score based on weight optimization) If the preselected word is w_i, then the preselected word score is shown as follows.

$$Score(w_i) = \lambda_1 PMI(x, y) + \lambda_2 |\min(H_L(w_i), H_R(w_i))|. \tag{5}$$

Wherein, $\min(H_L(w_i), H_R(w_i))$ represents the minimum value of entropy, which are negative. The smaller the entropy, the more stable relationship between the two characters and the greater the probability of becoming a new word. The calculation formulas for λ_1 and λ_2 are as follows.

$$\lambda_1 = \frac{\mathrm{PMI}(x, y) - \overline{\mathrm{PMI}}}{\overline{\mathrm{PMI}}}, \tag{6}$$

$$\lambda_2 = \frac{2 * \sqrt{\left(\mathrm{H}_L(w_i) - \overline{\mathrm{H}_L}\right)^2 + \left(\mathrm{H}_R(w_i) - \overline{\mathrm{H}_R}\right)^2}}{\overline{H_L} + \overline{\mathrm{H}_R}}. \tag{7}$$

In Equation (6) and Equation (7), $\overline{\mathrm{PMI}}$ refers to the average pointwise mutual information value of pre-selected words. $\overline{\mathrm{H}_L}$ and $\overline{\mathrm{H}_R}$ represent the average information entropy value of pre-selected words. $\overline{\mathrm{PMI}}$ and $\min(H_L(w_i), H_R(w_i))$ is optimized by λ_1 and λ_2 to determine influence degree of the pre-selected words. If $\lambda_1 > \lambda_2$, it means that $\overline{\mathrm{PMI}}$ has a greater influence on the preselected words than $\min(H_L(w_i), H_R(w_i))$; if $\lambda_1 < \lambda_2$, it means that $\overline{\mathrm{PMI}}$ has a smaller influence on the preselected words than $\min(H_L(w_i), H_R(w_i))$.

Preselected Word	Before Weight Optimized	After Weight Optimized
Danmaku	4.75	5.23
Yinghe	Unrecognized	11.7
Gaoneng	Unrecognized	11.7
Xiaoku	6.35	6.93

Table 2. Comparison of preselected word score (part)

Table 2 shows that the recognition rate of new words (such as "Yinghe") are improved after the weights are optimized. In addition, the scores of some new words (such as "Xiaoku") are improved.

2.3 OBTM Based on Biterm Filtering

The danmaku text contains some low-frequency words, which have no special meaning and have little effect on the text topic. When using external knowledge to expand the features of the danmaku, noise word pairs are introduced. These noise word pairs are related to danmaku new words but not related to the topic of the danmaku. Noisy word pairs reduce the accuracy of clustering. Therefore, the OBTM word-pair filtering method based on $tf - idf$ (F-oBTM) is proposed. The specific process is used to calculate the $tf - idf$ value of a word pair. Furthermore, whether the $tf - idf$ value is within the threshold range is a judgment. Finally, all noise word pairs are found in accordance with the judgment result. The $tf - idf$ is a statistical method that can evaluate the importance of words to text data sets. The calculation formula of $tf - idf$ is as follows.

$$tf - idf\,(b_i) = \frac{nb_i^d}{nd} * \left|\ln \frac{1+nD}{nb_i^D}\right| (d \in D)\,. \tag{8}$$

In Equation (8), nb_i^d represents the number of b_i in a certain document d. nd represents the total number of words contained in d. $\frac{nb_i^d}{nd}$ is the probability of occurrence of b_i in d. nb_i^D represents the total number of b_i in text set D. nD represents the total number of documents contained in D. $\ln \frac{1+nD}{nb_i^D}$ is the inverse document frequency of b_i.

Definition 3. (Threshold-based word pair selection) Assuming that the $tf - idf$ value of word pair is $tf - idf_{bi}$. The threshold values are ε and δ ($\varepsilon > \delta$). The word pair before filtering is b_i. Then the word pair filtered by the $tf - idf$ method is expressed as $b_i^{*(t)}$ and the formula of $b_i^{*(t)}$ is as follows.

$$b_i^{*(t)} = \{b_i | tf - idf_{bi} \geq \varepsilon, tf - idf_{bi} \leq \delta\} \left(b_i^{*(t)} \in B^{(t)}\right). \tag{9}$$

In Equation (9), f_{bi} is the $tf - idf$ value of the word pair b_i. The parameters ε and δ are based on the actual situation of the data set in the experiment. N_B represents total number of words. The word pair set is $B^{(t)}$ ($B^{(t)} = \left\{{b_i}^{*(t)}\right\}_{i=1}^{N_B}$, ${b_i}^{*(t)} = \left(w_{i,1}^{*(t)}, w_{i,2}^{*(t)}\right)$). The probability of $b_i^{*(t)}$ is shown as follows.

$$P(b_i^{*(t)} | \Theta^{(t)}, \Phi^{(t)}) = \sum_{k=1}^{K} \theta_k {\phi_{k,w}^{(t)}}^2. \tag{10}$$

In Equation (10), the Dirichlet distribution $\Theta^{(t)}$ and $\Phi^{(t)}$ represent document-topic distribution and topic-word pair distribution. In the case that time slice $t > 1$ and the hyperparameters $\alpha^{(t)}$ and $\beta^{(t)}$ are known, the distribution of θ_k and $\phi_{k,w}^{(t)}$

can integral to obtain probability of $b_i^{*(t)}$.

$$P(b_i^{*(t)}|\alpha^{(t)},\beta^{(t)}) = \iint \sum_{k=1}^{K} \theta_k \phi_{k,w}^{(t)} \,^2 \, \mathrm{d}\theta \, \mathrm{d}\phi. \tag{11}$$

Calculate the product of the probability of each word, that is, the probability of generating the entire corpus $B^{(t)}$.

$$P(B^{(t)}|\alpha^{(t)},\beta^{(t)}) = \prod_{i=1}^{N_B} \iint \sum_{k=1}^{K} \theta_k \phi_{k,w}^{(t)} \,^2 \, \mathrm{d}\theta \, \mathrm{d}\phi. \tag{12}$$

F-oBTM also uses Gibbs sampling to optimize the probability distribution. Gibbs sampling process of F-oBTM refers to [18].

3 DANMAKU TEXT CLUSTERING ALGORITHM BASED ON FEATURE EXTENSION AND WORD-PAIR FILTERING OBTM

The danmaku text suffers from low clustering accuracy, due to short texts and more new words. Therefore, the FEF-oBTM algorithm is proposed in this paper. In Figure 1, the FEF-oBTM algorithm is divided into the following four steps.

New word discovery stage: First, the jieba is used to segment the danmaku data set and the stop words are removed. Then, new words of danmaku are determined in accordance with Definition 2.

Feature extension stage: This stage is divided into three parts. In part I, new words are imported into the jieba dictionary. Then, the new jieba dictionary is used to segment danmaku text, and stop words are removed. Finally, the word2vec is used to extract the top5 related words of new words in the danmaku corpus. In part II, the external knowledge of danmaku new words is obtained by using the Baidu Encyclopedia entry and using word2vec to extract the top5 related words of new words in the external knowledge. In part III, the danmaku word set by the top5 feature words in the danmaku corpus and external knowledge, which are related to the new words, is extracted.

Topic modeling stage: The F-oBTM is used to extract topic words from the danmaku word set.

Text clustering stage: Feature words are clustered using the iterative Single-Pass algorithm.

3.1 New Word Discovery

The new-word recognition algorithm belongs to the danmaku preprocessing in the FEF-oBTM algorithm. Danmaku new words imply user emotions and opinions, which are not included in the word segmentation dictionary. Therefore, danmaku

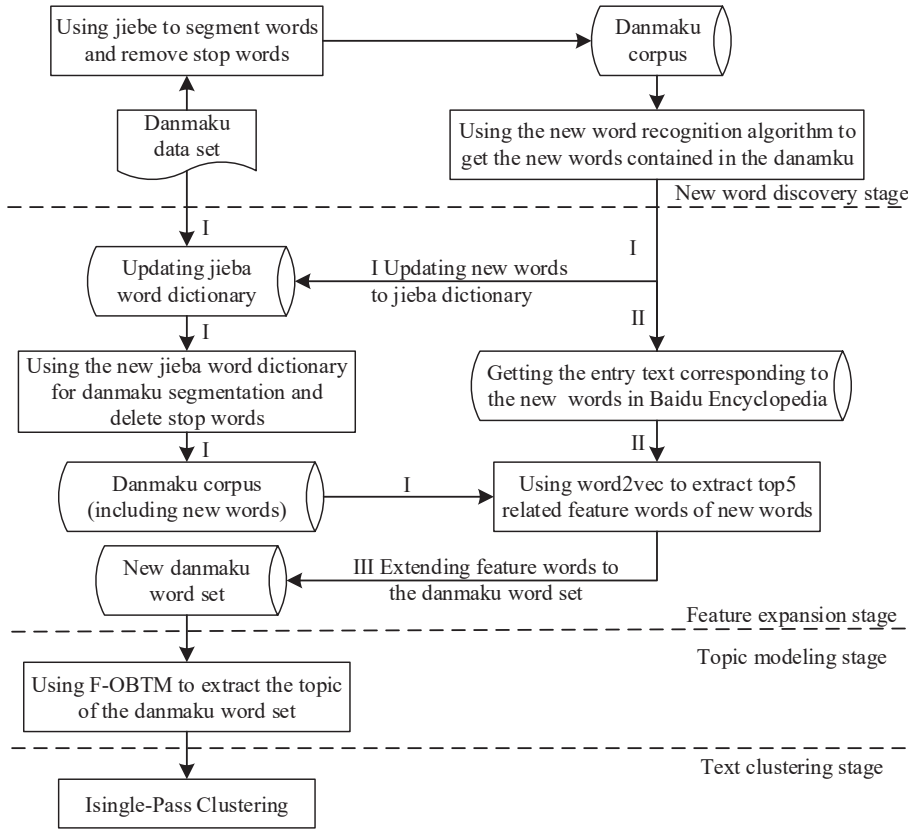


Figure 1. Danmaku text clustering algorithm based on feature extension and biterm filtering OBTM

preprocessing includes danmaku data set acquisition, format processing, word segmentation, stop word removal, new word recognition, word segmentation dictionary expansion, second word segmentation and second stop word removal. This preprocessing can retain new words of danmaku. The new-word recognition algorithm is as follows.

3.2 Feature Extension

In Section 3.1, the new words of danmaku, which improve the word segmentation effect, are retained. However, considering that the meaning of new words cannot be obtained literally, the features of new words should be extracted. Two common feature extension methods are available:

1. feature extension of short text based on external knowledge and

Algorithm 1 New Word Discovery Algorithm

Input: danmaku corpus corpus.txt
Output: preselected word w_i $(i = 1, 2, \ldots, N)$

Initializing dictionary tree T, $\lambda_1 - 0$, $\lambda_2 = 0$; /* T tree is a trie tree */
Recoding corpus w_i to T by line;
Traversing T to calculate the frequency of node P(node);
Traversing T to calculate the co-occurrence frequency of nodes bi_*node*;
Calculating PMI according to Equation (5);
Traversing T to calculate $H_L(w_i)$ according to Equation (3);
Traversing T to calculate $H_R(w_i)$ according to Equation (4);
Calculating λ_1 and λ_2 according to Equation (6) and Equation (7);
Calculating node score $Score(w_i)$ by Equation (4);
Output w_i.

2. feature expansion based on new words in the internal text.

Therefore, the feature extension method based on new-word processing, which combines the above two feature extension methods, is proposed. This method extracts the feature of new words of danmaku text and extracts the feature of new words in external knowledge. The algorithm is as follows.

Algorithm 2 Feature Extension Algorithm

Input: new word w_i, danmaku data set D, Baidu Encyclopedia entry set entry.txt
Output: danmaku word set D^*

Updating w_i to the jieba dictionary;
Using the new jieba word segmentation to preprocess D;
Writting the preprocessing result to D';
Using word2vec to extract Top5 related words T_i of w_i in entry.txt;
Using word2vec to extract Top5 related words R_i of w_i in D';
Writting T_i and R_i into the corresponding position of w_i in the danmaku word set to get D^*;
Output D^*.

In Algorithm 2, the Baidu Encyclopedia is selected as external knowledge base because the Baidu Encyclopedia meets the following general conditions:

1. Baidu Encyclopedia contains a large amount of data and rich content and
2. Baidu Encyclopedia entry covers most of the new words.

3.3 Topic Modeling Based on F-oBTM

The colloquialization of danmaku and the introduction of external knowledge lead to the increase of noise features. Therefore, the OBTM topic model based on word-pair filtering (F-oBTM) is proposed. This method improves the accuracy of topic

extraction by reducing noise word pairs. The F-oBTM model diagram is shown below.

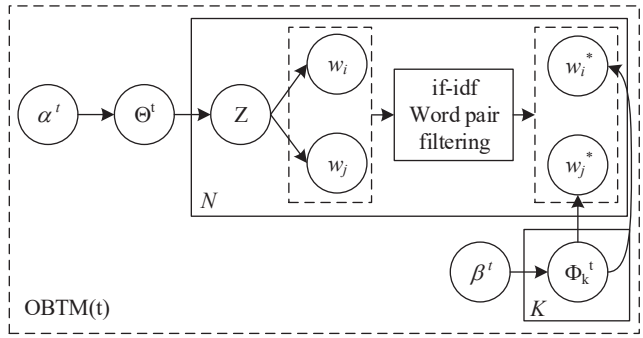


Figure 2. F-oBTM topic model

In Figure 2, the word pair b_i ($w_i \in b_i \,\&\, w_j \in b_i$) is filtered using $tf - idf$. First, the $tf - idf$ values of the danmaku word set are calculated according to Equation (8). Then, thresholds ε and δ ($\varepsilon > \delta$) are set to delete the unqualified word pairs. Finally, the remaining word pairs continue to perform F-oBTM modeling. The word pair filtering process is shown in Algorithm 3.

Algorithm 3 Word Pair Filtering Algorithm based on $tf - idf$

Input: Word pair set biterm.txt

Output: $B^{(t)}\left(b_i^{*(t)} \in B^{(t)}\right)$

1: Calculating the $tf - idf\,(b_i)$ according to Equation (8);
2: Setting $B^{(t)} = []$ and threshold $\varepsilon = 6.024$ and $\delta = 5.454$;
3: Filtering out $b_i^{*(t)}$ according Equation (10);
4: **if** $tf - idf\,(b_i) \in (\delta, \varepsilon)$ **then**
5: Deleting b_i;
6: **end if**
7: $B^{(t)} \leftarrow$ Organizing formats of w_i, w_j to $w_i\_w_j$:z; /* $(w_i, w_j) = b_i^{*(t)}$ */
8: Output $B^{(t)}$.

In Algorithm 3, $\varepsilon = 6.024$ and $\delta = 5.454$ are obtained through simulation experiment results.

3.4 ISingle-Pass Clustering

The Single-Pass algorithm [25] is simple and fast, resulting in its wide use in large-scale data processing. However, the Single-Pass algorithm is sensitive to the order of text arrival, which leads to unstable clustering results. The iterative Single-

Pass algorithm (ISingle-Pass) is proposed to reduce the influence of the text order (Figure 3).

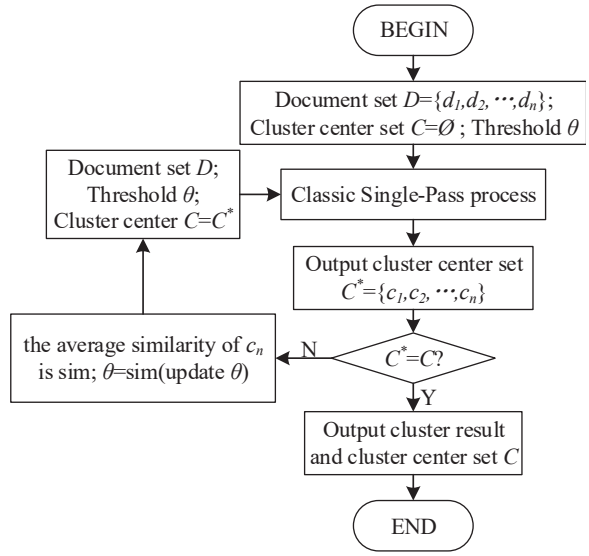


Figure 3. Iterative process of cluster center

The above figure shows the iterative process of cluster center C. Threshold θ represents the judgment condition of clustering. θ is constantly updated during the iteration. The average cluster similarity of the clustering result is calculated as sim, which is assigned to θ. C represents the result of the last clustering. If the current clustering result C^* is inconsistent with the last clustering result C, then C^* is re-input to the classic Single-Pass algorithm. If C^* and C are consistent, then the clustering ends. The ISingle-Pass clustering process is shown as Algorithm 3.

In Algorithm 4, the Pcorpora is a two-dimensional matrix (Pcorpora[i][j]). Wherein, [i][j] represents the j^{th} word in the i document. The Pcorpora[i][j] is processed by methods of doc2bow and $tf - idf$.

4 EXPERIMENT

To verify the advantages of the FEF-oBTM algorithm in the accuracy of danmaku text clustering, two data sets of Tsinghua NLP Classic Chinese News THUCNews and Bilibili Video danmaku are selected for comparison experiments. Four sets of contrast algorithms are designed: OBTM, Word2vec + BTM, OurE.Drift*, FEF-oBTM. The experimental method is described as follows.

OBTM topic model: This model was first proposed in the paper [14]. It directly models the short texts from slices of time. To obtain the topic-word pair dis-

Algorithm 4 ISingle-Pass Clustering Algorithm

Input: Pcorpora[i][j], Threshold $\theta = 0.005$

Output: Clusters number $n\_cluster$, center set C^*

```
1: n_cluster ← 1, maxValue ← 0, sim ← 0, i ← 0, C = {c_i}, C* = ∅;
   /* cluster set: cluster_n = []; maximum similarity maxsim = 0; average cluster
   similarity asim = 0 */
2: while C* != C do
3:    C ← C*;
4:    for (Pcorpora[i][j]) do
5:       Calculating the similarity maxsim between Pcorpora[i] and c_i;
6:       if maxsim > θ then
7:          Pcorpora[i] belongs to c_i, updating C*;
8:       else
9:          n_cluster ← n_cluster + 1, Pcorpora[i] belongs to c_{i+1}, writting to C*;
10:      end if
11:      i ← i + 1;
12:   end for
13:   θ ← maxsim;
14: end while
15: Output n_cluster, C*.
```

tribution, the probability of word pairs is counted for a certain time slice. The OBTM is the online method of the BTM topic model.

Word2vec + BTM: This model was first proposed in the paper [19]. It uses word2vec to train word vectors and then uses BTM to extract features from the word vectors.

OurE.Drift*: This method was derived from the text classification in the OurE. Drift algorithm proposed in the paper [22]. The external knowledge base is used to obtain text related to short texts content, which is long text and analyzed by LDA. The content is then added to the short texts for text expansion. Finally, the topic of short texts is extracted through OBTM.

FEF-oBTM: new words are identified in short texts, which includes entity nouns and network popular words. The relevant content of new words is obtained from Baidu Encyclopedia. Next, BTM is used to obtain the corresponding feature words. Then, the feature words are replaced or weighted in the original text. Furthermore, the topic is extracted through F-oBTM. Finally, the topic words are used for clustering by ISingle-Pass.

All experiments are carried on a PC, which the memory is 8.0 GB and the CPU is Intel(R) Core(TM) i5 1.60 GHz. The oBTM and F-oBTM models run in Ubuntu 16.04. The text preprocessing algorithms run in Windows 10. The algorithms are compiled on eclipse-Java-2019-09 and Visual Studio Code 1.39.2 (configuration: Python 3.7, C/C++).

4.1 Experimental Data Collection

This paper selects two data sets:

1. The classic Chinese news text data set THUCNews;
2. Bilibili video danmaku data set.

The THUCNews data set is provided by the Tsinghua NLP group. It includes a large amount of data. There are 14 categories about finance, lottery, real estate, stock, home, education, technology, society, fashion, politics, sports, constellation, games, entertainment. There are 6500 texts in each category. In order to ensure that the data volume of the two data sets is consistent, we selected the headline parts of the 6 news texts in THUCNews (Table 3). The news headline is regarded as a kind of short texts data, which is saved as THUCNews.txt.

Original format	0.txt	Motorola: GPON has more advantages than EPON in FTTH In 2009, under the fierce domestic trend...
	1.txt	France welcomes science students business school students are easy to find jobs In the lecture on studying abroad at the French Cultural Open Day on March 24...
Selected content	Motorola: GPON has more advantages than EPON in FTTH France welcomes science students business school students are easy...	

Table 3. Example of THUCNews data set

The danmaku data set is sourced from the video danmaku of a website (bilibili.com). The tool Fiddler 4 is used to parse the video web page content and find the XML file corresponding to the video danmaku (The URL of the XML danmaku file is: api.bilibili.com/x/v1/dm/list.so ? oid = *). The danmaku are exported in XML. Then the preprocessed danmaku text is saved as danmaku.txt. The result is shown in Table 4.

XML format	<d p="2421.65300,1,25,16777215,1542016731,0,9cabda7e,166307 99812001792">One Belt One Road, one waterway, one land road, long-term vision.</d> <d p="44.27800,1,25,15138834,1561558671,0,d5395ff9,180441560 49883136">No one can get good, only hurt the harmony, ask for trouble.</d>
pre-processed format	One Belt/One Road/ /one/waterway/one/land road/ /vision/ long-term/ / no/good/hurt/harmony/ /ask for/trouble/ /

Table 4. Example of danmaku data set

The data set related information is shown in Table 5.

Data Sets (Name)	Number of Texts (Entry)	Average Text Length (Characters)	Number of Categories (pcs)
Danmaku	6 000	9.05	none
THUCNews	6 000	11.7	6

Table 5. Data set related information

4.2 Evaluation Methods

In order to objectively explain the effectiveness of the four methods, perplexity, F1-measure and NMI are used to evaluate the experimental results.

The experiments involve with topic model, so perplexity [26] is chosen as the evaluation method to select the optimal parameter. For the topic model, the aggregation of a document into a class depends on the distribution of the model being trained. This distribution is a probabilistic result and does not have unique certainty [27]. This uncertainty is reflected in the perplexity. The lower the perplexity, the better the effect of topic model. The calculation formula is as follows.

$$Perplexity\,(D) = \exp\left(-\frac{\sum \log\left(P\left(w\right)\right)}{N}\right), \tag{13}$$

$$P(w) = \sum P(z)P(z|d)P(w|z). \tag{14}$$

In Equation (13), $P(w)$ is calculated by Equation (14). Wherein, $P(z)$ refers to the probability of a topic. $P(z|d)$ refers to a document Probability of all included topics. $P(w|z)$ refers to the probability of a word in a topic. N refers the total number of words included in the danmaku text.

F1-measure is a commonly used and effective cluster evaluation method. The F1-measure combines the two criteria of precision and recall. The formula is as follows.

$$\text{F1-measure} = \frac{2\text{PR}}{\text{P} + \text{R}}. \tag{15}$$

Wherein, P refers to the accuracy rate for a document belonging to a certain type of cluster, and R refers to the completeness of a certain type of cluster containing a certain type of document.

In order to comprehensively evaluate the clustering accuracy, it is also necessary to analyze the internal correlation degree of the words clustered. NMI [28] can be used to describe the correlation of words in the same category. Therefore, NMI is to be used to internally evaluate the clustering accuracy. The formula is as follows.

$$\text{NMI}(x,y) = \frac{-2}{\sum p(x)\log p(x) + \sum p(x)\log p(x)} * \sum\sum p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right). \tag{16}$$

In Equation (16), $p(x,y)$ represents the joint probability distribution of word x and word y in the same category that appear simultaneously in a sliding window.

The $p(x)$ or $p(y)$ is the edge probability that the word x or word y appears in the sliding window within the range of the edge probability distribution.

4.3 Experimental Process and Analysis

4.3.1 The Optimal Weight λ

In this section, the optimal weights λ_1 and λ_1 are determined by calculating F1-measure. The weights λ_1 and λ_1 are set as shown in Table 6. Set weights to four cases:

1. $\lambda_1 = 0$, $\lambda_2 = 1$;
2. $\lambda_1 = 1$, $\lambda_2 = 0$;
3. $\lambda_1 = 0.5$, $\lambda_2 = 0.5$;
4. $\lambda_1 =$ Equation (6), $\lambda_2 =$ Equation (7).

F1-measure can be calculated according to Equation (15). P = number of correct recognition/total number of recognitions, R = number of correct recognition/total number of new words. The relationship between P, R and weights is shown in Figure 4.

	Weight	P	R
1	$\lambda_1 = 0$, $\lambda_2 = 1$	0.509	0.58
2	$\lambda_1 = 1$, $\lambda_2 = 0$	0.533	0.8
3	$\lambda_1 = 0.5$, $\lambda_2 = 0.5$	0.616	0.69
4	$\lambda_1 =$ Equation (6), $\lambda_2 =$ Equation (7)	0.692	0.9

Table 6. P and R under different weights

In Figure 4, when $\lambda_1 =$ Equation (6), $\lambda_2 =$ Equation (7), the values of precision and recall are the highest, and the effect is the best. The curve of F1-measure shows that the proportion of pointwise mutual information is smaller than that of information entropy. Because of the sparsity of danmaku text, the pointwise mutual information has less influence on the discovery of new words in the danmaku text. Actual effect of new word discovery is shown in Table 7.

4.4 Selection of the Optimal Parameter

In the experiment, it was found that the effect of the experiment using the default parameter values of the topic model was not ideal. Therefore, the parameter value needs to be adjusted. The parameters involved in the topic model include: hyper-parameters β and α, number of iterations n_iter, number of time slices day, decay factor λ and number of topics K.

The danmaku data set crawled the danmaku text within 4 days. The number of time slices is set to $day = 4$. The number of iterations and the decay factor

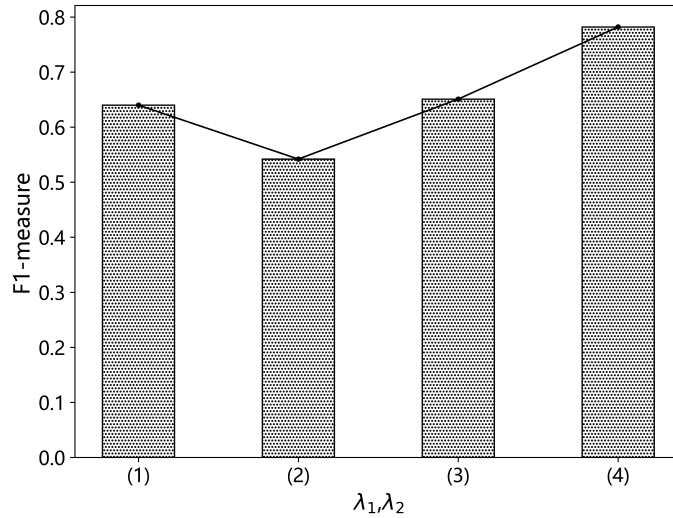


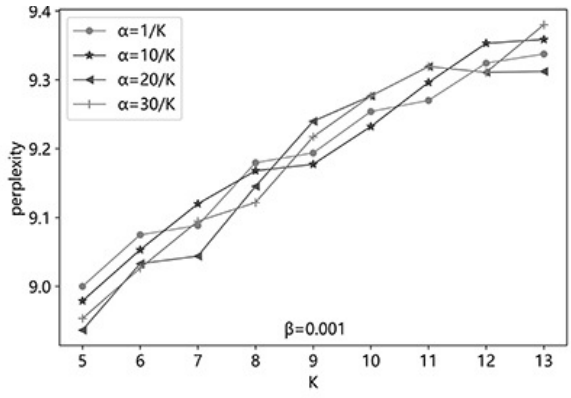
Figure 4. F1-measure of different weight λ

Preselected words	λ_1 = Equation (6) λ_2 = Equation (7)	$\lambda_1 = 0$ $\lambda_2 = 1$	$\lambda_1 = 1$ $\lambda_2 = 0$	$\lambda_1 = 0.5$ $\lambda_2 = 0.5$
Offshore capital	21.378	14.294	14.205	15.84
Diplomatic principle	20.856	20.514	21.065	23.604
Basic principles of WTO	20.356	17.139	17.334	17.756
Lanshouxianggu	19.401	15.139	15.189	15.606
Zhaxin Laotie	19.922	unrecognized	17.64	17.64
PANTA	19.398	unrecognized	unrecognized	15.544
Trade deficit	17.383	unrecognized	unrecognized	unrecognized
Bolixin	17.331	unrecognized	17.578	19.228
Yinghe Debut	17.328	unrecognized	14.201	14.578
Huawei Hongmeng	16.718	unrecognized	14.356	14.356

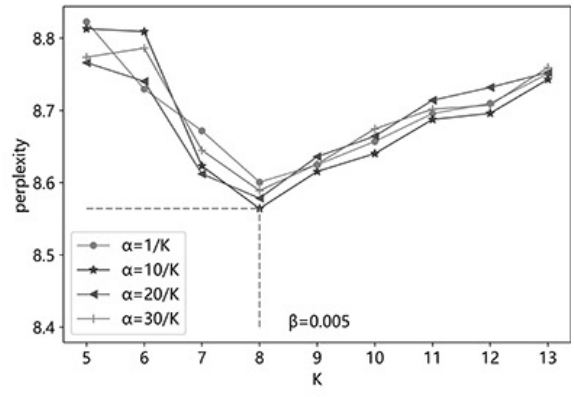
Table 7. New Word Recognition Algorithm Effect Display Score(w_i)

are set to experience values $n\_iter = 1\,000$ and $\lambda = 1$. We conducted experiments using the control variates method in the range of $\alpha = \{30/K, 40/K, 50/K, 60/K\}$, $\beta = \{0.0010.0050.010.05\}$ and $K = \{5, 6, 7, 8, 9, 10, 11, 12, 13\}$. Perplexity is used to evaluate the model to obtain the optimal parameter values. The perplexity is calculated according to Equation (13). The relationships among β, α, K and perplexity is shown in Figure 5.

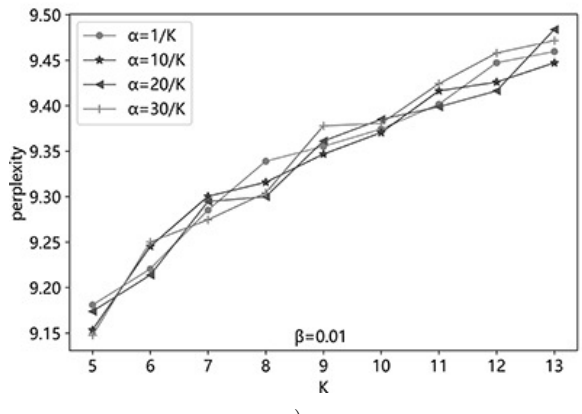
In Figure 5, when $\beta = 0.005$, $\alpha = 10/K$ and $K = 8$, the perplexity is minimal. It has proved that it is the optimal parameter of danmaku data set. The perplexity of $\beta = \{0.001, 0.01, 0.05\}$ changes more than that of $\alpha = \{1/K, 10/K, 20/K, 30/K\}$. Moreover, the $\beta = 0.005$ has little effect on the topic model.


α=1/K
α=10/K
α=20/K
α=30/K
perplexity
β=0.001
K

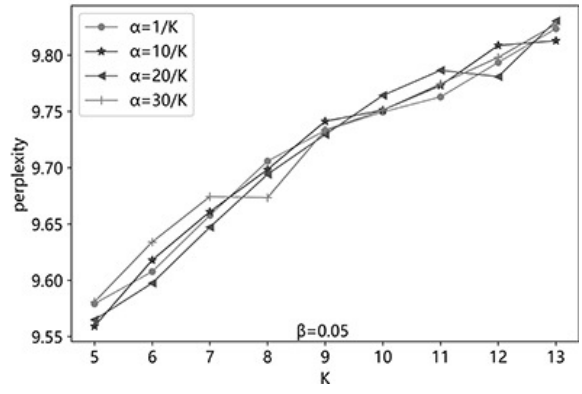
a)


α=1/K
α=10/K
α=20/K
α=30/K
perplexity
β=0.005
K

b)


α=1/K
α=10/K
α=20/K
α=30/K
perplexity
β=0.01
K

c)



d)

Figure 5. Perplexity of different β, α, K (danmaku data set)

The THUCNews data set includes 6 categories of news headlines, therefore, the number of topics is $K = 6$. The number of time slices, the number of iterations and the decay factor are set to empirical values $day = 3$, $n\_iter = 1\,000$ and $\lambda = 1$. We used the control variates method to carry out the experiment in the range $\alpha = \{30/K, 40/K, 50/K, 60/K\}$ and $\beta = \{0.001, 0.005, 0.01, 0.05\}$. The relationships among β, α and perplexity is shown in Figure 6.

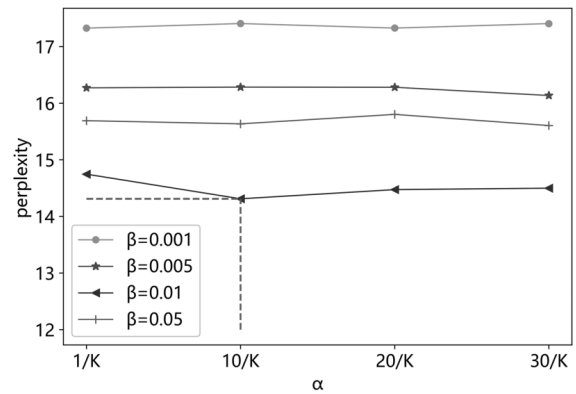


Figure 6. Perplexity of different n_iter (THUCNews data set)

In Figure 6, when $\beta = 0.01$ and $\alpha = 10/K$, the perplexity is minimal. It proved that the model works best in this case. Judging from the trend of the polyline, the α has little effect on the perplexity, which is also reflected in the danmaku data set.

Overall, the perplexity of the danmaku data set is smaller than that of the THUCNews data set. The text length of each danmaku is greater than the news

headline, resulting in low sparsity in modeling. Therefore, the model effect of the danmaku data set is better than that of the THUCNews data set.

The results of parameter selection on the two data sets are shown in Table 8.

Data Set (Name)	β	α	n_iter	day	K	λ
danmaku	0.005	$10/K$	1 000	4	8	1
THUCNews	0.01	$10/K$	1 000	3	6	1

Table 8. Optimal parameter of topic model

4.4.1 Clustering Accuracy Analysis

In this section, F1-measure and NMI are used to evaluate the clustering accuracy of the experimental method. The topic extraction results of OBTM, Word2vec + BTM and OurE.Drift* algorithms are clustering results. FEF-oBTM algorithm uses ISingle-Pass algorithm to perform clustering on the basis of F-oBTM topic extraction. First, save the top500 feature words of the F-oBTM topic extraction result. Then, the feature words of the 8 topics are divided into 800 documents in units of 5 words. Finally, the ISingle-Pass algorithm is used to cluster 800 documents.

The F1-measure needs to compare the clustering results with the actual classification results. Therefore, it is necessary to manually classify the danmaku data set. Referring to the manual classification result, the F1-measure was calculated. The THUCNews data set is known as 6 categories. The F1-measure results of the four experimental methods under different data sets are shown in Figure 7.

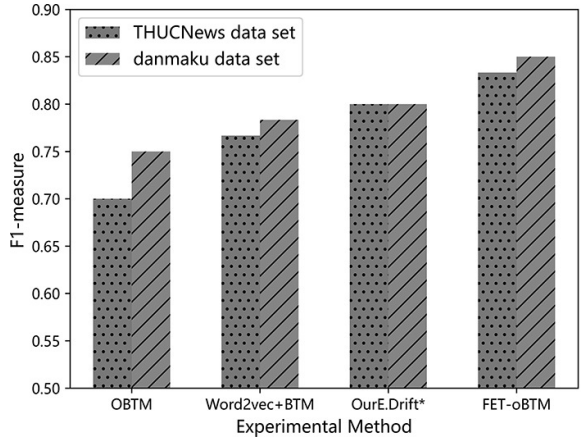


Figure 7. F1-measure comparison of experimental methods

In Figure 7, the F1-measure of the clustering from high to low is as follows: FEF-oBTM, OurE.Drift*, Word2vec+BTM, OBTM, the F1-measure of the barrage data

set is greater than that of the THUCNews data set. FEF-oBTM has the highest F1-measure degree, indicating that FEF-oBTM has the best clustering accuracy. The F1-measure improvement rate is shown in Table 9.

Data Set (Name)	Compared with OBTM	Compared with Word2vec + BTM	Compared with OurE.Drift*
danmaku	13.33 %	8.52 %	6.25 %
THUCNews	19.04 %	8.69 %	4.59 %

Table 9. Optimal parameter of topic model

In Table 9, FEF-oBTM combined feature extension method of OurE.Drift* and Word2vec + BTM, which made the clustering accuracy the best. OurE.Drift* used external knowledge for feature expansion. Word2vec + BTM used word2vec to vectorize text to refine and extract the internal information features. In addition, the F1-measure of the THUCNews data set is faster than that of the danmaku data set. The reason is that the news headlines in the THUCNews data set are more grammatical than the danmaku. The standardized text helped to extract the external knowledge of new words. The feature extension method is more effective for dealing with formal short texts.

The NMI results of the four experimental methods under different data sets are shown in Figure 8.

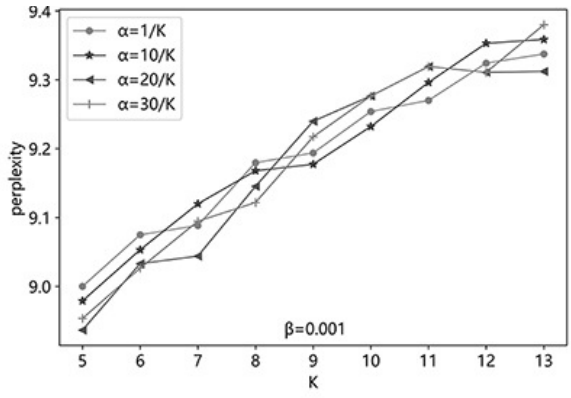
In Figure 8, the clustering NMI results from high to low are FEF-oBTM, OurE.Drift*, Word2vec+BTM, OBTM. The NMI of the danamku data set is greater than that of the THUCNews data set. The NMI of FEF-oBTM is the highest, which indicated that the feature words in each topic of FEF-oBTM have a high degree of relevance, that is, the clustering results are dense and the clustering accuracy is good. The NMI improvement rate is shown in Table 10.

Data Set (Name)	Compared with OBTM	Compared with Word2vec + BTM	Compared with OurE.Drift*
danmaku	17.06 %	9.38 %	3.55 %
THUCNews	20.73 %	16.11 %	4.80 %

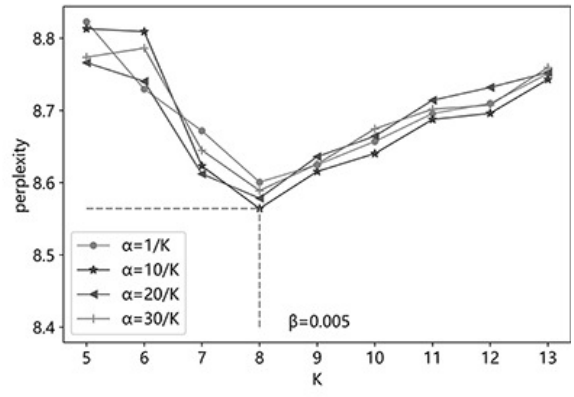
Table 10. Optimal parameter of topic model

In Table 9, in order to reduce the feature noise introduced by external knowledge, the FEF-oBTM algorithm filters word pairs. Reducing noise word pairs makes the danmaku feature words more relevant and further optimizes the clustering accuracy. In addition, the improvement rate of F1-measure of the THUCNews data set is faster than that of the danmaku data set, because the relevance of each type of text in the THUCNews data set is inherently high.

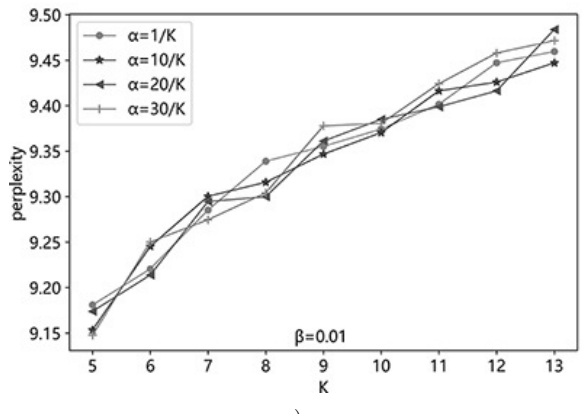
Moreover, the experimental effect of danmaku data set is better than THUCNews data set. The reason is that the danmaku texts is longer than the news headline of THUCNews data set and the topics of the danmaku texts are more coherent.


α=1/K
α=10/K
α=20/K
α=30/K
perplexity
β=0.001
K

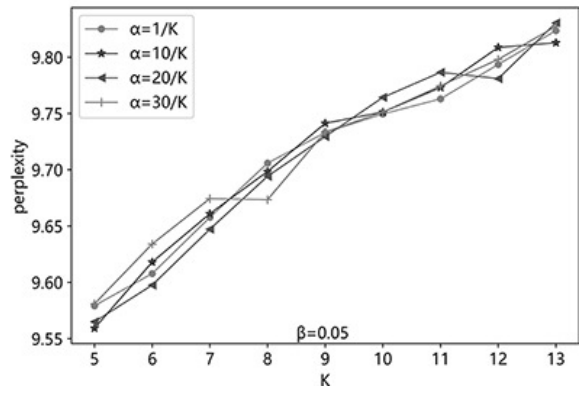
a)


α=1/K
α=10/K
α=20/K
α=30/K
perplexity
β=0.005
K

b)


α=1/K
α=10/K
α=20/K
α=30/K
perplexity
β=0.01
K

c)



d)

Figure 8. NMI comparison of experimental methods

The OBTM algorithm has the lowest F1-measure and NMI. The reason is that the text length of the data sets is too short. After processing with OBTM, the text features are very sparse, resulting in unsatisfactory experimental results.

The Word2vec + BTM method uses the word2vec algorithm to extract the features of the short text. The OurE.Drift* and FEF-oBTM methods enrich the short text by extracting external knowledge. These methods reduced the text sparsity. Therefore, the F1-measure and NMI of these methods are higher. Experiments show that the feature expansion method based on external knowledge is better than the feature expansion method based on internal information extraction.

In summary, FEF-oBTM has the highest F1-measure and NMI values. Feature expansion and word pairs filtering can remarkably improve the clustering accuracy of danmaku. The FEF-oBTM algorithm can effectively improve the accuracy of short text clustering.

5 CONCLUSIONS

Considering the problems of the danmaku text, i.e., being short and having new words, the danmaku text clustering algorithm based on feature extension and word-pair filtering OBTM is proposed. The algorithm uses the Baidu Encyclopedia knowledge to expand text features, uses the $tf - idf$ deleted noise word pairs of OBTM to reduce the text dimension, and improves the Single-Pass algorithm by using the cluster center iteration for the clustering of short text. In the experiment, we obtain the optimal values of experimental parameters by analyzing the perplexity results of the topic model. Experiments on the THUCNews and the danmaku data sets show that the FEF-oBTM algorithm is superior to the compared algorithms in terms of F1-measure and NMI. Compared with previous work, the FEF-oBTM algorithm is more suitable for the clustering of danmaku text and effectively improves the cluster-

ing accuracy. The danmaku data also contains attributes such as time, relationship, and number of likes. The use of these attributes to obtain the emotional changes of the danmaku is an interesting topic for future work.

Acknowledgements

This work is supported by Research Projects of the Nature Science Foundation of Hebei Province (No. F2020402003, F2019402428, F2021402005), National Natural Science Foundation of China (No. 62101174).

REFERENCES

[1] Bai, Q.—Hu, Q.—Fang, F.—He, L.: Topic Detection with Danmaku: A Time-Sync Joint NMF Approach. In: Hartmann, S., Ma, H., Hameurlain, A., Pernul, G., Wagner, R. R. (Eds.): Database and Expert Systems Applications (DEXA 2018). Springer, Cham, Lecture Notes in Computer Science, Vol. 11030, 2018, pp. 428–435, doi: 10.1007/978-3-319-98812-2_39.

[2] Ye, J.—Zhao, H.: A Public Opinion Analysis Model Based on Danmu Data Monitoring and Sentiment Classification. Journal of East China Normal University (Natural Science), Vol. 2019, 2019, No. 3, pp. 86–100, doi: 10.3969/j.issn.1000-5641.2019.03.010 (in Chinese).

[3] Wang, S.—Chen, Y.—Ming, H.—Huang, H.—Mi, L.—Shi, Z.: Improved Danmaku Emotion Analysis and Its Application Based on Bi-LSTM Model. IEEE Access, Vol. 8, 2020, pp. 114123–114134, doi: 10.1109/ACCESS.2020.3001046.

[4] Li, J.—Li, Y.: Constructing Dictionary to Analyze Features Sentiment of a Movie Based on Danmakus. In: Li, J., Wang, S., Qin, S., Li, X., Wang, S. (Eds.): Advanced Data Mining and Applications (ADMA 2019). Springer, Cham, Lecture Notes in Computer Science, Vol. 11888, 2019, pp. 474–488, doi: 10.1007/978-3-030-35231-8_34.

[5] Hong, Q.—Wang, S.—Zhao, Q. et al.: Video User Group Classification Based on Bullet Screen Sentiment Analysis and Clustering Algorithm. Computer Engineering and Science, Vol. 40, 2018, No. 6, pp. 1125–1139 (in Chinese).

[6] Peng, F.—Feng, F.—McCallum, A.: Chinese Segmentation and New Word Detection Using Conditional Random Fields. Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), 2004, pp. 562–568, doi: 10.3115/1220355.1220436.

[7] Huang, D.—Zhang, J.—Huang, K.: Automatic Microblog-Oriented Unknown Word Recognition with Unsupervised Method. Chinese Journal of Electronics, Vol. 27, 2018, No. 1, pp. 1–8, doi: 10.1049/cje.2017.11.004.

[8] Roul, R. K.—Sahoo, J. K.—Arora, K.: Modified TF-IDF Term Weighting Strategies for Text Categorization. 2017 14th IEEE India Council International Conference (INDICON), IEEE, 2017, pp. 1–6, doi: 10.1109/INDICON.2017.8487593.

[9] Pazienza, M. T.—Pennacchiotti, M.—Zanzotto, F. M.: Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. In: Sirmakessis, S. (Ed.):

Knowledge Mining. Springer, Berlin, Heidelberg, Studies in Fuzziness and Soft Computing, Vol. 185, 2005, pp. 255–279, doi: 10.1007/3-540-32394-5_20.

[10] Zhikov, V.—Takamura, H.—Okumura, M.: An Efficient Algorithm for Unsupervised Word Segmentation with Branching Entropy and MDL. Information and Media Technologies, Vol. 8, 2013, No. 2, pp. 514–527, doi: 10.11185/imt.8.514.

[11] Pathak, V.—Joshi, M.: Rule Based Approach for Word Normalization by Resolving Transcription Ambiguity in Transliterated Search Queries. 2019, doi: 10.48550/arXiv.1910.07233.

[12] Chen, K. J.—Ma, W. Y.: Unknown Word Extraction for Chinese Documents. Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), Vol. 1, 2002, doi: 10.3115/1072228.1072277.

[13] Miao, G.—Xu, J.—Li, Y.—Li, S.—Chen, Y.: An Unknown Word Processing Method in NMT by Integrating Syntactic Structure and Semantic Concept. In: Wong, D., Xiong, D. (Eds.): Machine Translation (CWMT 2017). Springer, Singapore, Communications in Computer and Information Science, Vol. 787, 2017, pp. 43–54, doi: 10.1007/978-981-10-7134-8_5.

[14] Cheng, X.—Yan, X.—Lan, Y.—Guo, J.: BTM: Topic Modeling over Short Texts. IEEE Transactions on Knowledge and Data Engineering, Vol. 26, 2014, No. 12, pp. 2928–2941, doi: 10.1109/TKDE.2014.2313872.

[15] Blei, D. M.—Ng, A. Y.—Jordan, M. I.: Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol. 3, 2003, pp. 993–1022.

[16] Li, X.—Zhang, A.—Li, C.—Guo, L.—Wang, W.—Ouyang, J.: Relational Biterm Topic Model: Short-Text Topic Modeling Using Word Embeddings. The Computer Journal, Vol. 62, 2019, No. 3, pp. 359–372, doi: 10.1093/comjnl/bxy037.

[17] Nakamura, T.—Shirakawa, M.—Hara, T.—Nishio, S.: Wikipedia-Based Relatedness Measurements for Multilingual Short Text Clustering. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), Vol. 18, 2018, No. 2, Art. No. 16, doi: 10.1145/3276473.

[18] Shirakawa, M.—Nakayama, K.—Hara, T.—Nishio, S.: Wikipedia-Based Semantic Similarity Measurements for Noisy Short Texts Using Extended Naive Bayes. IEEE Transactions on Emerging Topics in Computing, Vol. 3, 2015, No. 2, pp. 205–219, doi: 10.1109/TETC.2015.2418716.

[19] Wu, D.—Zhang, M.—Shen, C.—Huang, Z.—Gu, M.: BTM and GloVe Similarity Linear Fusion-Based Short Text Clustering Algorithm for Microblog Hot Topic Discovery. IEEE Access, Vol. 8, 2020, pp. 32215–32225, doi: 10.1109/ACCESS.2020.2973430.

[20] Li, W.—Feng, Y.—Li, D.—Yu, Z.: Micro-Blog Topic Detection Method Based on BTM Topic Model and K-Means Clustering Algorithm. Automatic Control and Computer Sciences, Vol. 50, 2016, No. 4, pp. 271–277, doi: 10.3103/S0146411616040040.

[21] Zhao, W. X.—Jiang, J.—Weng, J.—He, J.—Lim, E. P.—Yan, H.—Li, X.: Comparing Twitter and Traditional Media Using Topic Models. In: Clough, P., Foley, C., Gurrin, C., Jones, G. J. F., Kraaij, W., Lee, H., Mudoch, V. (Eds.): Advances in Information Retrieval (ECIR 2011). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 6611, 2011, pp. 338–349, doi: 10.1007/978-3-642-20161-5_34.

[22] Hu, X.—Wang, H.—Li, P.: Online Biterm Topic Model Based Short Text Stream Classification Using Short Text Expansion and Concept Drifting Detection. Pattern Recognition Letters, Vol. 116, 2018, pp. 187–194, doi: 10.1016/j.patrec.2018.10.018.

[23] Jia, Y.—Liu, L.—Chen, H.—Sun, Y.: A Chinese Unknown Word Recognition Method for Micro-Blog Short Text Based on Improved FP-Growth. Pattern Analysis and Applications, Vol. 23, 2020, No. 2, pp. 1011–1020, doi: 10.1007/s10044-019-00833-z.

[24] Chen, J.—Becken, S.—Stantic, B.: Lexicon Based Chinese Language Sentiment Analysis Method. Computer Science and Information Systems, Vol. 16, 2019, No. 2, pp. 639–655, doi: 10.2298/CSIS181015013C.

[25] Guha, S.—Mishra, N.—Motwani, R.—O'Callaghan, L.: Clustering Data Streams. Proceedings 41st Annual Symposium on Foundations of Computer Science, IEEE, 2000, pp. 359–366, doi: 10.1109/SFCS.2000.892124.

[26] Neubig, G.: Neural Machine Translation and Sequence-to-Sequence Models: A Tutorial. 2017, doi: 10.48550/arXiv.1703.01619.

[27] Newman, D.—Lau, J. H.—Grieser, K.—Baldwin, T.: Automatic Evaluation of Topic Coherence. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10), 2010, pp. 100–108.

[28] McDaid, A. F.—Greene, D.—Hurley, N.: Normalized Mutual Information to Evaluate Overlapping Community Finding Algorithms. 2011, doi: 10.48550/arXiv.1110.2515.

Di Wu received her B.Sc. and M.Sc. degrees in computer application technology from the Hebei University of Engineering, and her Ph.D. degree in computer application technology from Yanshan University. She is currently Associate Professor with the School of Software Engineering, Hebei University of Engineering, China.

Zhuyun Huang is currently pursuing her Master's degree with the Department of Information and Electronic Engineering, Hebei University of Engineering, China.