Short Text Classification: A Survey

Ge Song, Yunming Ye, Xiaolin Du, Xiaohui Huang, and Shifu Bie
Shenzhen Key Laboratory of Internet Information Collaboration, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, 518055, China
Email: carroll0708@qq.com, yeyunming@hit.edu.cn, {duxiaolinhitsz, hxh016}@gmail.com, bieshifu001@sina.com

***Abstract*—With the recent explosive growth of e-commerce and online communication, a new genre of text, short text, has been extensively applied in many areas. So many researches focus on short text mining. It is a challenge to classify the short text owing to its natural characters, such as sparseness, large-scale, immediacy, non-standardization. It is difficult for traditional methods to deal with short text classification mainly because too limited words in short text cannot represent the feature space and the relationship between words and documents. Several researches and reviews on text classification are shown in recent times. However, only a few of researches focus on short text classification. This paper discusses the characters of short text and the difficulty of short text classification. Then we introduce the existing popular works on short text classifiers and models, including short text classification using sematic analysis, semi-supervised short text classification, ensemble short text classification, and real-time classification. The evaluations of short text classification are analyzed in our paper. Finally we summarize the existing classification technology and prospect for development trend of short text classification.**

***Index Terms*—Short Text; Text Classification; Feature Selection; Semantic Analysis; Integrated Learning; Semi-Supervised Learning**

I. Introduction

With the explosion of e-commerce and online communication, short texts become available in many application areas, such as Instant Messages, online Chat Logs, Bulletin Board System Titles, Web Logs Comments, Internet News Comments, SMS, twitter etc. Therefore, successfully processing them becomes increasingly important in many Web and IR applications. However, it is a new challenges that classifying these sorts of text and Web data.

Unlike normal documents, these text and Web segments are usually noisier, less topic-focused, and much shorter, that is, they consist of from a dozen words to a few sentences [1]. Because of the short length, they do not provide enough word co-occurrence or shared context for a good similarity measure [40]. Therefore, normal machine learning methods, which are rely on the word frequency, enough word co-occurrences or shared context to measure the similarity of documents [41], usually fail to achieve desire accuracy due to the data sparseness.

New classifying methods on short text are appeared, such as sematic analysis, semi-supervised short text classification, ensemble models for short text, and real-time classification. However, compared with a lot of reviews and surveys on text classification, only few of surveys are appeared to discuss the recent researches on short text classification. This paper analyzes the challenges associated with classifying short text and systemic summarizes the existing related methods to short text classification using analytical measures.

After the analysis of the feature and difficulty of short text, we point out the process of short text classification in section II. In section III, short text classification based on semantic analysis is introduced. We describe some algorithms on semi-supervised short text classification in section IV. Section V and Section VI introduce the ensemble model for classifying short text and online short text classification, respectively. We analyze relevant evaluating measures in Section VII. In Section VIII, we summarize the methods for classifying short text.

II. Backgrounds

A. Feature of Short Text

Short text has been widely used in many fields, such as mobile short message, instant message, BBS title, news title, online chat record, blog comment, news comment, etc. And its main characteristic of the text length is very short, no longer than 200 characters. As mobile short message which we common used daily is no more than 70 characters, BBS title and news title less than 30. Instant messaging (IM) software supports longer message. For sending message quickly and ensuring it safely, IM software also limits its length, such as Windows Live Messenger of Microsoft allowing the longest message 400 characters. In fact, in daily communication, the instant message is only dozen words.

Generally, the features of short text are as follows [1] [2]:

Sparseness: a short text only contains several to a dozen words with a few features, it does not provide enough words co-occurrence or shared context for a good similarity measure. It is difficult to extract its valid language features.

Immediacy: short texts are sent immediately and received in real time. In addition, the quantity is very large.

Non-standardability: The description of the short text is concise, with many misspellings, non-standard terms and noise.

doi:10.4304/jmm.9.5.635-643

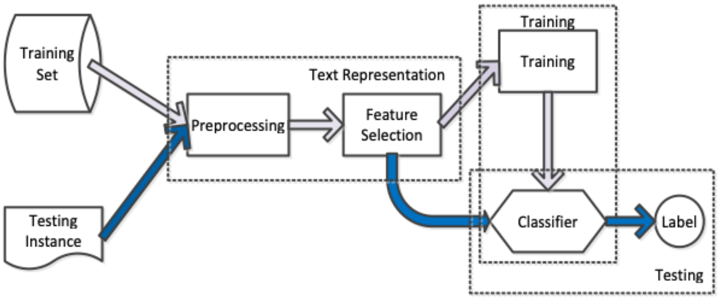

Figure 1. The processing of short text classification

Noises and imbalanced distribution: The application background (such as network security) needs to deal with massive amounts of short textual data. However, we may focus on only a small part (detecting objects) among the large-scale data. Therefore, useful instances are limited, and the distribution of short text is imbalanced.

Large scale data and labeling bottlenecks: It is difficult to manually label all of the large scale instances. Limited-labeled instances may only provide limited information. So how to make full use of these labeled instances and other unlabeled instance has become a key problem of short text classification.

Most of the traditional methods (such as SVM, BAYES, and KNN) are based on the similarity of term frequency, ignoring the feature of short text. These traditional methods may be not deal with the short text classification. Most of them (such as BAYES) may fail to obtain high accuracy if the labeled information is insufficient. In addition, some classification methods based on Vector Space Model (SVM) should use semantic information to improve the performance of the classifiers.

B. Short Text Classification

With the growing use of digital devices and the fast growth of the number of pages on the World Wide Web, text categorization is a key component in managing information. Automated categorization of text documents plays a crucial role in the ability of many applications to classify and provide the proper documents in a timely and correct manner.

Text classification can be defined simply as follows (Fig. 1): Given a set of documents D and a set of classes (or labels) C, define a function F that will assign a value from the set of C to each document in D. For example in short text classification, D might consist of the set of all classified advertisements in a newspaper, and C would therefore be the set of headings in the classified section of that same newspaper.

Learning to classify text and Web documents has been intensively studied during the past decade. Many learning methods, such as k nearest neighbors (k-NN), Naive Bayes, maximum entropy, and support vector machines (SVMs), have been applied to a lot of classification problems with different benchmark collections and achieved satisfactory results. However, traditional classification method was not good at short text classification, because of the character and difficulty of short text. Therefore, how to reasonably represent and choose features items, effectively reduce the spatial dimension and noises, and increase classification accuracy become the problem of short text classification.

III. SHORT TEXT CLASSIFICATION BASED ON SEMANTIC ANALYSIS

At present, the solution of reducing the feature spatial dimension is mainly based on the semantic features and semantic analysis. This is because the processing of text classification is generally in Vector Space Model (VSM) , which has the basic assumption that the relationships of words are independent, neglected the correlation between texts. However, short text has weaker capacity of semantic expression, which is needed this correlationship. While traditional classification cannot distinguish language fuzziness of natural language, cognates and synonyms, all of which are abundant in short text. Therefore traditional classification methods usually fail to achieve expected accuracy of short text.

Semantic analysis pays more attention to the concept, inner structure semantic level, and the correlation of texts to obtain the logic structure, which is more expressive and objectivity. In the existing researches, classification based on the Latent Semantic Analysis occupies an important position. Using the statistic method, latent semantic analysis extracts potential semantic structure, eliminates the synonymous influence, and reduces feature dimension and noises. Thus, many algorithms based on semantic analysis are proposed to deal with short text classification (More detailed information is shown in Table I) [3] [5-11] [44]. Zelikovitz [3] applies it to short text classification. Qiang Pu etc [5] combine LSA and Independent Component Analysis (ICA) [8] [42] together. Xuan - Hieu Phan etc. [7] establishes large-scale short text classification framework. The framework is mainly based on recent successful latent topic analysis models (such as pLSA and LDA) and powerful machine learning methods like maximum entropy and SVMs. Bing-kun WANG etc. [9] present a new method to tackle the problem by building a strong feature thesaurus (SFT) based on latent Dirichlet allocation (LDA) and information gain (IG) models. Language independent semantic (LIS) kernel [10] is proposed to enhance the language-dependency, when exploiting syntactic or semantic information. It is able to effectively compute the similarity between short-text documents without using grammatical tags and lexical databases. Mengen Chen etc.

[11] propose the method that extracts topics at multiple granularities, which can model the short text more precisely. Transductive LSA [3] [4] is another example of short text classification based on LSA. Transduction makes use of the test examples in choosing the hypothesis of the learner. The recreation of the space with the incorporation of the test examples does choose a representation based upon the test examples. The reduction of dimensionality of the training/test set combined allows the smaller space to more accurately reflect the test set to which it will be applied for classification. The inclusion of the test examples into the original matrix allows LSA to calculate entropy weights of words with the vocabulary and examples and co-occurrences of words in the test set.

In the following subsection, we describe detailed on definition and procession of LSA, pLSA, and LDA. Then we analyze the advantages and disadvantages of these three sematic analysis methods, respectively.

TABLE I. COMPARISON OF SEMANTIC APPROACHES FOR SHORT TEXT CLASSIFICATION

Reference	Model	Datasets	Measure
[3] [4]	Transductive LSA	Physics, NetVet , Business, News, Thesaurus	Accuracy
[5]	ICA, LSA	400 Chinese short-text documents	WS,BS,N, Similarity
[7]	pLSA, LDA	Wikipedia data, MEDLINE	Error, Accuracy,
[9]	SFT, LDA, IG	Chinese corpus, BBS	Precision, Recall, F1
[10]	LIS kernel	English and Korean datasets	Accuracy
[11]	LDA, Multi-Granularity Topics	Search Snippets, Chinese corpus, BBS	Accuracy

A. Short Text Classification Using Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) [12] [13] [43] is based upon the assumption that there is an underlying semantic structure in textual data, and that the relationship between terms and documents can be redescribed in this semantic structure form [14]. LSA transforms the vector space into the semantic space. Based on statistical method, LSA extracts and quantifies the semantic structure, eliminates the correlation between terms. LSA can reduce the high-dimensional vector matrix to construct the low-dimensional subspace which can effectively describe the relationship of term-document. Many Dimension reduction methods in LSA are proposed [15] [16], such as Singular Value Decomposition (SVD), Semi-discrete Decomposition (SDD), and Nonnegative Matrix. The most common method, SVD, should be introduced in our paper.

The process of LSA based on SVD is as follows [15] [16]:

Short documents are represented as vectors in a vector space. Therefore, the term-document matrix represents as $A_{mn} = \left[a_{ij} \right]_{m \times n}$, with each position corresponding to the absence or weighted presence of a term (a row i) in a document (a column j). This matrix is typically very sparse, as most documents contain only a small percentage of the total number of terms seen in the full collection of documents.

Compute the weight of a_{ij}. In order to focus on the contribution of each term (or document), we should compute the weight of a_{ij}. The traditional method is: $a_{ij} = LW_{ij} \times GW_{ij}$, where LW_{ij} is a local weight of the term I in the document j, GW_{ij} is the global weight of the term I in the entire dataset. We should obtain the local weight of a term by computing the log of the total frequency of the term I in the document j. The global weight of a term equal to the entropy of the term in the dataset. This entropy is based upon the number of occurrences of this term in each document.

The singular value decomposition (SVD) of $A_{mn} = \left[a_{ij} \right]_{m \times n}$ matrix. According to the above analysis, $A_{mn} = \left[a_{ij} \right]_{m \times n}$ matrix is very sparse. Moreover, in this very large space, some documents seem to be closer with each other by sharing common words. But these documents may be not related to each other semantically. Meanwhile, many documents that appear very distant by not sharing any term may actually closer. Because the same concept can be represented by many different words and words can have ambiguous meanings. LSA reduces this large space, and hopefully captures the true relationships between documents. To do this, LSA uses the singular value decomposition (SVD) of the term by $A_{mn} = \left[a_{ij} \right]_{m \times n}$ matrix. SVD of $A_{mn} = \left[a_{ij} \right]_{m \times n}$ is the product of three matrices:

$$A = \sum_{i=1}^{r} u_i \sigma_i v_i = \left[u_1 \cdots u_r \right] \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{bmatrix} \quad (1)$$

where u and v are the matrices of the left and right singular vectors and σ is the diagonal matrix of singular values. The diagonal elements of σ are ordered by magnitude, and therefore these matrices can be simplified by setting the smallest k values in σ to zero. The columns of T and D that correspond to the values of σ that were set to zero are deleted.

$$A = U_r \sum\nolimits_r V_r^T \approx U_k \sum\nolimits_k V_k^T \quad (2)$$

Classification based on LSA. According to SVD, The form of the reduced space $U_k \sum_k V_k^T$ is similar to the form of the original vector. Therefore, all of the classification algorithms that are suitable for vector space model can also apply to LSA classification model [17]. Many classification methods that combine LSA and traditional algorithms [18] [19] [20], such as sequence classification algorithm, Naive Bayes, KNN [4], and SVM are proposed to improve the accuracy of classifying short text.

To sum up, the advantage of the latent semantic analysis as follows:

Reduce the feature distribution and noise. According to LSA, K-dimensional semantic space is extract. This semantic space not only keeps the most information of the original vector space, but also reduces the dimension. Meanwhile, LDA eliminate the noise by discarding useless feature. Therefore, LSA can effectively deal with large-scale short test dataset;

Strengthen the semantic relationship. In latent semantic space, vectors no longer simply mean the frequency and distribution. They can describe the semantic relationship between terms and documents. LSA create less dependency on polysemy, synonyms, and common words, which may lead to low accuracy in traditional vector space. Previous researches show that at a relatively low-dimensional space with higher semantic expression, the performance of classification will be improved by similarity analysis

Flexibility. Since terms and documents are mapped to the same K-dimensional space, the LSA model can not only analyze the similarity between term-term (as traditional models do), but also obtain the better effect on analyzing the similarity between document-document and term-document.

Although the LSA model is the effective for short text classification, it still has some defects in the as follows:

Lose the structural information as reducing the dimension. Since feature space is the semantic space based on the text information, it may retain the main global information of the original feature matrix, ignoring some features that contribute a lot in the local space, but are insignificance in the global space.

SVD does not have the strict mathematical sense. In addition, it cost more computing time and space complexity in high-dimensional space.

The meaning of document is represented by the linear summation of vectors, ignoring grammar information of words in the phase of information extraction.

LSA can only deal with the visible variable. However, some meanings such as metaphor and analogy cannot be calculated [20] [21].

B. Short Text Classification Using Probabilistic Latent Semantic Analysis (pLSA)

Probabilistic Latent Semantic Analysis (pLSA) [20] [21] is proposed by Hofmann at 1999. The principle of pLSA is stronger than LSA. Moreover, It explicitly introduced the concept of “latent topic”. The latent topic is defined as the latent variable during a random process. According to fitting the training data, the probability model P is shown below:

$$P\left(d,w\right)=P\left(d\right)P\left(w\middle|d\right)=\sum_{z\in Z}P\left(w\middle|z\right)P\left(z\middle|d\right)P\left(d\right) \qquad (3)$$

where z is the latent class variables, d (document) and w (word) are observed variables, which are independent on the conditional probability of latent topic z. The probability model P can be expressed form of SVD as follows [28]:

$$P=U_k\sum\nolimits_k V_k^T \qquad (4)$$

where

$$\begin{aligned} U_k &= \left(P\left(d_i\middle|z_k\right)\right)_i \\ \sum\nolimits_k &= diag\left(P\left(z_k\right)\right)_k \\ V_k^T &= \left(P\left(w_j\middle|z_k\right)\right)_{j,k} \end{aligned} \qquad (5)$$

For a particular training set, the probability of document $P\left(d\right)$, the word probability $P\left(w\right)$ is known, and conditional probability $P\left(z\middle|d\right)$, $P\left(w\middle|z\right)$ is unknown. According to the principle of maximum likelihood estimation the maximum of the logarithm likelihood function is calculated by expectation maximum algorithm (EM) to fit the following model [20] [21]:

$$L=\sum_{i=1}^{N}\sum_{j=1}^{M}f\left(d_i,w_j\right)\log P\left(d_i,w_j\right) \qquad (6)$$

where $f\left(d_i,w_j\right)$ is the frequency of term w_j in document d_i.

The feature vector should be considered as the left singular vectors (the relationship between document and latent factors) or right singular vector (the relationship between terms and latent factors) in the procedure of classification.

Compared with the LSA model, pLSA model has the following advantages:

The pLSA model is the approximation model based on probability function with polynomial sampling. pLSA acquires the more stable and better approximation effect. However, the LSA model involves in Gaussian noise hypothesis.

The probability distribution of each variable has been clearly defined in the pLSA model, but LSA fails to define normal probability distribution.

pLSA can determine the optimal k-dimension using the existing statistical method, while LSA is mainly based on heuristics to determine the optimal k-dimension with more computational complexity for model selection.

pLSA model, however, still has some shortcomings:

The pLSA model needs to acquire the prior probability of label, which is computed by training set. However, it does not have a suitable prior probability to describe the unknown test corpus

The parameter space is increased with increasing training instances in pLSA model. This phenomenon may lead to excessive fitting problems. Meanwhile, too many discrete features that are only suitable for training set is existed, but these features cannot properly describe the unknown testing set.

C. Short Text Classification using Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) [7] is a probabilistic generative model that builds the linear discriminant function on the input variable. It is looking for a kind of transformation to acquire the maximum separability between classes and minimum difference within the class, LDA is a generative graphical model as shown in Fig. 2. It can be used to model and discover underlying topic structures of any kind of discrete data in which text is a typical example. LDA was developed based on an assumption of document generation process.

The main process of LDA is shown as: Assume that the text corpus contains several linear latent topics. According to probability inference algorithm, each document should be represented as the probability form on these latent topics. In the training phase, since LDA is the generative model, generative classification algorithms (such as MaxEnt algorithm [23]) are used to build the classification model. That is, the documents in each class independently train the sub-LDA model and share a set of topics, while the topics belong to different class are separated. In the predicting phase, the testing instance is generated by all of the sub-LDA models, and predicts its label by calculating the testing instance generative probability on each class.

It is worth to point out that the key problem of constructing LDA model is how to acquire the distribution information of latent topics within the document. Some algorithms such as Calculus of variations [24], Expectation Maximization (EM) [25], and Gibbs sampling algorithm [26] are used to deal with this question.

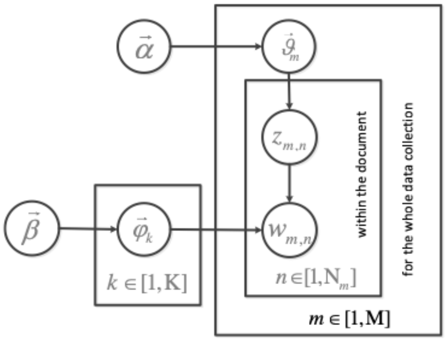

Figure 2. The structure of LDA

Compared with LSA and pLSA, LDA can find the latent structure of "topics" or "concepts" in a test corpus. The advantage of LDA is shown as

The Dirichlet probability distribution is used in LDA. This probability distribution (a continuous distribution) can give the unknown instance a probability of belonging to some topic set;

LDA directly chooses a suitable topic set from the topic distribution compared with pLSA that need a prior probability of unknown instance.

The LDA model has stronger ability of describing the realistic sematic. The LDA model which inherites all the advantages of the pLSA generative model, is more close to realistic sematic environment.

IV. Semi-Supervised Short Text Classification

Semi-supervised learning refers to the use of both labeled and unlabeled data for training. It contrasts supervised learning (data all labeled) or unsupervised learning (data all unlabeled). Other names are learning from labeled and unlabeled data or learning from partially labeled/classified data. It is found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy. The acquisition of labeled data for a learning problem often requires a skilled human agent to manually classify training examples. The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive. In such situations, semi-supervised learning can be of great practical value [27].

Most of the semi-supervised short text classifications are flexible semi-supervised learning. It can utilize unlabeled data to improve the classifier. However, unlike traditional semi-supervised learning algorithms, the universal data and the training/test data are not required to have the same format. In addition, once estimated, a topic model can be applied to more than one classification problems provided that they are consistent.

Reference [28] proposes a new semi-supervised short text classification algorithm. It uses the unlabeled corpus as "background knowledge" for the learner. A concrete example can be seen in the task of assigning topic labels to technical papers. Any title containing a word on galaxy (such as galaxy) should be easily classified correctly as an astrophysics paper, since the feature term in the title is common. However, an article on a less common topic, as for example old white dwarfs, should be able to correctly classify by utilizing a corpus of unlabeled paper abstracts from the same field. Those unlabeled paper abstracts are most similar to both old white dwarfs and to various training titles, each of which is quite dissimilar to it when compared directly.

Background knowledge may provide with a corpus of text that contains information both about importance of words and joint probability of words [29] [6]. We can use this background combined with the training examples to label a new example.

```
SELECT Test.instance, Train.label
FROM Train AND Test AND Background
WHERE Train.instance SIM Background.value
AND Test.instance SIM Background.value
```

If the features of the sample are special fewer, some background information that connected training samples and testing samples may not be found. We can use it as follows:

```
SELECT Test.instance, Train.label
FROM Train AND Test AND Background as B1
AND Background as B2
WHERE Train.instance SIM B1.value
AND Test.instance SIM B2.value
AND B1.value SIM B2.value
```

Reference [30] establishes information flow corpus of related topics in HAL space to classify the document title (short text). In that reference, Latent semantic relationship (such as co-occurrence relationship) among topics is represented by information flow. The feature HAL vector is extracted in the HAL model. Feature HAL vector is very sensitive to semantic information (especially to context semantic information). Then, the degree of the vector is calculated as follows:

$$\deg ree\left(c_i \lhd c_j\right) = \frac{\sum\limits_{p_l \in \left(QP_\mu(c_i) \wedge QP_\mu(c_j)\right)} w_{c_i p_l}}{\sum\limits_{p_k \in QP_\mu(c_i)} w_{c_i p_k}} \qquad (6)$$

According to dynamic corpus, information inference model is established.

Another algorithm based on contextual information is Distributional term representations (DTRs) [31]. DTRs represent terms by means of contextual information, which are given by document occurrence and term co-occurrence statistics. This algorithm enriches document representations that help to overcome the small-length and high-sparseness short text to some extent. More detailed on semi-supervised approaches are shown in Table II.

TABLE II. COMPARISON OF SEMI-SUPERVISED APPROACHES FOR SHORT TEXT CLASSIFICATION

Reference	Model	Datasets	Measure
[28]	background knowledge	Technical papers, News, Web page titles, Companies	Error rate
[30]	HAL	Reuters	Precision
[31]	DTRs	reduced Reuters R8 news corpus, EasyAbstract, CICLing2002	Error, Accuracy,

V. ENSEMBLE SHORT TEXT CLASSIFICATION

Single classifier generally based on similarity of items feature to classification is difficult to obtain the very good prediction result of short text classification since it is difficult to compute the similarity if the feature space is sparseness. Ensemble learning methods, on the other hand, through assigning a right weight for each weak classifier, get the weight of each feature; it is suitable for solving short text classification problem.

Reference [1] proposes a new dynamic assembly classification algorithm for short text, to solve the problems of sparse features and unbalanced data of the short text. In this method, to reduce the impact of the sparse features and unbalanced data, a treelike assembly classifier was constructed to support the classification. Then a dynamic strategy is presented to adjust the combinational structure in an adaptive way. Aixin Sun is proposed a short text classification method using very few words [32]. The predicted category label in this method is the majority vote of the search results, which are obtained by searching for a small set of labeled short texts best matching the query words.

A new model [33] is proposed to directly measure the correlation between a short text instance and a domain instead of representing short texts as vectors of weights. Firstly domain knowledge for each user-defined domain is drawn using an external corpus of longer documents. Secondly, the correlation is calculated. Finally, if the correlation is greater than a threshold, the instance will be classified into the domain.

To address this problem that short texts in Twitter do not provide sufficient word occurrences, another algorithm [34] is proposed with a small set of domain-specific features extracted from the author's profile and text in twitter. The proposed approach effectively classifies the text to a predefined set of generic classes. More detailed information on ensemble models are shown in Table III.

TABLE III. COMPARISON OF ENSEMBLE MODELS FOR SHORT TEXT CLASSIFICATION

Reference	Model	Datasets	Measure
[1]	Dynamic ensemble	Chinese BBS	Precision, Recall, F1
[32]	IR, Voting	Web snippet dataset	Accuracy
[33]	Domain knowledge,	External Corpus, Micro-blog Dataset	Precision, Recall, F1
[34]	Domain - specific features	recent tweets from random users	Accuracy

VI. REAL-TIME CLASSIFICATION OF LARGE SCALE SHORT TEXT

Immediacy is another feature of short text, which means Short texts are sent immediate and received in real time and usually the quantity is very large. So how to classify large-scale short text data immediately also becomes an important problem. Currently, comparing with several classic classification algorithm, it is often chose Bayes algorithm as online classifier. Naive Bayes algorithm judges categories through calculating probability of text belongs to each category, which is a simple, accuracy and widely used algorithm [39].

TABLE IV. COMPARISON OF REAL-TIME MODELS FOR SHORT TEXT CLASSIFICATION

Reference	Model	Datasets	Measure
[35]	Online and offline	messages	Precision,Recall, Error,
[36]	Extent Bayesian filtering technique	English and Spanish SMS spam collections	ROC curves
[37]	naïve Bayesian	SMS spam collections	Precision, Recall

Reference [35] proposed a spam message filtering system which combined online filtering with offline classifying. The system could filter the messages efficiently according to the features of sending behavior and the content of the messages using Naive Bayes algorithm. Further, the system uses a feedback self-learning mechanism, so the classifiers could improve themselves according to the filtering result. Another spam messages filter systems are based on Native Bayes and SVM [36] [37]. Native Bayes advantage of rapid statistics classification and SVM incremental training feature in this system, and updates the keywords database in time to enhance the self-adaptability. More detailed description on real-time algorithms are shown in Table IV.

VII. Evaluation of Short Text Classification

How to evaluate the model is another important problem in short text classification. Some measures that are used to compare and evaluate the classification method mainly include [38]:

Accuracy,
Precision and recall
F-measure.
Macro average and micro average

A. Accuracy

Accuracy is the most basic classification evaluation measure. The accuracy is computed:

$$accuracy = \frac{TP + TN}{N} \tag{7}$$

where N is the total number of the testing instances, TP, FP, FN, TN is described in the Table V:

TABLE V. The contingency table for category C

		True Label	
		Yes	No
Classifier Label	Yes	TP	FP
	No	FN	TN

$$TP = \sum_{i=1}^{C} TP_i , FN = \sum_{i=1}^{C} FN_i ,$$

$$FP = \sum_{i=1}^{C} FP_i , TN = \sum_{i=1}^{C} TN_i ,$$

where C is the total number of the classes.

B. Precision and Recall

Classification effectiveness is usually measured in terms of precision and recall, which are described as follows:

$$\Pr ecision = \frac{TP}{TP + FP} \tag{8}$$

$$\operatorname{Re} call = \frac{TP}{TP + FN} \tag{9}$$

C. F-measure

The precision and recall is reciprocal relationship. The purpose of classification is to get precision and recall together. F-measure is considered both of them, whose weather is described by parameters b.

$$F_\beta = \frac{\left(\beta^2 + 1\right) \times \Pr ecision \times \operatorname{Re} call}{\beta^2 \times \Pr ecision \times \operatorname{Re} call} \times 100\% \tag{10}$$

If $\beta = 1$, it is called breakeven point, which is widely used:

$$F_1 = \frac{2 \times \Pr ecision \times \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \times 100\% \tag{11}$$

D. Macro Average and Micro Average

Precision, recall and F_1 are aimed at a class with only local significance. Now commonly comprehensive measures (macro average and micro average) are shown below:

$$Micro \operatorname{Re} call = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + FN_i} \times 100\% \tag{12}$$

$$Micro \Pr ecision = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + FP_i} \times 100\% \tag{13}$$

$$MicroF_1 = \frac{2 \times Micro \Pr ecision \times Micro \operatorname{Re} call}{Micro \Pr ecision + Micro \operatorname{Re} call} \times 100\% \tag{14}$$

$$Macro \operatorname{Re} call = \frac{1}{C} \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + FN_i} \times 100\% \tag{15}$$

$$Macro \Pr ecision = \frac{1}{C} \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + FP_i} \times 100\% \tag{16}$$

VIII. Summary

Nowadays, the rapid development of information dissemination and media, especially the rise and development of instant communication lead to short text are widespread use, such as its application in topic tracking and discovery, catchword analysis, and network security. Short text has its own features, such as sparseness, large-scale, immediacy, and non-standardability. Therefore, normal machine learning methods usually fail to achieve desire accuracy due to the data sparseness.

At present, short text classification algorithm mainly classify in the following aspects:

Reduce feature dimension and extract feature using semantic relationship. Such as an LSA, LDA classification model.

Combined with plenty of unlabeled text, use semi-supervised classification algorithm to solve label bottleneck problems.

Use ensemble classification to improve classification accuracy.

Combine online classification and off-line classification to deal with large scale short texts.

However, short text classification is a challenging field, because many technologies are in the initial stage as well as the difficulties of classification didn't get the excellent solution such as how to design dynamic short text stream classification model. According to the passage of the application of short text, it produces several other problems such as multi-label short-text classification, comment emotional classification spam filtering, and topic tracking and control.

ACKNOWLEDGMENT

This research was supported in part by NSFC under Grant No. 61073195, Shenzhen Science and Technology Program under Grant No. CXY201107010163A, Shenzhen Strategic Emerging Industries Program under Grants No. JCYJ20130329142551746 and No. JCYJ20120613135329670.

REFERENCES

[1] YAN Rui, CAO Xian-bin, LI Kai, "Dynamic Assembly Classification Algorithm for Short Text," *ACTA ELECTRONICA SINICA*, Vol. 37(5), pp. 1019-1024, 2009.

[2] SU Jin-Shu, ZHANG Bo-Feng, and XU Xin. "Advances in Machine Learning Based Text Categorization," *Journal of Software*, Vol. 17(9), pp. 1848−1859, 2006.

[3] Zelikovitz, S, Marquez, F, "Transductive Learning for Short-Text Classification Problems using Latent Semantic Indexing," *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. l9(2), pp. 143-163, 2005.

[4] Zelikovitz, S, "Transductive LSI for Short Text Classification Problems," *Proceedings of the 17th International Flairs Conference*, pp. 556—561, 2004.

[5] Qiang Pu and Guo-Wei Yang, "Short-Text Classification Based on ICA and LSA," *Advances in Neural Networks - ISNN 2006*, pp. 265-270, 2006.

[6] S. Zelikovitz and H. Hirsh, "Using LSI for text classification in the presence of background text," *In Proceedings of 10th International Conference on Information and Knowledge Management*, pp. 113-118, 2001.

[7] Xuan-Hieu Phan, Le-Minh Nguyen, Susumu Horiguchi, "Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections," *WWW 2008 Refereed Track: Data Mining – Learning*, pp. 91-100, 2008.

[8] BACH F JORDAN M I, "Kernel independent component analysis," *The Journal of Machine Learning Research*, Vol. 3(11), pp. 1-48, 2003.

[9] Wang, Bing-kun and Huang, Yong-feng and Yang, Wan-xia and Li, Xing, "Short text classification based on strong feature thesaurus," *Journal of Zhejiang University SCIENCE C*, Vol. 13(9), pp. 649-659, 2012.

[10] Kim, Kwanho and Chung, Beom-suk and Choi, Yerim and Lee, Seungjun and Jung, Jae-Yoon and Park, Jonghun, "Language independent semantic kernels for short-text classification," *Expert Systems with Applications*, Vol. 41(2), pp. 735-743, 2012.

[11] Chen, Mengen and Jin, Xiaoming and Shen, Dou, "Short text classification improved by learning multi-granularity topics," *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume*, Vol. 3, pp. 1776-1781, 2011.

[12] Dumais, S. "Combining evidence for effective information filtering," *In AAAI Spring Symposium on Machine Learning and Information Retrieval, Tech Report SS-96-07*, 1996.

[13] M WBerry, S T Dumais, GW O'Brien, "Using Linear Algebra for Intelligent Information Retrieval," *SIAM Review*, December 1995.

[14] Ding C. H., Q. A, "Dual Probabilistic Model for latent semantic indexing information retrieval and filtering," *Proceedings of the 22nd Annual Conference on Research and Development in Information Retrieval (ACMSIGIR)*, 1999.

[15] Deerwester S., Dumais S. T., Furnas G. W., Landaure T. K., and Harshman, R.. "Indexing by Latent Semantics Analysis," *Journal of the American Society for Information Science*, Vol. 41(6), pp. 391-407, 1990,

[16] Landaues T. K., Foltz P. W., Laham D., "An Introduction to Latent Semantic Analysis," *Discourse Processes 25*, pp. 259-284, 1998.

[17] Andreas Hotho, Steffen Staab, Gerd Stumme., "Explaining Text Clustering Results Using Semantic Structures," *In Principles of Data Mining and Knowledge Kiscovery. 7th European Conference*, pp. 217-228, 2003.

[18] Baker L D, McCallum A K., "Distributional clustering of words for text classification," *Proc. ACM -SIGIR-98, Australia: ACM Press*, pp. 96-103, 1998.

[19] Park H, Howland P, Jeon M. "Cluster Structure Preserving Dimension Reduction Based on the Generalized Singular Value Decomposition," *SLAM Journal on Matrix Analysis and Applications*, Vol. 25 (1), pp. 165-179, 2003.

[20] Hofmann T., "Probabilistic latent semantic analysis," *In UAI99, Morgan Kaufmann*, pp. 289-296, 1999.

[21] Hofmann T., "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, Vol. 42, pp. 177-196, 2001.

[22] Chen Ning, Chen An, Zhou Long xiang, "An Incremental Grid Density Based Clustering Algorithm," *Journal of Software*, Vol. 13 (1), pp. 1-7, 2002,.

[23] Zhao, Wayne Xin and Jiang, Jing and Yan, Hongfei and Li, Xiaoming, "Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid," *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 56-65, 2010.

[24] Blei DM, Andrew YN, Michael IJ, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022, 2003.

[25] Minka T, Lafferty J., "Expectation Propagation for the Generative Aspect Model," *Proc of UAl2002*, 2002.

[26] Griffiths TL, Steyvers M., "Finding scientific topics," *The National Academy of Sciences*, Vol. 101, pp. 5228-5235, 2004.

[27] K. Nigam, A. K. McCallum, S. Thrun, T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, Vol. 39(2-3), pp. 102-134, 2000.

[28] Zelikovitz S, Hirsh H, "Improving Short-text Classification Using Unlabeled Background Knowledge to Assess Document Similarity," *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000.

[29] Zelikovitz, Sarah, "Using background knowledge to improve text classification," 2002.

[30] D Song, P D Bruza, Z Huang et al. "Classifying Document Titles Based on Information Inference," *Proceedings of the 14th International Symposium on Methodologies for Intelligent System*, pp. 297-306, Japan, 2003.

[31] Cabrera, Juan Manuel and Escalante, Hugo Jair and Montes-y-mez, Manuel, "Distributional term representations for short-text categorization," *Computational Linguistics and Intelligent Text Processing*, pp. 335-346, 2013.

[32] Sun, Aixin, "Short text classification using very few words," *Proceedings of the 35th international ACM SIGIR*

conference on Research and development in information retrieval, pp. 1145-1146, 2012.

[33] Feng, Xiao and Shen, Yang and Liu, Chengyong and Liang, Wei and Zhang, Shuwu, "Chinese Short Text Classification Based on Domain Knowledge," *International Joint Conference on Natural Language Processing*, pp. 859–863, 2013.

[34] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, Murat Demirbas, "Short Text Classification in Twitter to Improve Information Filtering," *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, Jul. 2010

[35] Huang Wbnlian, Li Shijian, li Jiuxinl, Xu Congfu, "c," *Jorunal of Beijing University of posts and telecommunications*, Vol. 31(3), 2008.

[36] José María Gómez Hidalgo, Guillermo Cajigas Bringas, Enrique Puertas Sánz, "Content Based SMS Spam Filtering," *DocEng'06*, Amsterdam, The Netherlands, 2006.

[37] WEI-WEI DENG, HONG PENG, "Research on a naïve Bayesian based short message filtering system," *Proceedings of the Fifth International Conference on Machine Learning and Cybernetics,* Dalian, pp. 13-16 August 2006.

[38] Han Jiawei, Kamber M. "DATA MINING-Concepts and Techniques," 2001.

[39] S. Dubussion, E. Devoine. M. Masson, "A solution for facial expression representation and recognition," *Signal Processing Image Communication*, Vol. 17 (9), pp. 657-673, 2002.

[40] Xiaojun Quan, Gang Liu et al., "Short text similarity based on probabilistictopics." *Knowl Inf Syst*, pp. 473-491, 2009.

[41] Rafeeque P C, Sendhilkumar S. "A survey on Short text analysis in Web," *Advanced Computing (ICoAC), 2011 Third International Conference on. IEEE*, pp. 365-371, 2011.

[42] Li H, "Text Classification Retrieval Based on Complex Network and ICA Algorithm," *Journal of Multimedia,* Vol. 8(4), pp. 372-378, 2013.

[43] Hu J, Li J, Zeng Z. "SWSCF: a semantic-based Web service composition framework," *Journal of Networks*, Vol. 4(4), pp. 290-297, 2009.

[44] Jun H Y, Xin J J, You C H. "Chinese Short-Text Classification Based on Topic Model with High-Frequency Feature Expansion." *Journal of Multimedia*, Vol. 8(4), pp. 425-431, 2013.