Postprint: Harmful Text Filtering in Uyghur Web Pages Using n-gram Model and Class-imbalanced SVM

Authors: Ruxianguli · Abudurexiti, Yasin Eziz, Guo Wenqiang

Date: 2018-10-11T00:00:00+00:00

Abstract

With the development of network infrastructure in the Xinjiang region, a substantial volume of Uyghur-language web pages has emerged. To foster a healthy online environment, this paper proposes a Uyghur text filtering method that combines an n-gram statistical model with a class-imbalanced Support Vector Machine (SVM) classifier. First, web page texts undergo preprocessing, wherein word stems are initially extracted through an n-gram statistical model; then, semantic analysis is performed on the stems to aggregate those with similar meanings into a single category, thereby reducing stem dimensionality; finally, a parameter controlling the distance between hyperplanes is introduced into the traditional SVM to construct a class-imbalanced SVM, enabling effective classification of Uyghur texts characterized by non-linear inseparability and class imbalance. Experimental results demonstrate that the proposed method can accurately classify harmful texts while exhibiting relatively short classification time.

Full Text

Preamble

Title: Reactionary Text Filtering Method Based on n-gram and Class-Unbalanced SVM for Uyghur Webpages

Authors: Ruxianguli · Abudurexiti[1], Yasen Aizezi[1]†, Guo Wenqiang[2]
[1] Dept. of Information Security Engineering, Xinjiang Police College, Urumqi 830013, China
[2] School of Computer Science & Engineering, Xinjiang University of Finance and Economics, Urumqi 830013, China

Abstract: Along with the construction and development of the network in Xinjiang, a large number of Uyghur webpages have been produced. In order to construct a healthy network environment, this paper proposes a Uyghur text filtering method combining an n-gram statistical model and a class-unbalanced support vector machine (SVM) classifier. Firstly, it preprocesses the webpage text and extracts stems initially using the n-gram statistical model. Then, it carries out semantic analysis of the stems and aggregates stems with similar meanings into one class, thereby reducing the stem dimension. Finally, it introduces a parameter that controls the distance between hyperplanes in the traditional SVM and constructs a class-unbalanced SVM to classify Uyghur texts with nonlinear indivisibility and imbalance. The experimental results show that the method can accurately classify harmful texts and has a shorter classification time.

Keywords: Uyghur webpage; reactionary text filtering; n-gram stem extraction; class-unbalanced SVM

0 Introduction

With the rapid development and popularization of the Internet, a large volume of diverse short texts has emerged, such as web forums, tweets, news feeds, books, and movie summaries. Classifying these short texts is crucial for network information filtering, as identifying and filtering out unhealthy content related to drugs, pornography, and other harmful categories can purify the online environment and help maintain social stability. Short texts are typically unstructured, presented in brief conversational forms consisting of multiple short sentences. Due to their sparse feature vectors and class imbalance characteristics, traditional classification techniques cannot achieve high accuracy for short text classification. An imbalanced dataset refers to one where different classes have unequal sample sizes. A class with many samples is called the "majority class," while a class with few samples is called the "minority class." When classifying imbalanced datasets, classifiers often achieve high accuracy for the majority class but low accuracy for the minority class.

In recent years, with the economic and educational development of Xinjiang, many Uyghur-language websites have emerged. Filtering harmful text information from Uyghur websites is of great significance for stability and healthy development in Xinjiang. Currently, research on Uyghur text classification and filtering methods is limited, primarily conducted by Xinjiang University. For instance, reference [5] uses a purely statistical method that relies solely on word n-grams for classification. Reference [6] applies a supervised method using a maximum entropy classifier to categorize documents into known classes, and an unsupervised learning method to group unlabeled documents, with feature vectors composed of original words and their n-grams. Reference [7] employs a K-nearest neighbor (KNN) classifier combined with three different distance met-

rics (cosine, Euclidean, and Jaccard). Reference [8] proposes using a method for feature extraction and adopts support vector machine (SVM) as the classifier, using statistics to select features—if features and classes are independent, the value is zero. This approach results in high-dimensional feature space because document vectors are sparse, with few relevant features.

Additionally, traditional SVM does not yield good results for imbalanced dataset classification. To address this, scholars have made various improvements. For example, reference [9] uses different loss functions for majority and minority classes in SVM (i.e., using the square of the L2 norm instead of the L1 norm) to penalize misclassification of minority samples. Reference [10] attempts to correct the bias learned by the classifier by introducing correction factors into the support vectors of minority class samples to reduce SVM model bias. Reference [11] proposes fuzzy support vector machine (FSVM) as a learning tool. These techniques require fine-tuning of user-defined parameters and have high complexity. Reference [12] adopts the synthetic minority over-sampling technique (SMOTE), a relatively new method for imbalanced dataset learning that artificially synthesizes minority class samples to increase their proportion and balance sample differences. However, SMOTE requires fine-tuning many user-defined parameters, making it difficult to obtain a suitable parameter set. Reference [13] proposes an improved SVM method called MINSVM, which deletes some majority class samples and provides greater weight to minority class samples to make them more attention-worthy than majority classes, causing the resulting hyperplane to be as close as possible to the majority class. However, its sample deletion operation inevitably affects learning effectiveness to some extent.

Therefore, this paper combines the n-gram statistical model and a class-unbalanced SVM classifier to propose a new framework for classifying short Uyghur texts in webpages. Experimental results demonstrate that the proposed method can effectively classify harmful texts and provides a good foundation for filtering harmful information on webpages.

---

1 Proposed Text Filtering Framework

Uyghur is a script based on Arabic letters with high agglutinative characteristics. The Uyghur alphabet has 32 letters with diverse forms, typically containing four representation forms, resulting in relatively complex morphological changes. Uyghur words consist of stems and affixes, with different affixes added before or after the same stem to express different meanings. These features create certain difficulties for Uyghur text information processing, such as high feature dimensionality.

The proposed framework consists of two parts: text processing and classifier. Figure 1 [Figure 1: see original paper] shows the workflow of the method. First, digital text is saved in UTF-8 format text files. Then, text processing converts

raw text into feature vector representation. Finally, the proposed improved SVM classifier (CUB-SVM) is employed, training the classifier to build a classification model for categorizing test samples.

The main innovations of this method are: (a) Considering the characteristics of Uyghur, this paper adopts the N-gram statistical model for stem extraction and uses a semantic similarity method to further categorize stems, reducing the number and sparsity of text features; (b) To improve classification accuracy for imbalanced data, this paper develops an improved SVM based on traditional SVM.

---

2 Text Preprocessing and Vector Representation

The text processing and vector representation procedure is illustrated in Figure 2 [Figure 2: see original paper].

1) Word Segmentation: Applying traditional techniques to classify short texts produces large and sparse feature vectors. First, the Stanford word segmenter [16] is used to segment raw text, separating prepositions and pronouns from original words and delimiting any punctuation.

2) Stop Word Filtering: The text stop word filtering process involves removing stop words, pronouns, numbers, punctuation, and other non-Uyghur symbols. These components only increase the size of feature vectors without helping to distinguish texts.

3) Stem Extraction: Uyghur is a stem-based language, meaning almost every word is its own root or derived from three-letter or four-letter roots. Words derived from the same root have similar meanings and can therefore be grouped according to their roots. Thus, stems can be considered as features, reducing the length of feature vectors. A statistical method more suitable for the Uyghur environment is adopted to extract stems. The employed statistical method is the n-gram statistical model [17]. Word segmentation is performed at the letter level, where consecutive N letters serve as a gram unit. In the n-gram model, for a harmful letter in the text, its occurrence probability is assumed to be related to the occurrence of the previous N-1 letters. Therefore, the probability of letter sequence occurrence is:

$$P(l_1, l_2, ..., l_N) = \prod_{i=1}^{N} P(l_i | l_{i-1}, ..., l_{i-N+1})$$

The setting of N in the n-gram model needs to be combined with the specific language environment. For Uyghur, since each word is formed by combining multiple letters, a small N cannot effectively represent word attributes, while a larger N (such as 3 or 4) has stronger representativeness.

In the process of using the n-gram statistical model to extract stems, to reduce word dimensionality and redundancy, this paper first deletes the most common affixes in words according to the Uyghur dictionary. Then, it calculates the similarity between two words [18] to extract stems. To demonstrate the process of extracting words using the n-gram statistical model, an example with N=2 is presented, calculating the similarity between two words: (education) and (educational).

First, decompose the words into N=2 letter combinations:

→ , , ,

→ , , , ,

After removing common affixes:

→ , , ,

→ , , ,

The similarity between these two words is then calculated as:

$$S = \frac{2C}{A + B + 2C}$$

where A represents the number of letter combinations present in the first word but not in the second word; B represents the number of letter combinations present in the second word but not in the first word; and C represents the number of identical letter combinations present in both words. If the similarity between two words exceeds a set threshold, the two words are merged into one stem.

4) Semantic Grouping: Stemming helps group words belonging to the same stem, but some words with similar meanings do not share the same stem. Therefore, this paper uses a semantic method to group stems with similar meanings as follows. First, each stem from the dataset is used as a query word, and a synonym dictionary returns synonyms containing that stem. Then, synonyms are extracted from the text source and stored in a list containing stems and their synonyms. Finally, stems in the list are compared. If one stem shares a synonym with another stem, these stems are considered to have similar meanings and are grouped together. If a stem shares a synonym with a stem already in a group, the new stem is added to the existing group. This process performs only one iteration, meaning the resulting groups are not further aggregated. Figure 3 [Figure 3: see original paper] illustrates the semantic grouping process. By the end of this stage, stems with similar meanings are grouped together and can be considered as one feature.

5) Feature Vector Construction: At this stage, feature vectors are constructed for each text in the dataset based on the obtained stem features, and the entire dataset can be presented in tabular form.

3.1 Traditional SVM Classifier

SVM classifier is a machine learning method based on statistics. Its basic idea is to map data into a high-dimensional feature space through a nonlinear mapping and then perform linear regression. Given an input dataset $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^n$ is the input feature vector and y_i is the expected output value, for binary classification problems, the SVM mapping function is represented as $f(x) = w^T\phi(x) + b$, where $\phi(x)$ maps data to high-dimensional space. The values of parameters w and b are obtained by minimizing:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-)$$

subject to the constraints:

$$y_i - w^T\phi(x_i) - b \le \varepsilon + \xi_i^+$$

$$w^T\phi(x_i) + b - y_i \le \varepsilon + \xi_i^-$$

$$\xi_i^+, \xi_i^- \ge 0$$

For convenience, two slack variables ξ_i^+ and ξ_i^- are introduced, transforming the problem into estimating parameter values by minimizing:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-)$$

3.2 Improved Class-Unbalanced SVM (CUB-SVM)

In this work, to enable SVM to effectively handle Uyghur texts with nonlinear inseparability and imbalance, this paper extends the traditional SVM classifier to form the Class-Unbalanced SVM (CUB-SVM).

In addition to integrating the kernel into SVM, this paper introduces a new parameter τ that minimizes the distance between majority data samples and the separating hyperplane while maximizing the distance between minority data samples and the separating hyperplane, as shown in Figure 4 [Figure 4: see original paper].

The objective formulation of CUB-SVM is:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C^+\sum_{i|y_i=+1}\xi_i^+ + C^-\sum_{i|y_i=-1}\xi_i^- + \tau^+\sum_{i|y_i=+1}\xi_i^+ - \tau^-\sum_{i|y_i=-1}\xi_i^-$$

subject to:

$$w^T\phi(x_i) + b \geq 1 - \xi_i^+ \quad \text{for } y_i = +1$$
$$w^T\phi(x_i) + b \leq -1 + \xi_i^- \quad \text{for } y_i = -1$$
$$\xi_i^+, \xi_i^- \geq 0$$

where the subscript "+" represents the majority class, "-" represents the minority class, and τ is a scaling parameter.

The Lagrangian equation for this problem is:

$$\mathcal{L} = \frac{1}{2}\|w\|^2 + C^+ \sum_{i|y_i=+1} \xi_i^+ + C^- \sum_{i|y_i=-1} \xi_i^- + \tau^+ \sum_{i|y_i=+1} \xi_i^+ - \tau^- \sum_{i|y_i=-1} \xi_i^- - \sum_{i|y_i=+1} \lambda_i(1 - \xi_i^+ - w^T\phi(x_i) - b) - \sum_{i|y_i=-1} \mu_i(1$$

where $\lambda_i, \mu_i, \alpha_i, \beta_i, \gamma_i, \delta_i$ are Lagrange multipliers. By finding the KKT conditions and substituting them into the Lagrangian function, we obtain a dual problem:

$$\max_{\lambda,\mu} \sum_{i|y_i=+1} \lambda_i + \sum_{i|y_i=-1} \mu_i - \frac{1}{2} \sum_{i,j|y_i=y_j=+1} \lambda_i\lambda_j y_i y_j K(x_i, x_j) - \frac{1}{2} \sum_{i,j|y_i=y_j=-1} \mu_i\mu_j y_i y_j K(x_i, x_j) + \sum_{i,j|y_i=+1,y_j=-1} \lambda_i\mu_j y_i$$

subject to:

$$\sum_{i|y_i=+1} \lambda_i y_i + \sum_{i|y_i=-1} \mu_i y_i = 0$$
$$0 \leq \lambda_i \leq C^+$$
$$\mu_i \geq D^-$$

where $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ is the kernel function. After solving this problem, we can find the separating hyperplane, where:

$$w = \sum_{i|y_i=+1} \lambda_i y_i \phi(x_i) + \sum_{i|y_i=-1} \mu_i y_i \phi(x_i)$$

and the classifier becomes:

$$f(x) = \text{sign}\left(\sum_{i|y_i=+1} \lambda_i y_i K(x_i, x) - \sum_{i|y_i=-1} \mu_i y_i K(x_i, x) + b \right)$$

4 Experiments and Analysis

4.1 Experimental Setup

To test various harmful topics in web text, 500 texts were collected from various Uyghur website forums, divided into 4 categories: (1) drug-related texts (143 samples); (2) pornographic texts (78 samples); (3) gambling texts (107 samples); and (4) normal texts (172 samples). These texts and their categories are imbalanced. The character length statistics of the text set are shown in Table 1 .

The proposed method was implemented on a PC with an Intel Core i5 5250@2.7GHz CPU, 24 GB memory, and Windows 7 64-bit, using MATLAB R2013b and the CVX 2.1 toolbox.

4.2 Performance Metrics

Five-fold cross-validation was performed on the dataset. To measure classifier performance, three metrics were used, where TP represents correctly classified positive samples, FP represents misclassified positive samples, TN represents correctly classified negative samples, and FN represents misclassified negative samples.

a) Accuracy measures the classifier' s correct classification performance:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

b) F-measure evaluates the classifier' s overall performance, calculated from Precision and Recall:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TN}{TN + FN}$$

$$F\text{-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

c) AUC Area: The Receiver Operating Characteristic (ROC) curve plots the true positive rate against the false positive rate. The area between the curve and the X-axis is the AUC area, ranging from 0.5 to 1, reflecting classifier effectiveness. Larger values indicate better classification performance.

4.3 Classification Results Analysis

First, this paper analyzes the effect of the preprocessing stage. In traditional text processing methods, words are directly used as features without filtering and stemming. For the Uyghur text dataset used in this paper, the traditional method produces a feature vector of length 6204. After stemming and semantic

grouping, the proposed method obtains a feature vector of length 1163, reducing it by nearly 5.3 times.

To evaluate the performance of CUB-SVM, it is compared with standard SVM, MINSVM proposed in reference [13], and SMOTE-SVM proposed in reference [12] to highlight differences between various improved classification methods. Table 2 presents the average performance of various methods on the dataset. Figure 3 [Figure 3: see original paper] shows the ROC curves of various methods.

The results demonstrate that the CUB-SVM classifier outperforms traditional SVM and some improved SVM classifiers, with F-measure improved by 16% compared to traditional SVM. This is because the preprocessing process in the proposed method can effectively reduce data dimensionality, and CUB-SVM can well classify imbalanced short texts. For SMOTE and MINSVM methods, data resampling techniques do not improve SVM classifier performance because changing data distribution may introduce more outliers that degrade SVM performance. Additionally, SVM with different cost functions does improve classifier precision, but this improvement comes at the cost of reduced recall, resulting in lower F-measure values and overall accuracy.

Table 3 presents the classification time of various classifiers. For fair comparison, the proposed filtering and stemming steps are used in preprocessing for all Uyghur text data. SMOTE-SVM has the highest running time because it requires oversampling of data, which is time-consuming and generates more data. MINSVM has the lowest processing time because it randomly deletes some data samples, resulting in a smaller dataset and thus shorter processing time. The CUB-SVM classifier' s processing time is slightly longer than standard SVM but does not significantly increase overhead.

4.4 Statistical Analysis

Statistical analysis is used to test the significance of accuracy differences between classifiers. Given two classifiers, a statistical test determines whether they have the same expected error rate. To perform statistical analysis, K-fold cross-validation experiments are employed.

In K-fold cross-validation, K training/test sets are obtained from the original dataset. A classifier is trained on training set D^i_{train} and tested on test set D^i_{test}. The error rates on training and test sets are denoted as p^i_{train} and p^i_{test}, respectively. If classifiers have the same error rate, they should have the same mean, meaning the difference between their means should equal 0. For the K-fold cross-validation test, the difference in the i-th error rate is $p^i = p^i_1 - p^i_2$, yielding a distribution of K points. Assuming p^i_1 and p^i_2 are normally distributed, their difference is also normally distributed.

Let μ be the mean of this distribution. Under the null hypothesis $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$, there is a statistic that follows a t-distribution with $K - 1$ degrees of freedom:

$$t = \frac{\bar{p}}{S/\sqrt{K}}$$

where $\bar{p} = \frac{1}{K}\sum_{i=1}^{K} p^i$ and $S^2 = \frac{1}{K-1}\sum_{i=1}^{K}(p^i - \bar{p})^2$. If this value falls outside the range $(-t_{\alpha/2,K-1}, t_{\alpha/2,K-1})$, the test rejects the hypothesis at significance level α. When $\alpha = 0.1$ and $K = 5$, the confidence level is 90% and the range is (-2.132, 2.132).

The statistical analysis results of test error rates and overall error rates are shown in Table 4 . Due to dataset imbalance, some categories have small sample sizes. The statistical results show that all tests reject the hypothesis for minority categories. For majority categories, the hypothesis is rejected on two datasets and accepted on three others, meaning error rates differ on two datasets but not on the other three. For overall error rates, all data accept the hypothesis, indicating no difference in overall error rates. In summary, the CUB-SVM classifier achieves good accuracy on minority categories without sacrificing overall accuracy, demonstrating that it can effectively handle imbalanced data.

---

5 Conclusion

This paper proposes a method for filtering texts on Uyghur websites, employing harmful text preprocessing and stemming steps to reduce text feature dimensionality and vectorize the text. To better classify imbalanced texts, an improved SVM classifier (CUB-SVM) is proposed to achieve high-precision classification of Uyghur texts for harmful text filtering. Experimental results show that the proposed method can accurately classify harmful webpage texts and can be applied to the management and purification of Uyghur webpages.

References

[1] Yu Bengong, Zhang Lianbin. Chinese short text classification based on CP-CNN [J]. Application Research of Computers, 2018, 35 (4): 1001-1004.

[2] Huang Faliang, Feng Shi, Wang Daling, et al. Mining topic sentiment in microblogging based on multi-feature fusion [J]. Chinese Journal of Computers, 2017, 40 (4): 872-888.

[3] Gu Xiaoqing, Jiang Yizhang, Wang Shitong. Zero-order TSK-type fuzzy system for imbalanced data classification [J]. Acta Automatica Sinica, 2017, 43 (10): 1773-1787.

[4] Bhowan U, Mark Johnston, Zhang Mengjie, et al. Evolving diverse ensembles using genetic programming for classification with unbalanced data [J]. IEEE Trans on Evolutionary Computation, 2013, 17 (3): 368-386.

[5] Maimaitiyiming Hasimu, Wushouer Silamu, Weinila Mushajiang, et al. Research on N-gram based Uyghur text classification technique [J]. Application Research of Computers, 2015, 32 (7): 1986-1988.

[6] Turdi Tohti, Akbar Pattar, Askar Hamdulla. Semantics-based feature extraction and its application in Uyghur text classification [J]. Journal of Chinese Information Processing, 2014, 28 (6): 140-144.

[7] Turdi Tohti, Ahmatjan Ablat, Muyassar Aniwar, et al. Combined algorithm of GAAC and K-means for Uyghur text clustering [J]. Computer Engineering and Science, 2013, 35 (7): 149-155.

[8] LI Xiang, Tuergen Yibulayin, Kahaerjiang Abiderexiti, et al. Emotion analysis of active learning based on SVM in Uyghur language [J]. Journal of Xinjiang University (Natural Science Edition), 2015, 32 (4): 447-452.

[9] Pal M, Mather P M. Support vector machines for classification in remote sensing [J]. International Journal of Remote Sensing, 2015, 26 (5): 1007-1011.

[10] Yuan Yubo, Fan Weiguo, Pu Dongmei. Spline function smooth support vector machine for classification [J]. Journal of Industrial & Management Optimization, 2017, 3 (3): 529-542.

[11] Ma Hongyan, Wang Liling, Shen Bo. A new fuzzy support vector machines for class imbalance learning [C]// Proc of International Conference on Electrical and Control Engineering. Piscataway, NJ: IEEE, 2011: 2871-2874.

[12] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16 (1): 321-357.

[13] Ajeeb N, Nayal A, Awad M. Minority SVM for linearly separable imbalanced datasets [C]// Proc of International Joint Conference on Neural Networks. Piscataway, NJ: IEEE, 2014: 1-5.

[14] Abdusalam Dawut, Hussein Yusuf, Askar Handulla. Emotion recognition from Uyghur sentences based on combinations of class discrimination words and a sentiment dictionary [J]. Journal of Tsinghua University: Science and Technology, 2017, 57 (2): 197-201.

[15] Han Junbing, Halidan · Abudureyimu, Gulnur Arken, et al. Improved information gain algorithm based on Uyghur feature selection [J]. Computer Engineering and Applications, 2017, 53 (23): 34-38.

[16] Sun Rong, Zhou Wen, Liu Zongtian. Using language rules to improve the performance of word segmentation [C]// Proc of International Congress on Image and Signal Processing. Piscataway, NJ: IEEE, 2014: 1665-1669.

[17] Yu Jie. Research on definition extraction based on Spark and DN-gram model [J]. Journal of Beijing Information Science & Technology University, 2017, 32 (4): 64-68.

[18] Dong Yangyi, Li Weihua, Yu Hui. Hierarchical relation mining of Chinese text based on mixed cosine similarity [J]. Application Research of Computers, 2017, 34 (5): 1406-1409.

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