ISSN (Print) : 2319-8613
ISSN (Online) : 0975-4024

Supervised Algorithm Based Automatic Bill Classification and Prediction

Anjana K.P [#1], Padmavathi S*[2]
Department of Computer Science and Engineering, Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidhyapeetham, Tamil Nadu, India
[1] anjanakalesh90@gmail.com
[2] s_padmavathi@cb.amrita.edu

***Abstract*— Text classification is one of the major research areas in the field of text mining. It is the process of automatically classifying text documents in to predefined categories. Our objective is to classify the collection of bill documents in to predefined categories based on the bill contents. Bills will be either in the form of electronic format or printed documents. In this paper, Tesseract Optical character recognition (OCR) tool is used for converting bills in to text format. Then, Feature vector representation is done using Bag of Words and Term Frequency-Inverse Document Frequency (TF-IDF) methods. We compared different supervised classification algorithms that have been used in the text classification and suitable algorithm for bill classification is suggested based on their performance. All the algorithms are evaluated using standard evaluation metrics.**

Keyword - Text Categorization, OCR, TF-IDF, Bag of Words, Supervised Classification

I. INTRODUCTION

The amount of digital data is growing exponentially day by day. The data may be in structured or unstructured in nature. Text mining includes the process of extraction of knowledge from unstructured text data [1]. Text categorization is the process of automatically assigning the appropriate category to each document. The text documents may be in the form of email, news, business documents, bills, research papers etc. Each document may belong to multiple, exactly one or no category at all. The main objective is to classify the collection of bills into predefined categories based on the contents of bills. Categories of bills include Computer and electronics purchase bills, Hotel-accommodation bills, Hotel-food bills, Medical bills, Transport bills etc. The users have to provide only the images of bills as input to the system. In this paper we used Tesseract optical character recognition (OCR) tool to extract text from images. Then we have applied text classification techniques on those text data. Fig. 1 shows the sample image of a transport bill document.

Fig. 1. Sample image of Transport bill

Text classification is broadly classified as supervised document classification and unsupervised document classification. In supervised classification the documents are already labelled, then the system will classify the documents in which category they belongs and then find the accuracy of the system. On the contrary, unsupervised document classification is completely done without any human interference. Here we are focusing on supervised text classification algorithms.

Text classification algorithms are normally applied on feature vectors of text data. So, before doing classification we need to convert the text dataset in to feature vector representation. Representing text document in feature vector form is normally called as vector space model (VSM). Bag of Words (BoW) and Term frequency-Inverse document frequency (TF-IDF) are used for feature vector representation of this dataset. The bag-of-words model is a common and simple representation used for document classification. In this model, the presence or the number of occurrence of each word is considered as a feature for a classifier. For bill classification the presence or the number of occurrences of each word gives relevant information to distinguish between different categories. TF-IDF is a weighting scheme which finds the importance of a word to a document in a document collection. TF-IDF reduces the weight of terms that occur repeatedly in the document set. For bill classification, supervised algorithms such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Multinomial Naive Bayes and Bernoulli Naive Bayes Classifiers are used. All these algorithms have significant applications in many fields. In this paper bill classification is done using all these algorithms and the suitable algorithm for bill classification is suggested based on their performance. Their Performances are listed in chapter 4 and conclusive remarks are given in chapter 5.

II. RELATED WORK

The Main goal of text classification is to discover a category assignment function $f: d \times c \rightarrow \{0,1\}$, where c is the predefined classes and d is the set of all text documents. The value of f is 1 if the document d belongs to category c. otherwise it is 0. Many researchers are working in this area. Bruno Trstenjak et. Al [2] proposed a method of KNN algorithm with TF-IDF for document classification. The weight matrix is formed by taking the relations between each unique words and documents. TF-IDF method finds the relative occurrence of words in a specific document using inverse proportion of the word over the document set. In KNN it is required to determine K value. K value indicates required number of documents from the collection which is nearest to the selected document. Tested with 500 online documents and showed good result but it is sensitive to the type of documents. Also the time required for data processing is increased with the increasing amounts of data. An improved classification based on predictive association rules (CPAR) is proposed in [3]. It joins the advantages of both associative classification and traditional rule-based classification. Association rule classification includes rule generation, rule selection and classification. Features of improved CPAR are Class weighting adjustment and Post processing with SVM. Class weighting adjustment adjusts the weight and controls the classification method of each class. After classification with the rules, there may have some test documents satisfying no rule. Post-processing with SVM is used for the purpose of classification of such documents. Dataset used is Chinese text classification corpus.

In [4] they compared text classification using both SVM and artificial neural network. They experimented with Reuters News Data Sets. It shows SVM has good performance on large data sets. An improved document classification through enhanced Naive bays algorithm is proposed in [5]. They also implemented 2 sub classification algorithms named hierarchical sub classification and sub categorization using document similarity method. To reduce the computational complexity of text classification [6] proposed a hybrid algorithm which uses KNN with Principal Component Analysis (PCA).

III. PROPOSED WORK

In this paper, bill classification is done using various supervised machine learning algorithms. K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Multinomial Naive Bayes and Bernoulli Naive Bayes Classifiers are used for bill classification. Even if all these algorithms give significant success rate, it purely depends on the dataset in which these algorithms are applied. So for finding the most suitable algorithm for bill classification all these algorithms are applied in the bill dataset and their performance is calculated using standard evaluation metrics such as accuracy, precision, recall and F measure.

Fig. 2 shows the architecture diagram for bill classification. There are five main steps in bill classification. Initially images of bill documents are given to the system as input for the classification. Then Optical Character Recognition is done using tesseract tool. It is one of the most precise open source OCR engines currently available and is maintained by Google. This tool takes the images of the bill documents as input and extracts the textual content from the bill for further processing.

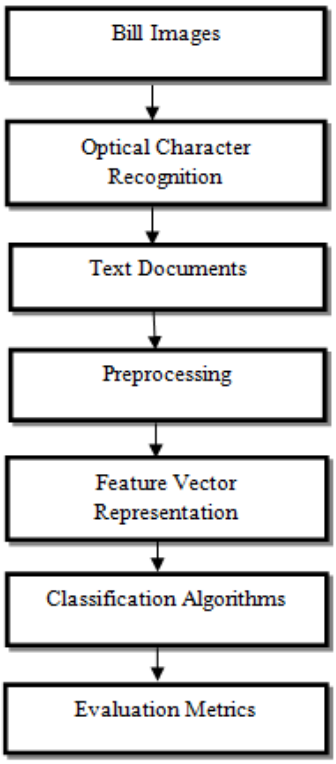

Fig. 2. Architecture Diagram for bill classification

Before applying classification algorithms we need to perform pre processing on text documents and then transform text data into feature vector representation. Pre processing includes tokenization, removal of digits and special characters and n-gram construction. Tokenization is the process of dividing a text into individual elements that take as an input for various algorithms. In the n-gram model, a token can be defined as a collection of n items. Unigram is the simplest model in which each word consists of exactly one word, letter, or symbol. Here we tested the classifier with 1-gram and 2-gram model. Other pre processing tasks include stop words removal and stemming. In bill documents stop words removal and stemming is not required. Because almost all words in bill documents should be in its root form and the presence of stop words are very less in bill documents. Hence we can reduce the effort and time for these pre processing steps. Next step is the conversion of text data to feature vector representation.

A. Feature Vector Representation

In this paper for doing feature vector representation two methods are implemented. They are Bag of Words and Term frequency Inverse Document frequency (TF-IDF).

1) *Bag of Words (BoW):* It is a generally used vector representation method in Natural Language Processing. In this method the textual contents of bill documents are taken as input. Bag of Words method considers the bill document as a collection of words. First it creates a vocabulary of all words that occur in the training set. The number of occurrences, that is, the frequency of that particular word in the particular bill content is found. Hence the whole content of bill is represented as a set of words associated with the count of occurrences. Consider an example of BoW representation of text documents given below. Let D_1 and D_2 be the 2 bill documents.

D_1 (Transport bill) : "Bus arrival time ticket charge amount"

D_2 (Hotel-food bill) : "Hotel rice chapatti curry juice amount "

The vocabulary of words can be written as

V={bus:1, arrival:1, time:1, ticket:1, charge:1, amount:2, hotel:1, rice:1, chapatti:1, curry:1, juice:1}

Bag of words representation of the 2 documents D_1 and D_2 is shown below.

D_1 (Transport bill) = {1 1 1 1 1 1 0 0 0 0 0 }

D_2 (Hotel-food bill) = {0 0 0 0 0 1 1 1 1 1 1 }

$\sum$ = [1 1 1 1 1 2 1 1 1 1 1]

2) *Term Frequency-Inverse Document Frequency (TF-IDF):* This is a traditional method for finding term weight calculation of each word in vector space model.TF-IDF; TF reflects the distribution of terms within the text. IDF calculates the distribution of words in the whole text set [7]. This method helps us to eliminate high frequency and low discrimination words from the set because the terms with high frequency may not contain any useful data for classification. More than that, it can mislead us from the correct classification also. For example, consider the word "only". The "only" word will be having high frequency as it occurs many times in a bill data. But this word really does not have any relevant information for classification. These kind of terms are called bogus terms. So this method is considered as effective one for term weight calculation. TF-IDF weight of a word can be calculated using the given equation [8].

$$\mathrm{TF-IDF(w,d)} = \mathrm{TF(w,d)} \times \log(\mathrm{N/DF(w)})$$

Where $TF(w,d)is$ the frequency of word w in document d, N is the number of documents, and $\mathrm{DF}(w)$ is the number of documents that containing word w. Next step is classification of bill documents using various algorithms. Classification has two main phases: Training and testing, which is done using a set of labeled documents. In the training phase, a collection of labeled documents are given to the system. That is, it considered to belong to one of the predetermined classes. Using these labeled documents, the system will model a distribution. In the second phase, that is testing, a set of documents is given as input. The system will assign each of the documents to a certain predefined category according to the distribution model built in the training phase and then the actual label of test document is compared with the classified result to find the accuracy of the classifier.

B. Classification Algorithms

In this paper, supervised algorithms such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Multinomial Naive Bayes and Bernoulli Naïve Bayes classifiers are tested with bill dataset.

1) *K-Nearest Neighbor (KNN):* KNN is a simple and famous algorithm for text classification. To classify a document, the algorithm searches for the K nearest neighbors of the document. In this algorithm, when an unlabeled document comes for classification, the distance between the document and the neighbors are calculated using any distance formula. Here we used Euclidean distance formula. The new document is assigned to particular class according to k-value. If k=1, it assign the document to the class where the distance is minimum [9]. This implementation of this method is simple but is slow in performance as it required to comparing the test document with all training documents [10]. KNN is a multi-class classifier. That is, using KNN we can perform classification for more than two classes. In automatic bill classification we have 5 predefined classes. Consider an example of predicting the category of a test bill document. If k=5, it searches for the 5 nearest neighbors of the document using distance measures with minimum distance. It assigns the category of a test document to the maximum category of nearest 5 documents.

2) *Support Vector Machines (SVM):* SVM method is a popular and more powerful supervised method for text classification. Here the data plane is linearly separated using a hyper plane with maximum marginal distance. Hyper plane separates positive examples from negative examples. Equation given below shows the formula for hyper plane representation [11].

$$\mathrm{w^T x + b = 0}$$

Where b is an intercept term, w is a decision hyper plane normal vector and x is a data. The decision for classification of a new document is determined using below equation [11].

$$\mathrm{f(\vec{x}) = sign\left((\vec{w}^T \vec{x}) + b\right)}$$

Where w is a weight vector, b is an intercept term and value of f(x) can be positive or negative. A value of -1 indicates one class, +1 indicates other class. Automatic bill classification required multi-class classification. Since it contains 5 predefined categories such as Computer and electronics purchase bills, hotel accommodation bills, hotel food bills, transport bills and medical bills. The most common method for multi-class classification with SVM is to build one-versus-rest (ovr) classifiers and select the category which classifies the test document with greatest margin.

3) *Naive Bayes Classifier*: The Naive Bayes Classifier is a statistical Classifier method. It is based on Bayesian theorem which predicts class membership probabilities. The limitation of this classifier is, it assumes every feature word is independent from one other. The probability $\mathrm{P}(c|d)$ that the document d belongs to the class c can be calculated by using Bayes formula as given in equation (12).

$$P(c|d) = \frac{P(d|c) \times P(c)}{P(d)}$$

Where d is a document and c is a class, P(d|c) is likelihood and P(c) is the probability of having class c. The class label for a new document can be predicted using the below equation (12).

$$\text{predicted class label} \leftarrow \text{argmax P(c|d)}$$

For automatic bill classification two variations of naïve bayes classifier such as multinomial naive bayes and multivariate Bernoulli naive bayes are tested. In multivariate Bernoulli Event model the vector corresponds to each word is either 1(if word is present) or 0(otherwise). It finds the fraction of documents of category c_j in which word w appears. For example, in a bill document if the word 'juice' occurs 3 times then the multivariate Bernoulli event model represent it as 1(present). In Multinomial Event model the vector of each word represents the number of occurrence of that word in the document. It finds the fraction of times a word w appears in documents of category c_j. For the same example ie, for the word 'juice' multinomial event model represent the word vector as 3.ie, number of occurrence. For multi-class classification here we construct a separate binary classifier trained on positive samples from one particular class and negative samples from all other classes. Then for a test document d, it run all the classifiers and chose the label with the highest score.

IV. EXPERIMENT AND ANALYSIS

Classification is done using all the algorithms that is discussed earlier in this paper. For automatic bill classification we created our own dataset, which includes 250 bill documents (5 categories with 50 bills each) for training and 150 bill documents(5 categories with 30 bills each) for testing. The output and the performance of the various classifiers tested are depicted below. Table shows the accuracy, precision, recall and F-score of supervised classification algorithms we tested with bill dataset. Precision can be defined as the part of retrieved instances that are relevant. Precision can be calculated using the given equation [13].

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

Where TP is true positives and FP is false positives. Recall can be defined as the part of relevant instances that are retrieved. Recall can be calculated using the given equation [13].

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

Where FN is false negatives. The F-measure combines Precision and Recall into a single metric, and can be calculated using given equation [13].

$$\text{F} = \frac{2\text{PR}}{\text{P} + \text{R}}$$

Where P represents precision and R represents recall.

Here, various classification methods are tested using both bag of words (BoW) and Term Frequency Inverse Document frequency (TF-IDF). SVM shows an accuracy of 92.2% and it is the highest among all other methods. Bernoulli naive bayes with TF-IDF is also having good accuracy comparable to other methods.

TABLE I. Performance Measures of Different Classifiers

Sl No	Methods	Features	Accuracy (%)	Precision	Recall	F Measure
1	SVM	BOW	92.2	0.94	0.92	0.92
2	SVM	TFIDF	92.2	0.92	0.91	0.92
3	K Nearest Neighbor	BOW	87.2	0.90	0.85	0.87
4	K Nearest Neighbor	TFIDF	91.6	0.91	0.92	0.92
5	Multinomial Naive bayes	BOW	89.4	0.92	0.89	0.90
6	Multinomial Naive bayes	TFIDF	89.6	0.94	0.89	0.90
7	Bernoulli Naive bayes	BOW	91.8	0.92	0.91	0.90
8	Bernoulli Naive bayes	TFIDF	91.8	0.94	0.91	0.91

In the above table SVM with BoW and SVM with TFIDF show high accuracy, precision and recall value. High precision indicates that the classification algorithm returned considerably more relevant results than irrelevant, while high recall specifies that the algorithm returned most of the relevant results.

TABLE II. Training and Testing Time Required for Various Classifiers

Sl No	Methods	Features	Training Time(s)	Testing Time(s)
1	SVM	BOW	0.016	0.005
2	SVM	TFIDF	0.029	0.006
3	K Nearest Neighbor	BOW	0.012	0.011
4	K Nearest Neighbor	TFIDF	0.021	0.015
5	Multinomial Naive bayes	BOW	0.015	0.005
6	Multinomial Naive bayes	TFIDF	0.016	0.007
7	Bernoulli Naive bayes	BOW	0.013	0.005
8	Bernoulli Naive bayes	TFIDF	0.014	0.06

Table II shows the training time and testing time required for each of the algorithms. Testing time required for KNN is high compared to other classifiers and KNN with BoW shows comparatively less performance. Fig. 3 shows the accuracy graph of different classifiers tested. From this graph we can conclude that SVM shows high performance than other classifiers tested for bill dataset.

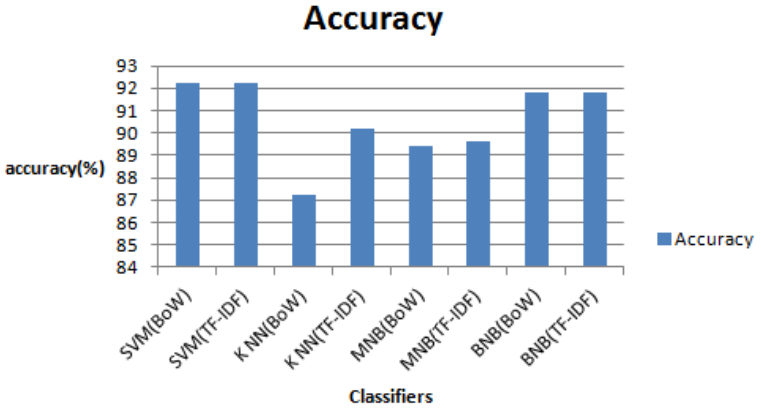

Fig. 3. Accuracy of Different Classifiers

Fig. 4 shows the result of prediction of category of a given bill document using SVM classifier. It is predicted as transport bills with a probability of 98.16 %.

```
Probability of prediction for bill 1
0.981636381221
'Travels India limited Travel in tours and travels Bus stand Kodaikanal Boarding
 pass Date Time Seat No inclusive of Service tax Happy journey in our Volvo Ac S
emi Sleeper Air Suspension buses Name Ardra Address talassery' => transport_bill
s
```

Fig. 4. Category prediction of a bill with its probability

Fig. 5 shows the graph of probability distribution of a test bill document in 5 categories. In this graph computer and electronics purchase bills, hotel food bills and medical bills shows nearly zero percent probability. But hotel accommodation bill has probability of less than 10% because of some common words present in both hotel accommodation bills and transport bills.

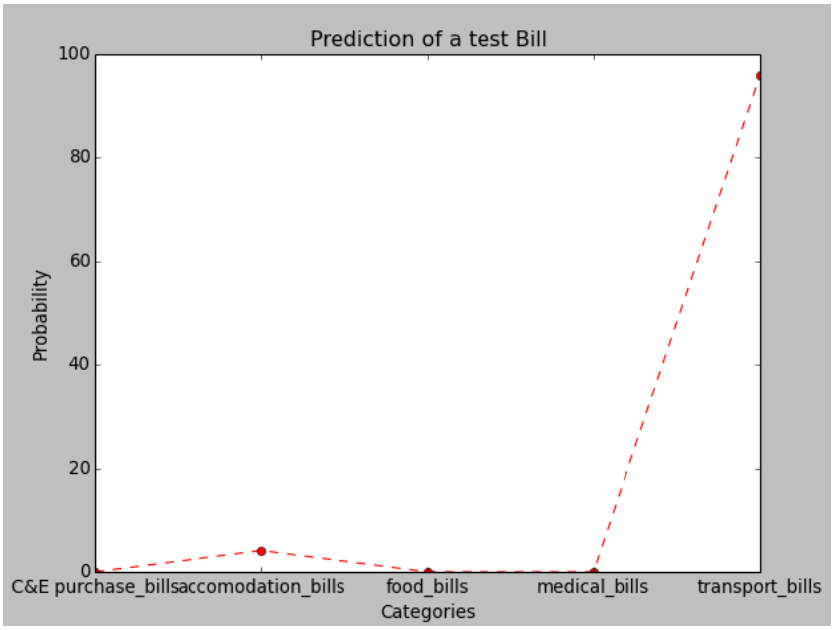

Fig. 5. Probability of Category Prediction of Bill Document in 5 Categories

V. CONCLUSION

Text categorization is one of the most important research areas in the field of text mining. To classify the collection of bill documents in to predefined categories based on the contents of bill we used SVM, KNN, and Naive Bayes classifier. The performance of these algorithms varies according to the dataset. From the experiment that carried out with the bill document dataset, it is found that SVM gives the best performance with an accuracy of 92.2 %. So SVM is the best suited algorithm for this bill document dataset.

REFERENCES

[1] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, Elsevier, 2000.
[2] Trstenjak, Bruno, Sasa Mikac, and Dzenana Donko. "KNN with TF-IDF based Framework for Text Categorization." Elsevier, Procedia Engineering 69 (2014): 1356-1364.
[3] Hao, Zhixin, et al. "Improved classification based on predictive association rules." Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on. IEEE, 2009.
[4] Basu, Atreya, Christine Walters, and M. Shepherd. "Support vector machines for text categorization." System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on. IEEE, 2003.
[5] Sathyadevan, Shiju, P. R. Sarath, U. Athira, and V. Anjana. "Improved document classification through enhanced Naive Bayes algorithm." In Data Science & Engineering (ICDSE), 2014 International Conference on, pp. 100-104. IEEE, 2014.
[6] Nedungadi, Prema, Haripriya Harikumar, and Maneesha Ramesh. "A high performance hybrid algorithm for text classification." In Applications of Digital Information and Web Technologies (ICADIWT), 2014 Fifth International Conference on the, pp. 118-123. IEEE, 2014.
[7] Jie, G., & Li-chao, C. (2010, December). Research of improved IF-IDF Weighting algorithm. In Information Science and Engineering (ICISE), 2010 2nd International Conference on (pp. 2304-2307). IEEE.
[8] "Text Categorization". Slideshare.net. N.p., 2015. [Online]. Available: http://www.slideshare.net/kanimozhiu/text-datamining-txtcat
[9] Gupta, V. (2011). Recent trends in text classification techniques. International Journal of Computer Applications, vol.35, No.6.
[10] Jindala, Rajni, and Shweta Tanejab. "Text Categorization–A Review" ,Elsevier, 2009.
[11] Nlp..stanford.edu.N.p.,2008.[Online].Available:http://nlp.stanford.edu/IR-book/html/htmledition/support-vector-machines-the-linearly-separable-case-1.html
[12] Raschka, S. (2014). Naive Bayes and Text Classification I-Introduction and Theory. arXiv preprint arXiv:1410.5329.
[13] Goutte, Cyril, and Eric Gaussier. "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation." In European Conference on Information Retrieval, pp. 345-359. Springer Berlin Heidelberg, 2005.

AUTHOR PROFILE

Anjana K.P has done her Master in Computer Application from ITEC Nileshwar and MTech in Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Coimbatore. Her current research interest includes Text mining and Image Analysis and Pattern Recognition.

Dr. S Padmavathi, Assistant Professor works at Amrita School of Engineering since 2001. She has done her ME in Computer Science Engineering from Government College of Technology, Coimbatore. She did her PhD in Amrita Vishwa Vidyapeetham, Coimbatore. Her research interests are allied with Digital Image Processing, Video processing, Image Analysis and Pattern recognition.