An Integrated Approach for Personality Analysis using OCR and Text Mining

Pavithree B. Shetty, Sanath R. Kashyap, Sneha V. Kamath, Supraja, Bharath Bhushan S. N*.

Department of Computer Science and Engineering, Sahyadri College of Engineering & Management, Mangaluru-575007

*Email: bharathbhushan.cs@sahyadri.edu.in

ABSTRACT

In this paper we aim at recognizing handwritten text from an image and convert that into an editable document and perform personality analysis. It is a challenging task because each individual's handwriting is unique. As a person's mind-set plays a major role when he pens down the content, in our work we use the content written by the person to classify contextually if the person is positive minded or negative minded. In order to convert handwritten characters into an editable document we have used optical character recognition (OCR) and machine learning techniques. Personality analysis is done based on the frequency of occurrences of words.

Keywords: Optical Character Recognition, Text Mining, Machine Learning

1. INTRODUCTION

From past few years there is a lot of data which is being generated which demands automated management without much human intervention. This paper mainly focuses on offline handwritten character recognition of English words by initially identifying individual characters. Today due to the advent of internet and its wider reach to public there are millions of text data which is being generated every day and this requires data management.

Optical character recognition (OCR) is the electronic or mechanical conversion of handwritten, typed or printed text images into machine-encoded text. This involves scanning of the text character by character and then the character image is translated into character codes, which is frequently used in data processing.

Given a scanned image of handwritten text, we aim to extract the text in that image using OCR algorithm and display it in a editable document along with identifying if the content has been written by a positive minded or negative minded person by identifying and understanding what has been written. Using text mining techniques we achieve the later part of the problem statement.

In the field of business, management, education, various billing systems and various ticket reservation systems lot of data is being generated but this does not provide us any useful conclusions. When this data is analysed we get valuable information from which we can get the current interests of people and improve the present business trends.

Text mining techniques can be used to derive useful insights from the wide range of data. Text mining is analysing the data which is contained in natural language text. From editable document using machine learning and text mining techniques we infer if a person is contextually positive minded or negative minded.

The aim of the text classification algorithm is to determine if the person is positive minded or negative minded based on the content given by the user. Suppose we have two classes of documents i.e., class1->content given by positive thinkers and class 2- content given by negative thinkers. We have to train our model with these documents. The text classification assigns a Boolean value to each pair Φ (Qd,S)=1 where, S is a set of predefined categories and Qd is the domain of documents. The task is to approximate the true function Φ= (Qd,S)->{0,1} (where its 1 if classifier classifies the document properly as positive minded or negative minded else it is 0) using a function^ ^ Φ= (Qd,S)->{0,1} such that the values of Φ and Φ has approximately similar values.

Following section of the paper is organized as follows. Section 2 gives a brief literature survey on the text classification and representation. Section 3 represents the proposed model for conversion of handwritten characters to editable format and in the later part we use text classifier to classify if a person is positive minded or negative minded using text mining approaches. Experimentation and comparative analysis performed on the proposed models will be discussed in the section 4. Finally we conclude the paper in section 5.

2. LITERATURE SURVEY

In [1], authors highlight the main techniques and methods used in text document classification. It emphasises the representation of text and machine learning techniques. The methods and theories of text mining and document classification is analysed in the paper. [2] This research article contains a B-Tree based classification methodology which is adapted for classification. The proposed compressed representation and B-Tree methodologies are verified on the publicly available large corpus to validate the effectiveness of the proposed models. [3] In this paper, a learning model of text classification for support vector machine (SVM) is evolved. It creates a bridge between the characteristics of text classification task and the generalisation performance of a SVM in a quantifiable manner. [4] Major problem such as handling large number of attributes, dealing with the unstructured text, and choosing a machine learning technique applicable to the text-classification application. [5] To increase the performance of the Centroid classifier, a novel batch-updated method is proposed in this paper. The aim of this approach is to successively update the classification model by batch, by taking advantage of training errors. [6] This paper explores a new technique of feature selection metrics using less number of keywords which is highly successful. [7] To deal with multi-label classification problems, this paper proposes the Probabilistic Neural Network (PNN) algorithm, and is compared with Ml-kNN algorithm. This application divides the MI-kNN algorithm into four parts which is used for multi-label categorization problems. [8] This paper describes a natural language processing system reinforced by the use of association of words and concepts, implemented as a neural network. Combining an associative network with a conventional system contributes to semantic disambiguation in the process of interpretation.[9] In this paper, a new text document classifier is implemented using the support vector machine (SVM) training algorithm and the K-nearest neighbor(KNN) classification approach combined together. The Support Vector Machine - Nearest Neighbor classification approach is named as SVM-NN. [10] In this paper, it takes the advantage of both longest common subsequence (LCS) and VSM algorithm and proposes integrated text retrieval (ITR) mechanism. LCS is used to evaluate the weight of terms and is the main idea of the ITR mechanism, so that the weight relationships and the sequence between the texts and the query can be examined concurrently. [11] This paper measures the virtual generalizing random access memory weightless neural networks (VG-RAM WNN), which is an efficient method for machine learning technique which is very simple to implement and faster in training and testing. To build automatic multi-label text categorization systems, VG-RAM WNN is used as a tool. The performance of the VG-RAM WNN is evaluated on two issues: 1) classification of the text characterization of commercial activities, 2) Web page classification. [12] This paper proposes an algorithm which will learn from the data set provided to perform speech recognition task and multiclass text task. This method is based on fresh and enhanced family for boosting the algorithms. Boos Texter, which is the new algorithm for boosting the performance, is used for text categorization task. [13] This paper presents a mathematical model of classification schemes and the one scheme which can be proved optimal among all those based on word frequencies. [14] This paper represents a method DP4FC which is used to choose appropriate feature to categorize and differentiate the appropriate documents from the inappropriate documents. DP4FC is combined with the other classifiers. After getting the appropriate document, the classifier creates the effective category groups and takes appropriate decisions in classifying and filtering. [15] In this paper, for dimension reduction, the phonological different words, grammatical words, and the stopwords are recognized and eliminated. There are two algorithms for dimension reduction. They are frequent term generation and improved stemming algorithms. [16] This paper explains the flow of the processing of the information and for text categorization. There are two efficient learning algorithms. They are Partial Least Squares (PLS) and Support Vector Machines (SVM) and is applied in other domain as well. [17] In this paper, the authors explain about the steps. They are rule generation, calculation of probability and pre-processing. The training set document is read in the rule generation. Negative and positive weights are calculated in the calculation of probability. The document which is given as input is divided into statements and paragraphs in pre-processing. [18] This paper makes use of statistical term clustering and syntactic processing to represent a document which is more accurate than obtained by using traditional keyword methods. [19] This paper compares the success rate of automatic learning algorithms by means of speed in learning, accuracy and speed in real time calculation for categorization of text. This paper also checks the size of training set and other representations of document. [20] This paper proposes transfer of knowledge method which is mapped from source to target domain based on feature representation. A new future space is created first, then feature representation map is built, and the target and source domain is reweighed. With the help of this, in the source domain, classifier models are trained which is used by target domain. [21] This paper tells the use of linear regression residual for binary text categorization. The main idea is to predict the given test vector using its k nearest neighbors in both positive and negative classes. [22] In this paper, the problem of classifying text by removing the information which is gained from clustering both testing and training is addressed. The knowledge which is gained from clustering is needed to increase the performance of the text classifier. [23] This paper explains the algorithm which combines the feature of k-nearest neighbor (KNN) and support vector machine (SVM) methods to improve the precision of classification of text which is based on variable precision rough sets (VPRS). [24] This paper explores on enhancing the *k*NN which is improved by implementing alternate distance functions which has weights to measure the data from various viewpoints. By using genetic algorithm, the weights for optimization are computed [25-29]. This paper gives the solution by Back propagation network, the techniques used for feature identification. The

back propagation network algorithm is used for the text classification.

3. PROPOSED METHOD

We have broadly classified our project into two stages where the first one deals with conversion of handwritten characters into editable format and another one deals with classifying people as positive minded or negative minded based on the content of what they write. We will see both the stages in details in the following section.

STAGE 1:

Optical Character Recognition:

Handwritten character recognition is broadly classified as online character recognition which is the real time acquisition and recognition of characters and offline character recognition which deals with recognition of characters which is written on a sheet of paper. This can be achieved with three techniques namely OCR (optical character recognition), MICR (magnetic ink character recognition), OMR (optical mark recognition).

Pre-processing and segmentation:

The image is given as a input to the OCR template matching algorithm where the characters are processed which involves segmenting the characters by using horizontal and vertical profiling using OCR techniques.

Feature Extraction:

For the given input image we get the corresponding vertical profile of complemented image from where we extract the required features of the segmented characters.

Text Classification:

We obtain matrix representation of recognized characters and thereby the text is classified. And hence we get the editable format which is the output of our first proposed model.

Algorithm 1: Character Recognition:

Input: A text image
Output: An editable document
Method:

```
for i=1 to length(Training_Samples)
    img=imread(dataset(i));
    No_Lines=HorizontalProfile(img);
    No_Char=VerticalProfile(img);
    Identified_Text=OCR(No_Char);
    Save("Identified_Text.txt");
  End
```

Stage 2:

As a person's mind contextually set plays a major role in the what a person writes, we are classifying them as positive minded or negative minded based on the content of the written matter.

Text mining and machine learning techniques:

The output of the previous model i.e., editable document is further processed in this model to get the final output. We use text mining techniques and machine learning techniques along with stop word elimination algorithm to classify if the person is positive minded or negative minded which our end result is.

A survey was conducted to understand the regional vocabulary of people and we collected English words from them which is classified as positive word and negative word by them according to their thinking. The detailed model is explained in the following section, here we use compression based integer representation based approach for classifying the extracted words as positive or negative. The task of classifying is a supervised task where we train the classifying algorithm with terms belonging to two major classes positive and negative. The result of this classification algorithm is to assign binary values {0, 1} (1->if the application recognises the terms properly else its 0). We are emphasizing at using integers based compression due to the fact that text terms occupies more space than integers. Once we are able to convert the terms to integers it will be very easy to handle the integer numbers and hence it contributes a lot to classification algorithm. The detailed explanation for this is presented in the corresponding subsections.

Classification Stage:

We first read the positive dictionary of words and negative dictionary of words and save it. Given the query document we first apply natural language processing methods and eliminate stop words. After which the set of positive and negative words of the query document will be compared with the dictionary. If found we keep the count of positive and negative words and based on the frequency of occurrence of words we classify the content of text as positive or negative. If the positive or negative word which is present in the query document is not found in the dictionary we update the dictionary with the new words.

Algorithm 2: Personality Analysis

Input: A query Document
Output: classifying the document as positive or negative
Method:

```
  Positive <- read pos_dictionary
  save Positive
  Negative <- read neg_dictionary
  save Negative

  [Prow Pcol]=size(Positive)
  [Nrow Ncol]=size(Negative)
  [Trow Tcol]=size(Reg_Text)
  for i<-1 to Prow
    for j<-1 to Trow
      if(strcmp(Positive(i), Reg_Text(j))=1)
        pos++
```

```
        end
    end
    for i<-1 to Nrow
        for j<-1 to Trow
            if(strcmp(Negative(i), Reg_Text(j))=1)
                neg++
            end
        end
    end
    End
```

4. EXPERIMENTATION

We have performed experimentation on three types of datasets. They include well-formed characters, partially well-formed characters and non-identifiable characters. Here we have mainly used two types of experimentation techniques to find the accuracy of the handwritten characters that are recognised.

1. The first one is using optical character recognition technique (OCR).
2. The second one is using Text Mining

4.1 Optical Character Recognition:

STEP 1: Here we initially give handwritten sample as input to OCR and find out the characters that are recognized.

STEP 2: Next we generate the Confusion matrix as follows:

- Create Rows and Columns of matrix using English Alphabets.
- Mark those cells in the matrix depending on how the alphabets are recognised.
- Calculate Row sum and Column sum, this will be our Recall and Precision for finding F Measure.

F measure calculated for partially well-formed characters is shown in Table 1.

4.2 Text Mining:

Steps:

1. Once the OCR data gets converted into editable text format and this will be stored in a .txt file.
2. The obtained data set will be divided into training and testing sample.
3. Stop words are removed from the training sample and integer representation is given to the data. Results of text mining stage is presented in Table 2.

5. CONCLUSION

This project can be useful for conversion of old handwritten documents into digital form. This will help a lot of organisations who have legacy documents in need of digitisation. This can also help students to digitize their notes. Right now there are scanners which are used to scan printed documents. But this is will be a single app which can do the work of a hardware device and its separate software. If this technology is used for other languages we can easily convert old books which are in need of restoration. This will convert the book to digital form and prevent the natural wear and tear that physical books are often subjected to. The personality analysis can be used by various organisations to judge a person before taking them into their organisation or to check the changes in a person's state of mind over time. This particular study can have huge applications in the field of psychology. We can see the similar works done in the references [26] to [51].

ACKNOWLEDGMENT

This research project was supported by Department of Computer Science & Engineering, Sahyadri College of Engineering & Management, Mangalore. We thank all the teaching and non-staff for their continuous support and encouragement.

REFERENCES

[1] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah khan, A Review of Machine Learning Algorithms for Text-Documents Classification, Department of Computer and Information Science, Universiti Teknologi PETRONAS, Tronoh, Malaysia.

[2] S. N. Bharath Bhushan, Ajit Danti and Steven Lawrence Fernandes. Integer Representation and B-Tree for Classification of Text Documents: An Integrated Approach.

[3] Thorsten joachims, GMD Forsehungszentrum IT, AIS.KD Schloss Birlinghoven, 53754 Sankt Augustin, Germany

[4] Mita K. Dalal, Mukesh A. Zaveri Automatic Text Classification: A Technical Review ,International Journal of Computer Applications (0975 – 8887) Volume 28– No.2, August 2011

[5] Songbo Tan, An improved centroid classifier for text categorization, 2007 Elsevier Ltd.

[6] Serafettin Tasc, Tunga Güngör , Comparison of text feature selection policies and using an adaptive framework, 2013 Elsevier Ltd.

[7] Elias Oliveira, Patrick Marques Ciarelli. Claudine Gon¸calves. A Comparison Between a kNN based Aproach and a PNN Algorithm for a Multi-Label Classification Problem, Universidade Federal do Esp´ırito Santo,Brazil

[8] KIMURA Kazuhiro SUZUOKA Takashi AMANO Sin-ya, Association-based Natural Language Processing with Neural Networks, Information Systems Laboratory Research and Development Center TOSHIBA Corp.

[9] Chin Heng Wan a, Lam Hong Lee b, Rajprasad Rajkumar b, Dino Isa, A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine, 2012 Elsevier Ltd.

[10] Cheng-Shiun Tasi, Yong-Ming Huang, Chien-Hung Liu, Yueh-Min Huang, Applying VSM and LCS to develop an integrated text retrieval mechanism, 2011 Elsevier Ltd.

[11] Alberto F. De Souza, Felipe Pedroni, Elias Oliveira, Patrick M. Ciarelli, Wallace Favoreto Henrique, Lucas

Veronese, Claudine Badue, Automated multi-label text categorization with VG-RAM weightless neural networks. Elsevier 2009.
[12] Robert E Schapire, Yoram Singer, Boos Texter: A Boosting-based System for Text Categorization, Machine Learning, 39(2/3):135-168, 2000.
[13] Louise Guthrie Elbert Walker, Document Classification by Machine: Theory and Practice.
[14] Rey-Long Liu, Dynamic category profiling for text filtering and classification.
[15] P. Ponmuthuramalingam and T. Devi, Effective Dimension Reduction Techniques for Text Documents, IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.7, July 2010 .
[16] Setu Madhavi Namburu, Haiying Tu, Jianhui Luo and Krishna R. Pattipati, Experiments on Supervised Learning Algorithms for Text Categorization IEEEAC paper #1260, Version 8, Updated December 10, 2004 .
[17] S. Subbaiah, Extracting Knowledge using Probabilistic Classifier for Text Mining, Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22.
[18] Tomek Strzalkowski and Barbara Vauthey, Fast Text Processing for Information Retrieval, Courant Institute of Mathematical Sciences New York University 251 Mercer Street New York, NY 10012
[19] Susan Dumais, John Platt, David Heckerman, Inductive Learning Algorithms and Representations for Text Categorization
[20] Jiana Meng, Hongfei Lin, Yanpeng Li, Knowledge transfer based on feature representation mapping for text classification, 2011 Elsevier Ltd.
[21] Hakan Altınçay, Using Linear Regression Residual of Document Vectors in Text Categorization, 2013 IEEE
[22] Antonia Kyriakopoulou, Theodore Kalamboukis, Using Clustering to Enhance Text Classification", SIGIR. Amsterdam, The Netherlands. ACM 978-1-59593-597-7/07/0007.
[23] Wen Li, Duoqian Miao, Weili Wang, Two-level hierarchical combination method for text classification, 2010 Elsevier Ltd
[24] Takahiro Yamada, Kyohei Yamashita, Naohiro Ishii, Text Classification by Combining Different Distance Functions with Weights", 2006 IEEE
[25] S.Ramasundaram, S.P. Victor, Text Categorization by Back propagation Network, International Journal of Computer Applications (0975 – 8887) Volume 8– No.6, October 2010.
[26] Bhushan Bharath S. N. and Danti Ajit. Classification of text documents based on score level fusion approach. Pattern Recognition Letters 94., 118–126. 2017.
[27] Danti Ajit and Bhushan Bharath S N. 2013, Document Vector Space Representation Model for Automatic Text Classification. In Proceedings of International Conference on Multimedia Processing, Communication and Information Technology, Shimoga. pp. 338–344
[28] Danti Ajit and Bhushan Bharath. Classification of Text Documents Using Integer Representation and Regression: An Integrated Approach.Special Issue of The IIOAB Scopus Indexed Journal.Vol. 7, No.2, pp. 45–50. 2016.

Table 1: F measure calculated for partially well-formed characters.

F Measure of recognized handwritten text in editable format		
PRECISION	RECALL	F MEASURE
1	0.9285	0.9523

Table 2: F measure calculated for recognized characters.

F Measure of recognized handwritten text in editable format		
PRECISION	**RECALL**	**F MEASURE**
0.91	0.9225	0.9162