DETECTING GENDER FROM HANDWRITING SAMPLES IN THE ENGLISH LANGUAGE

Burak Koryan[1],Chenhao Ding[2]

Department of Electrical and Computer Engineering,University of New Brunswick

15 Dineen Drive,Fredericton,NB E3B 5H5

[1]b.fk@unb.ca , [2]Chenhao.Ding@unb.ca

Abstract – For this paper, multiple handwriting samples,in the English language, from random females and males were gathered.LDA,QDA,k-NN,and Naïve Bayes classifiers were used to show that extracted features from datasets were distinct and handwriting samples can be classified by writer's gender with moderate to high classification accuracies.

Keywords : Handwriting Classification,Handwriting recognition,parametric classifiers,non-parametric classifers

I. Introduction

Handwriting recognition using pattern recognition tools has been becoming more popular due to their efficiency and speed and accuracy in classifying handwriting with computers.Handwriting recognition has been a necessity since early days of advancements in electronics and human-computer interaction.With increasing integration of digital devices into humans'lives create the necessity of recognizing identification of users and handwriting is one way of identifying civilized humans. Handwriting recognition has been used in many fields such as verifying identification through signatures or initials and forensic research.Handwriting is considered to be unique to every human being and has tens of distinct features depending on the language its written and and the writer's age,gender,and talent.A through overview of importance of handwriting recognition and writer identification is provided in [1] – [2].In human-computer interaction,the users'input into the computer can be in many forms such as keystroke,facial or physical expression or handwriting.Visually distinctive handwriting can give a lot of information about the writer's identification if thorough analysis and classification is done well enough.First thing most recognition systems do,if identifying a person required,is detecting the person's gender.With this information classification can be done a lot easier by eliminating unnecessary information.Many projects and research has been done on handwriting recognition and classification such as *An Off-line Cursive Handwriting Recognition System* by Robinson et al.In their work,features of handwriting such as endpoints,turning points,loops and dots have been examined for feature extraction and using neural networks open-vocabulary cursive handwriting samples were successfully classified with 87 percent recognition rate[3].Similar work has been done by Joseph et al. on Online *Handwritten Malayalam Character Recognition using LIBSVM in MATLAB.*Optical character recognition (OCR) of MATLAB has been used with help of support vector machine (SVM) classifier.Over 160 different handwriting styles have been used and SVM classifier with kernel of degree 3 as well as Gaussian kernel provided accuracy above 90 percent[4].Bayesian Decision Theory,the k-nearest neighbour algorithm(kNN),Linear Discriminant Analysis(LDA) and more statistical classifiers are used to separate data into classes.For instance,in *An intelligent Character Recognition System for Automatic mark Capturing* by Jordaan et al. about 100% accuracy rate was successfully achieved using kNN,Bayesian Decision Theory,and artificial neural networks when 400 samples were used in the

process[5].In this work,we are proposing that we can detect gender from handwriting samples.Handwriting samples from 56 females and 41 males were gathered.These people wrote 28 different words and letters on template printed out.Then these printed templates were scanned with flatbed scannter in 72 DPI with 1646 x 2328 pixels resolution.Pre-processing has been done using MATLAB's image processing functions and the samples were made available for feature extraction and classification.Only two features has been selected to simplify the process : Entropy and ratio of black pixels in the upper-half horizontal plane of sample images.These two features with appropriate labels were fed into four different classifiers,LDA,QDA,kNN,and Naive Bayes and collected results were plotted for graphical representation written in tabular form.

II. Methodology

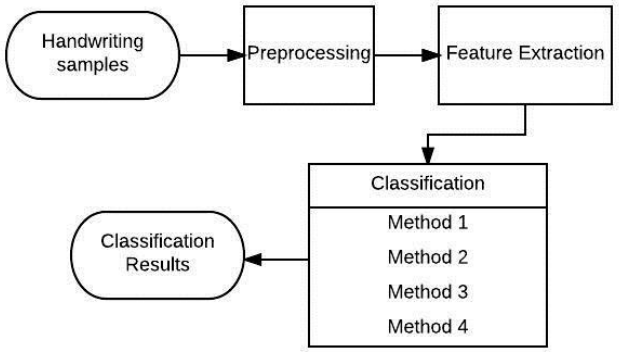

Fig 1. Block Diagram of classification system

- **Collection of handwriting samples:**

From 51 females and 46 males handwriting samples were collected in the English language using the template shown in Figure 1.There are 28 words and letters (8 letter + 20 words) in different vowel combinations. In total of 97 sample sheets were collected and used for this paper.

- **Image Pre-Processing:**

In order to make each sample word or letter usable for classification, there needed to be pre-processing done.First,using MATLAB's image processing function *imcrop()*,each word and letter were cropped automatically by writing their coordinates on the sample sheet in the source code and converted from JPG to TIF.After cropping,each sample were named differently to not cause confusion and were categorized.Next,because some handwriting samples were written in different color other than in black ink or for any other reason,as shown in Fig.3,the samples were converted into grayscale image from RGB.The sample images of words and letters were ready to convert them into binary as shown in Fig 4.With conversion into binary,black pixels in a sample image were assigned logic '1' and white pixels were assigned logic '0'.Using these matrices,in binary, that represented all samples,entropy and black pixel percentage in the upper-half plane of a sample image were calculated.

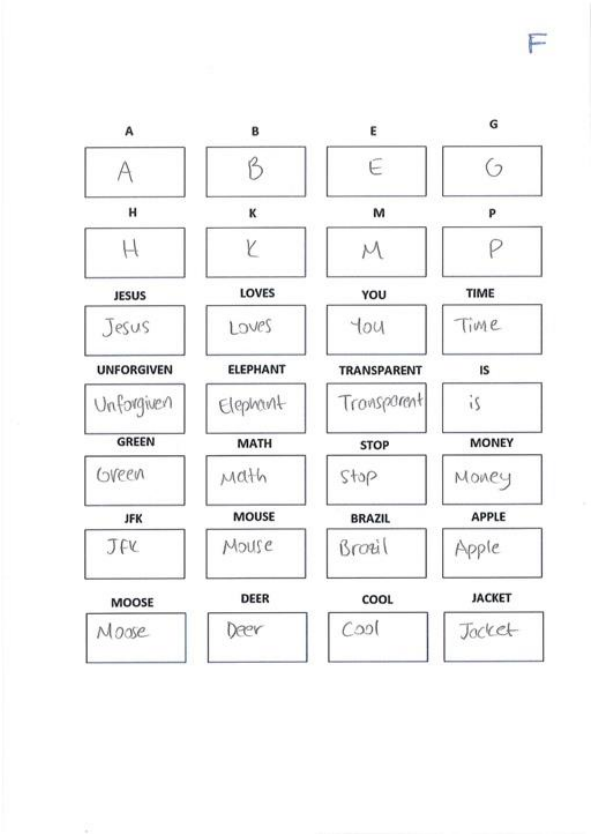

Fig 2. Handwriting Sample Template

Fig 3: Random Handwriting Samples

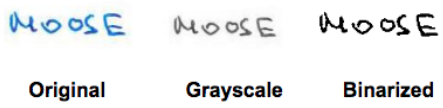

Fig 4: Converting sample image from its original to binary

- **Feature Extraction**

  A handwriting has many features that has been used for research and projects. The following handwriting features are considered in this paper but only two were selected :

  - Aspect ratio [6]
  - Percent of pixels above horizontal half point [6]
  - Number of strokes [6]
  - Average distance from image center [6]
  - Slant [7]
  - Word rotation [7]
  - Entropy

For our this project, entropy of selected sample image and percent of pixels above horizontal half point were chosen as two features. Entropy is a measure of randomness that can be used to characterize the texture of the input image[8].Entropy of sample images were found after they were converted to grayscale image and using MATLAB®'s *entropy()* function. The number of black pixels above horizontal half point were found by counting only black pixels above horizontal half point.The horizontal half point of an image is shown in Fig 5.Percent of pixels above horizontal half point is found using Eq. 1:

$$BP(\%) = \frac{\#\ of\ BP}{total\ \#\ of\ BP} \qquad (1)$$

where BP : black pixels above horizontal half point

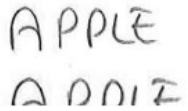

Fig 5: Sample image of the word "Apple" were cut from horizontal half point

- **Classification**
- LDA

- Linear Discriminant Analysis (LDA) is a method used in pattern recognition to find a linear combination of features that separates two or more classes. LDA has good separability for two classes have a linear boundary. It also is the most commonly used as dimensionality reduction technique in the pre-processing step.

In mathematics, the Gaussians for each class are assumed to share the same covariance matrix. In this project, both male and female class has 2-D data points and note that the mathematical formulation of the classification strategy parallels the MATLAB implementation associated with this work. The first step is compute the mean of each data set and mean of entire data set.

$$\mu_3 = p_1 \times \mu_1 + p_2 \times \mu_2 \qquad (1)$$

Where the p is the probabilities of the two classes, and the between-class scatter is computes using the following equation. [12]

$$S_b = (\mu_j - \mu_3) \times (\mu_j - \mu_3)^T \qquad (2)$$

Note that S_b is the covariance of data set whose members are the mean of each class. The test vectors are transformed and are classified using the Euclidean distance from each class mean.

Also the density has formula for a multivariate Gaussian distribution.

$$fk(x) = \frac{1}{(2\pi)^{P/2}|\Sigma k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}\mathbf{k})^T \boldsymbol{\Sigma}\mathbf{k}^{-1}(\mathbf{x}-\boldsymbol{\mu}\mathbf{k})} \qquad (3)$$

Where p is the dimension, Σk is the covariance matrix and x is the class vector. [11]

- QDA

Quadratic Discriminant Analysis (QDA) is closely related to LDA, where it is assumed that the measurements from each class are normally distributed, meanwhile in QDA there is no assumption that the covariance of each classes is identical. It has better separability for two classes have a non-linear boundary than LDA but it needs to estimate the covariance matrix for each classes.The uniform posterior probability is using the discriminant analysis model which the class k is 1 over the total number of classes.In mathematics, the QDA does not have much different from LDA except the different covariance matrix for each class. The quadratic discriminant has function:

$$\delta k(x) = -\frac{1}{2}\log|\Sigma k| - \frac{1}{2}(x-\mu k)^T \Sigma k^{-1}(x-\mu k) + \log \pi k \quad (4)$$

where Σk are the covariance matrix for the classes.[13]

- kNN

The k-nearest neighbor algorithm (kNN) is a non-parametric lazy learning algorithm used for classification. k is a user-defined constant and an unlabeled vector or test point is classified by assigning the label which is most frequent among the k training samples nearest to that query point. With a higher k value, the user can get a smoother boundary, but if the k is too high it will filter too much information. There is not a good way exist to decide the k value. So in this project, multiple k value were tested to find the best created boundary. It is usually an odd number if the number of classes is 2, but in this project k = 2,3,5,6,8,10 and keep increasing the k size and till k neighbors are captured.

In mathematics, the kNN classifier can be viewed as assigning a weight 1/k and others 0 weigh. The categorizing query points based on their distance to points in a training data set can be a simple way to classifier new points. For example if K = 1, then the kNN is simply assigned to the class of its nearest neighbor, also known as the Euclidean distance.

$$d = \sqrt{(x_s - y_t)V^{-1}(x_s - y_t)'} \quad (5)$$

Where mx-by-n data matrix X, which is treated as mx (1-by-n) row vectors x_1, x_2, ..., x_{mx}, and an my-by-n data matrix Y, which is treated as my (1-by-n) row vectors y_1, y_2, ...,$y_{my.}$ V is the n-by-n data diagonal matrix.[14]

- Naïve Bayes

Naïve Bayes classifiers are a probabilistic classifiers based on applying Bayes' theorem with naïve independence assumptions between the features.

$$P(A|B) = \frac{P(A|B)P(A)}{P(B)} \quad (6)$$

According to Bayes' theorem, where P(A|B) is the conditional probability which define the event A occurring given that B is true. P(A) and P(B) are the marginal probability of class A and class B. [15]

- Classification Error Rate Calculation:

In order to find error rate obtained from classifiers,using confusion matrix from each classification,error rate is calculated.To do so the two parameters are required from a confusion matrix.For instance for the word "time" the error rate was calculated as follows:

```
ldaResubCM =

    54     2
     9    32
```

Fig 6: Confusion matrix of LDA Classification of the sample word "time"

	Actual Class	
Predicted Class	**Female**	**Male**
Female	TP	FP
Male	FN	TN

Fig 7 : Typical Confusion matrix setup

where : TP : True Positive
FP : False Positive,
FN : False Negative,
TN: True Negative

$$\%\ error\ rate = \frac{FP + FN}{\#\ of sample\ data} * 100 \quad (7)$$

To verify our hand calculation results, MATLAB®'s *resubLoss()* function were used instead in the source code of the project.The *resubLoss()* function simply calculates classification error by resubstitution.[10]

III. Results:

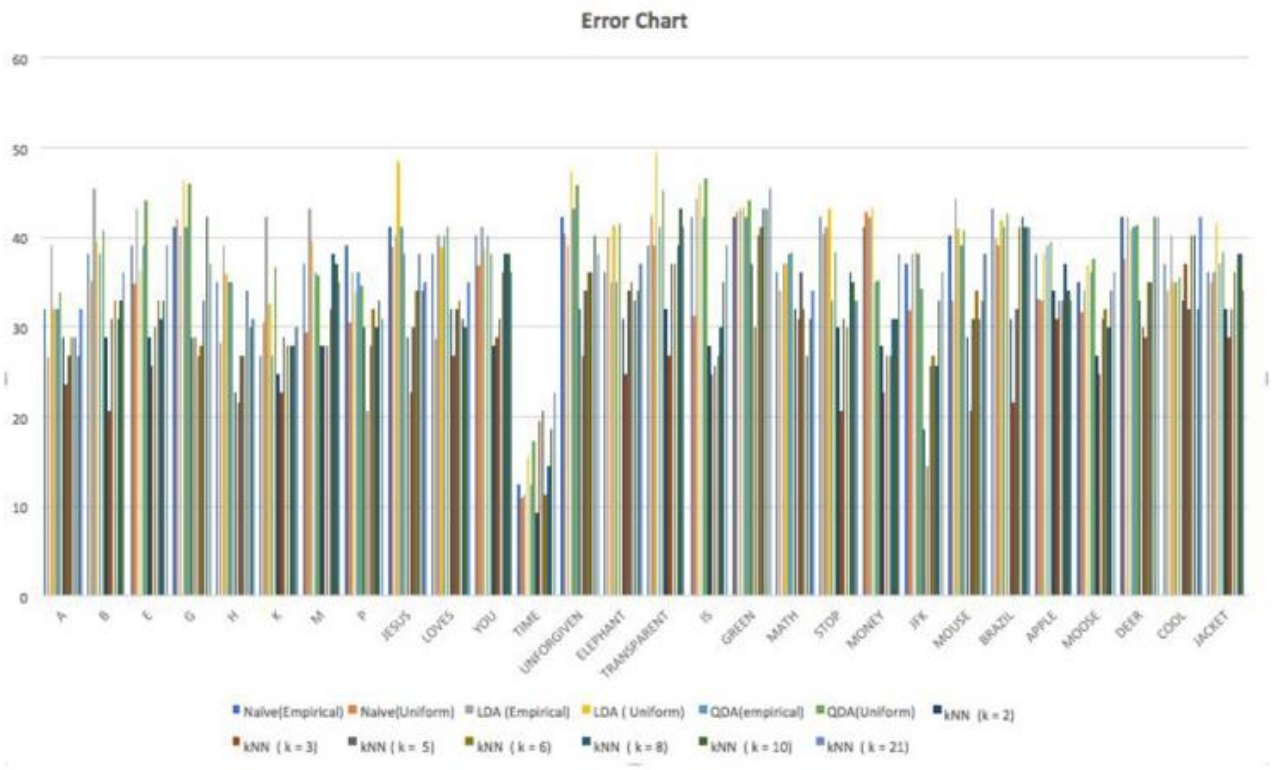

Figure 6: Error rates from all classifiers obtained using all samples words

From the classification error assessment by calculation error rate,the following conclusions can be made: the sample word “time” was classified best,with lowest error rate of 25% in general, with all classifiers used in the project as shown in Fig 6.Mean classification error for classifiers that used all sample words is shown in Fig 7.From this chart,it is necessary to say that the lowest error rates were obtained with k-Nearest Neighbour classifier when k = 2 and k = 3 and k = 4 23.70%,25.20% and 27.75% respectively.When the k value increased to 21,the error rate reached to 36% simply due to overfitting.Overfitting with increased value of k is shown in Fig 8 and Fig 9.

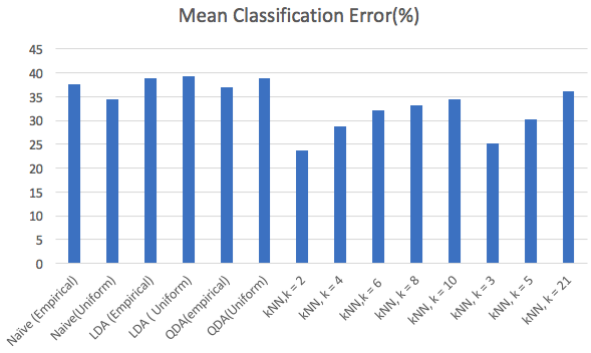

Fig 7 : Mean Classification Error Chart

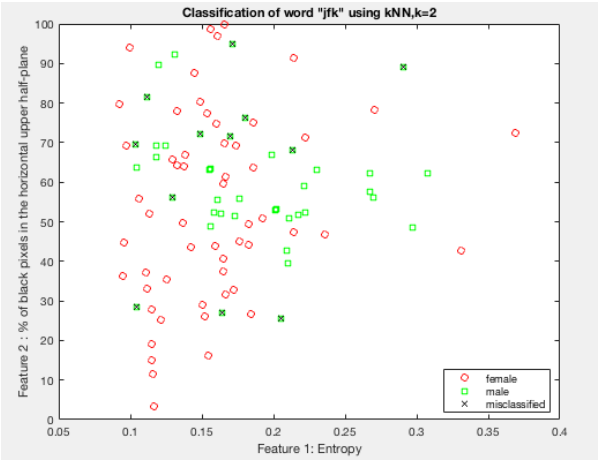

Fig 8 : Classification of Sample word “jfk” using k-NN with k = 2

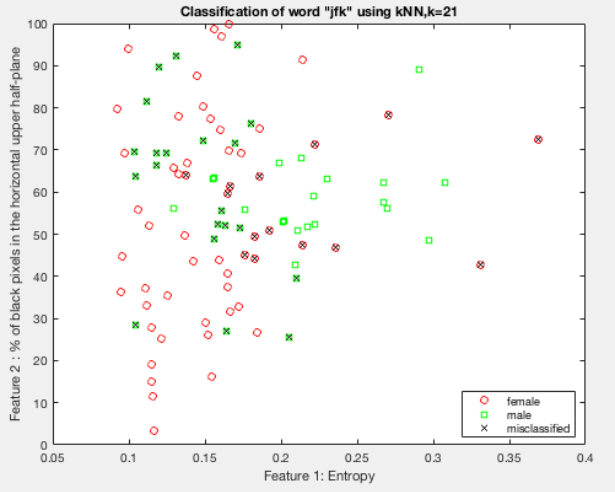

Fig 9 : Classification of Sample word “jfk” using k-NN with k = 21

As mentioned before,the sample word “time” was classified best with k-NN when k = 2 with classification error rate of 8.25% (therefore classification accuracy of 91.75%) and worst when k = 21 with classification error rate 22.68% (therefore classification accuracy of 77.32%).

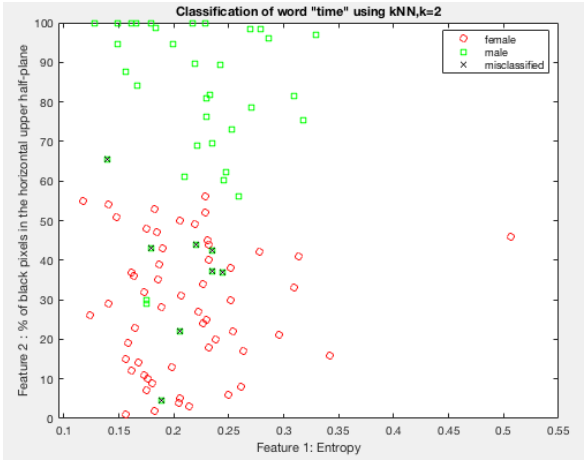

Fig 10 : Classification of Sample word “time” using k-NN with k = 2

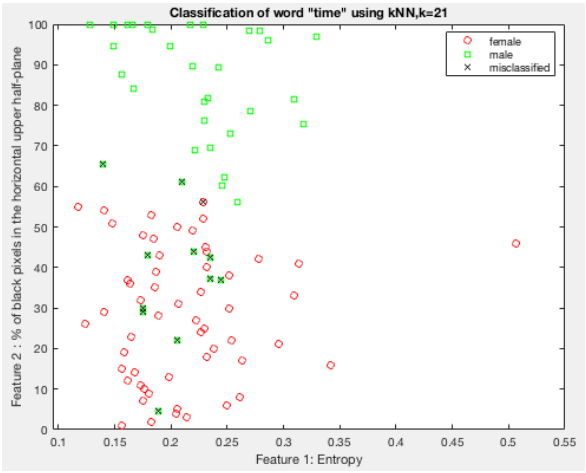

Fig 11 : Classification of Sample word “time” using k-NN with k = 21

In Fig 10 and Fig 11,classification of the sample word “time” is shown using k-NN with k = 2 and k = 21.The classification error obtained was 8.25% and 22.68%,respectively.This,again,proves that increasing k-value may cause increased classification error.

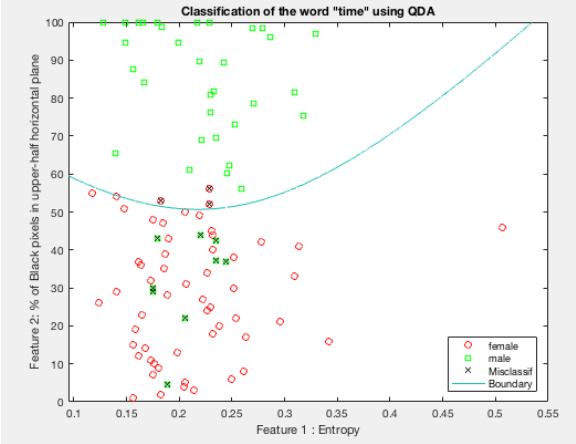

Fig 12: Classification of sample word “time” with QDA, Prior Probab:Empirical

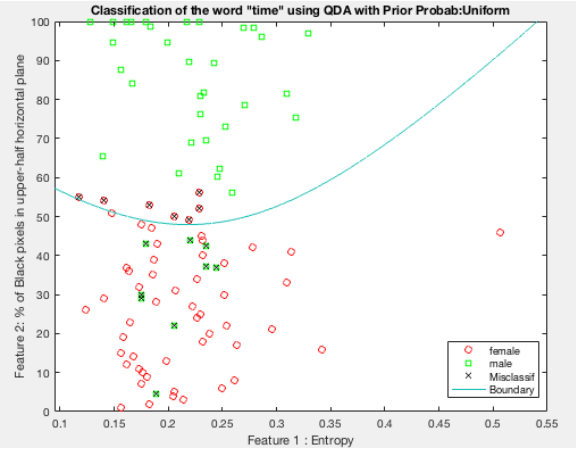

Fig 13: Classification of sample word “time” with QDA,Prior Probab: Uniform

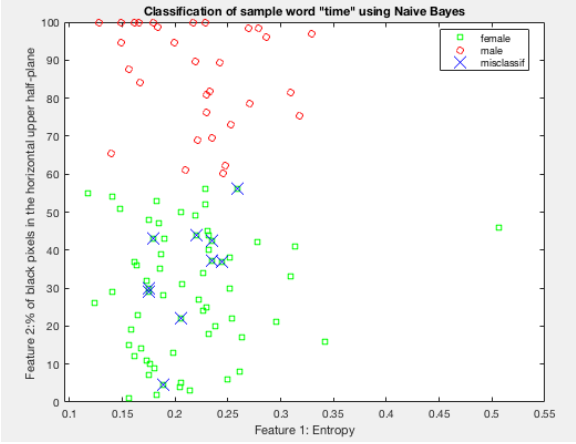

Fig 14: Classification of sample word “time” with Naïve Bayes Prior Probab:Empirical

In Fig 12 and Fig 13 classification of the sample word “time” is shown.In Fig 12,the sample was classified using QDA with empirical prior probability and 12.37% classification error occurred as a result.However,when the prior probability was changed to uniform,the classification error increased to 17.23% as misclassified data points is shown in Fig 13.From this result,we can say that changing prior probability does not always increase or decrease the classification error rate.This depends on the data samples used and classifier type.

In Fig 14,Naïve Bayes classification of the same sample data is shown.When the prior probability is set to empirical the classification error was 10.31% whereas when its set to uniform,as shown in Fig 15,the classification error was 10.98%.

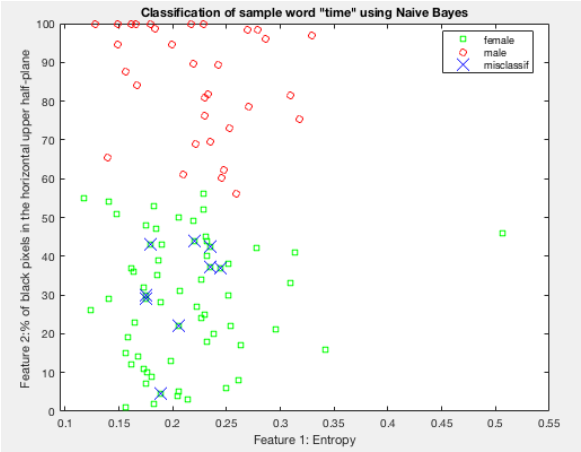

Fig 15: Classification of the word “time” using Naïve Bayes with prior probab: uniform

III. Conclusion:

With increased demand in human-computer interaction,identifying user has becoming more important than ever for protecting personal data or device.In this project,detection of gender from handwriting samples has been done using LDA,QDA,k-NN,and Naïve Bayes classifiers.From 56 female and 41 male,handwriting samples were gathered and classified.According to our results,it can be concluded that the most accurately classified sample word was “time” with classification accuracy of 91.75% when k-NN (with k = 2) was used.Another conclusion worth to mention is the prior probability assigned to classifiers.Changing prior probability definitely changes classification results by either

increasing or decreasing classification accuracy;so this should be taken into consideration according to a project's needs.By observing the overall results,as an interesting fact,it is worth mentioning that,by observing classification plots,it can be said that,surprisingly because the number of female data samples are a lot more (57 versus 41) than males,female handwritings are misclassified more than male handwritings.

Acknowledgements:

For his efforts and contributions to our knowledge in pattern recognition,we thank Dr.Erik Scheme.Burak Koryan dedicates his portion of work to Tekin Dincer for his everlasting smile in life.

References:

[1] J. Chapran, "Biometric Writer Identification: Feature Analysis And Classification," International Journal of Pattern Recognition and Artificial Intelligence, vol. 20, no. 04, pp. 483–503, 2006.

[2] H. Said, T. Tan, and K. Baker, "Writer identification based on handwriting," IEE Third European Workshop on Handwriting Analysis and Recognition, 1998

[3] A. Senior and A. Robinson, "An off-line cursive handwriting recognition system," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 3, pp. 309–321, 1998.

[4] S. Joseph and A. Hameed, "Online handwritten malayalam character recognition using LIBSVM in matlab," 2014 IEEE National Conference on Communication, Signal Processing and Networking (NCCSN), 2014.

[5] D. B. Mulindwa, S. Du, and J. A. Jordaan, "An intelligent character recognition system for automatic mark capturing," 2014 7th International Congress on Image and Signal Processing, 2014.

[6] A. Senior and A. Robinson, "An off-line cursive handwriting recognition system," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 3, pp. 309–321, 1998

[7] "Handwriting recognition," Wikipedia, 30-Nov-2017. [Online].Available:https://en.wikipedia.org/wiki/Handwriting_recognition. [Accessed: 02-Dec-2017].

[8] Entropy of grayscale image - MATLAB entropy. [Online]. Available:https://www.mathworks.com/help/images/ref/entropy.html. [Accessed: 02-Dec-2017].

[9] Definition of Prior Probability by Oxford English Dictionaries.[Online].Available:https://en.oxforddictionaries.com/definition/us/prior_probability. [Accessed: 02-Dec-2017].

[10] "resubLoss()," Classification error by resubstitution - MATLAB.[Online].Available: https://www.mathworks.com/help/stats/classificationensemble.resubloss.html.[Accessed:02-Dec-2017]

[11] Applied Data Mining and Statistical Learning – 9.2.2 Linear Discriminant Analysis. [Online]. Available: https://onlinecourses.science.psu.edu/stat857/node/74 [Accessed:02-Dec-2017].

[12] S. Balakrishnama and A.Ganapathiraju, "Linear Discriminant Analysis – A brief Tutorial" [Online].Available:https://www.isip.piconepress.com/publications/reports/1998/isip/lda/lda_theory.pdf.[Accessed:02-Dec-2017].

[13] Applied Data Mining and Statistical Learning – 9.2.8 Quadratic Discriminant Analysis. [Online]. Available: https://onlinecourses.science.psu.edu/stat857/node/80 [Accessed:02-Dec-2017].

[14] "Classification Using Nearest Neighbors" MathWorks [Online].Available:https://www.mathworks.com/help/stats/classification-using-nearest-neighbors.html [Accessed:02-Dec-2017].

[15] "Naïve Bayes classifier" Wikipedia, 30-Nov-2017. [Online].Available:https://en.wikipedia.org/wiki/Naive_Bayes_classifier [Accessed: 02-Dec-2017].