International Journal of Advanced Research in Engineering Technology & Sciences ISSN: 2394-2819
April-2016 Volume 3, Issue-4
Email: editor@ijarets.org www.ijarets.org

Text Character Detection

Preeti
M. Tech Scholar
Department of ECE
RIMT Rayat Bhahara Group
Chidana, Haryana

Parveen Khanchi
Assistant Professor
Department of ECE
RIMT Rayat Bhahara Group
Chidana, Haryana

ABSTRACT:
Use of characters in communication is very common now days to express feeling. There are many type of character like facial character and gestures, by write text, and by speech etc. There are many techniques exist for character detection like SVM, Naïve Bayes. In this paper we are going to study about these techniques and compare the results on basis of recognition of expressions of the six basic characters [1]. So for the comparison study data on which these techniques applied should be full of character to get more accurate result.

KEYWORD: - Text, Character, SVM, Naïve Bayes, Neural

I. INTRODUCTION:

Characters have fascinated researchers for long, as is evident in the vast body of research work related to character in fields of psychology, linguistics, social sciences, and communication. Human character manifests itself in the form of facial expressions, speech utterances, writings, and in gestures and actions. Consequently, scientific research in character has been pursued along several dimensions and has drawn upon research from various fields. Some psychologists have investigated facial expressions of character to identify the basic discriminable expressions among them, and mapped them to basic human characters. Basic characters are like sadness, happiness etc. [1].This work uses Ekman's character categories since these characters have been most widely accepted by the different researchers. Ekman's character categories have also been previously used in other computational approaches to character recognition [2, 3, and 4].
Human character can be sensed from such cues as facial expression, gestures, speech and writings. Research in character has focused on all these aspects. Computational approaches to character analyses have focused on various character modalities, resulting in a large number of multi-modal character-annotated data. Recognition and classification of character in text can be regarded as a sub-field of sentiment analysis.

II. TECHNIQUES USED FOR CHARACTER DETECTION:

Many techniques are available in market that is used for detection of character from text. Some of them we are going discussing in this paper. As we study with comparison of results that will find out which algorithm is worse. We take SVM and Naïve Bayes for the comparative study.

A. SUPPORT VECTOR MACHINE (SVM):

Support vector machine classifier is used to make segments of selected data on the basis of characters and simple text. Input data is presented in two sets of vectors in n-dimensional space, a separate hyper-plane is constructed in space due to which margin between two data sets maximize. There are many hyper-plane exist for classify data but we have to find that hyper-plane which provide maximize margin between two data sets. Due to set of support vectors, risk for structure minimizes. This is n-dimensional hyper-plane where define n is number of features of input vectors, that is necessary to define boundary of classes. Binary classes are required to classify training data. Main goal of performance for SVM is to minimize risk of structure. By

a training example as given (a_1, b_1), (a_2, b_2) …. (a_n, b_n), positive and negative examples separated by hyper-plane. If a point a_1 exists on hyper-plane then it satisfies:
($w.a_1$) + d = 0 where d represent distance from origin and w is normal to hyper-plane. There is shortest distance between negative and positive examples define by margin of hyper-plane. A kernel function is used to fit data on hyper-plane. We cannot directly fit data on hyper-plane without svm mechanism. User provides a function like a line, polynomial which select support vector along surface of this function. "Curse of dimensionality" is main property that is used to avoid upper bound on VC-dimension. VC-dimension is used to measure capacity of the machine. As shown in fig 1.

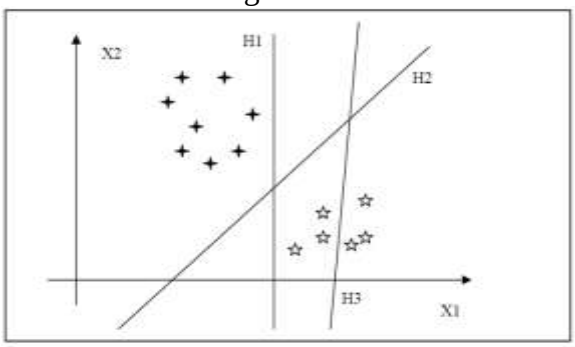

Fig 1 example for SVM

Main object of SVM is try to find nearest distance between point of same class and maximize with point of other class and draw hyper-plane in two categories very clearly as possible.

Kernel Function: During training a user need to define four standard kernels as following. A kernel function use of parameters such as γ, c, and degreethat defined by user during training

Kernel	Formula
Linear	$\mathbf{uv}$
Polynomial	$(\gamma\mathbf{uv}+c)^{degree}$
Radial Basis Function	$exp^{(\gamma\lvert\mathbf{uv}\rvert^2)}$
Sigmoid	$tanh(\gamma\mathbf{uv}+c)$

B. NAÏVE BAYES CLASSIFIER:

Naïve Bayes is used as text classifier because of its simplicity and effectiveness. Simple("naive")classifica1onmethodbased on Bayes rule [7]. The Bayes rule is applied on document for the classification of text. The rule which is following is:

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

This rule is applied for a document d and a class c. probability of A happening given to B can be find with the probability of B given to A. this algorithm work on the basis of likelihood in which probability of document B is same as frequency of words in A. on the basis of words collection and frequencies a category is represented. We can define frequency of word is number of time repetition in document define frequency of that word. We can assume n number of categories from C_0 to C_{n-1}. Determining which category a document D is most associated with means calculating the probability that document D is in category C_i, written $P(C_i|D)$, for each category C_i.

Using the Bayes Rule, you can calculate $P(C_i|D)$ by computing:

$P(C_i|D) = (P(D|C_i) * P(C_i)) / P(D)$

$P(C_i|D)$ is the probability that document D is in category C_i; in document D bag of words is given by probability, which create in category C_i. $P(D|C_i)$ is the probability that for a given category C_i, the words in D appear in that category.

$P(C_i)$ is the probability of a given category; that is, the probability of a document being in category C_i without considering its contents. P(D) is the probability of that specific document occurring. We can classify text with procedure that required using above discussed parameters is as following:

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c|d)$$ MAP is "maximum a posteriori" = most likely class

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d|c)P(c)}{P(d)}$$ Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}} P(d|c)P(c)$$ Dropping the denominator

III. NEURAL CLASSIFIER:

Approaches for object recognition like face or character detection is based on artificial neural network that approach is known as Gabor wavelets [5]. Main objective of this approach is to perform feature extraction. Text with Gabor wavelets is an input that convolves gray-level. A filtering operation is applied on input. This can be applied for whole text but done selectively for particular location. For each location, the magnitudes of the complex filter responses at a range of scales and orientations are combined into a vector (called a 'jet'). This jet characterizes the localized region of the face. Calculating such jets at discrete points in a lattice imposed on the character text yields a feature representation of the text.

Gabor wavelet filters are essentially sinusoidal planar waves restricted by a Gaussian envelope and can be described by

$$\psi_{\mathbf{k}}(\mathbf{x}) = \frac{\mathbf{k}^2}{\sigma^2} \exp\left(-\frac{\mathbf{k}^2\mathbf{x}^2}{2\sigma^2}\right)\left[\exp(i\mathbf{k}\mathbf{x}) - \exp\left(-\frac{\sigma^2}{2}\right)\right]$$

Where the parameter **k** is the characteristic wave vector, which determines the wavelength and orientation of the filter as well as the width of the Gaussian window It is defined as

$$\mathbf{k} = \begin{pmatrix} k_v \cos\phi_\mu \\ k_v \sin\phi_\mu \end{pmatrix}, \; k_v = 2^{-\frac{v+2}{2}}\pi, \; \phi_\mu = \frac{\pi}{8}$$

The filters possess both even (cosine) and odd (sine) parts. The convolution of the filter with the gray-level distribution I(**x**) of text is then given by

IV. COMPARATIVE RESULTS OF TECHNIQUES:

Four different sets of experiments were performed to test the effectiveness and contribution of the different feature groups:

1. Using only features from the General Inquirer (GI).

2. Using only features from WorldNet-Affect (WNA).
3. Combining features from the GI and WNA.
4. Combining all features (including the "other" features comprising of punctuations and emoticons).

Table 2 show the results performed by Naïve Bayes and SVM algorithm on fold of character/non character classification.

Features	Naïve Bayes Accuracy	SVM Accuracy
GI	71.45%	71.33%
WNA	70.16%	70.58%
GI+WNA	71.70%	73.89%
ALL	72.08%	73.89%

Table 2 Results of character/non-character classification

Overall the performance of the SVM classifier was found to be better than that of the Naïve Bayes classifier for this task. The highest accuracy achieved was 73.89%, which surpasses the baseline accuracy of 65.6%. The improvement is statistically significant (on the basis of a t-test, p=0.05). When all features are together then best result was found. There is no effect on result of SVM but it will improve performance of Naïve Bayes.

Table 3 shows result for all feature as following:

Model	Class	Precision	Recall	F-Measure	Baseline F-Measure
Corpus-based Unigrams	Happiness	0.743	0.377	0.500	0.469
	Sadness	0.476	0.341	0.397	0.368
	Anger	0.344	0.302	0.321	0.379
	Disgust	0.529	0.320	0.399	0.179
	Surprise	0.337	0.243	0.283	0.306
	Fear	0.538	0.374	0.441	0.506
	No-emotion	0.394	0.022	0.041	0.579
Roget's Thesaurus (RT) Features	Happiness	0.687	0.319	0.436	0.469
	Sadness	0.388	0.289	0.331	0.368
	Anger	0.400	0.201	0.268	0.379
	Disgust	0.604	0.169	0.264	0.179
	Surprise	0.388	0.226	0.286	0.306
	Fear	0.672	0.391	0.495	0.506
	No-emotion	0.267	0.013	0.025	0.579
Corpus-based Unigrams + RT Features	Happiness	0.690	0.386	0.495	0.469
	Sadness	0.368	0.434	0.398	0.368
	Anger	0.270	0.346	0.303	0.379
	Disgust	0.387	0.308	0.343	0.179
	Surprise	0.258	0.287	0.270	0.306
	Fear	0.360	0.426	0.390	0.506
	No-emotion	0.471	0.055	0.099	0.579
Corpus-based Unigrams + RT Features + WNA Features	Happiness	0.698	0.384	0.0496	0.469
	Sadness	0.361	0.422	0.389	0.368
	Anger	0.268	0.358	0.306	0.379
	Disgust	0.402	0.308	0.349	0.179
	Surprise	0.283	0.296	0.289	0.306
	Fear	0.366	0.426	0.394	0.506
	No-emotion	0.493	0.062	0.11	0.579

Table 5.5 Results of fine-grained classification using Naive Bayes

The results from ten-fold cross-validation experiments conducted using the WEKA [6] machine-learning package are shown in Table 3. The performance using the Naïve Bayes classifier was found to be worse than that of SVM.

V. IMPROVEMENT IN SVM:

An approach is developed by using SVM classifier to improve accuracy. Character models use theories which we have discussed. One of the major problems is the inclusion of subjectivity detection mechanisms which we have studied. Our first step towards this end is the use of gazetteer list in conjunction with syntactic data. Characters were detected based on keywords obtained from gazetteer lists that specifically deal with discovering character keywords. Semantic analysis was performed to enhance the detection accuracy. A prediction accuracy of 96.43% with LibSVM which would suggest that SVM can accurately predict the character class if aided with some human annotated examples that are knowledge rich. Furthermore, these results confirm that the standard SVM algorithm previously discussed is classifying data that is evenly distributed around the decision hyper-plane. The test set comprised of 560 instances (sentences) of which 540 were rightly classified. The training set comprised of 2340 instances classified into positive and negative classes. They found the optimum cost value for the SVM classifier to be 0.125 [7].

VI. CONCLUSION:

We demonstrated that a combination of corpus-based unigram features and features derived from character lexiconscan help automatically distinguish basic character categories in written text. When used together in an SVM-based learning environment, these features increased recall in all cases and the resulting F-measure values significantly for all character categories. With comparison of result we have find out that when features are small in amount than performance of SVM is better than Naïve Bayes but as all feature together then performance of Naive Bayes is better than SVM.

REFERANCES:

1. Ekman, P, **"An Argument for Basic Characters",** Cognition and Character, 6, 169-200, 1992.
2. Liu, H., Lieberman, H., Selker, T**. "A Model of Textual Affect Sensing using Real-World Knowledge",**In Proceedings of the International Conference on Intelligent User Interfaces, IUI 2003, Miami, Florida, USA.
3. Alm, C.O., Roth, D. and Sproat, R. **"Characters from text: machine learning for text based character prediction",** In Proceedings of the Joint Conference on Human Language Technology / Empirical Methods in Natural Language Processing Vancouver, Canada, pages 579-586, 2005.
4. Neviarouskaya, A., Prendinger, H., and Ishizuka, M. **"Analysis of affect expressed through the evolving language of online communication** "In Proceedings of the 12th International Conference on Intelligent User Interfaces (IUI-07), pages 278-281, Honolulu, Hawaii, USA, 2007.
5. B. Reeves and C. Nass. **"The Media Equation: How People treat Computers, Television, and New Media like Real People and Places".** Cambridge University Press, New York, NY, USA, 1996.
6. Witten, I.H. and Frank, E. **"Data Mining: Practical Machine Learning Tools and Techniques"**,(2nd Edition), Morgan Kaufmann, San Francisco, 2005.
7. Haji Binali, Chen Wu,VidyasagarPotdar, **" Computational Approaches for Character Detection in Text"**, 4th IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2010).