SindhUniv. Res. Jour. (Sci. Ser.) Vol.50 (001) 85-90 (2018)

https://doi.org/10.26692/surj-ss.v50i1.1299

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Online Text Categorization System Using Support Vector Machine

A. K. JUMANI, M. H. MAHAR, F. H. KHOSO*, M. A. MEMON**

Department of Computer Science, Shah Abdul Latif University, Khairpur Mirs Pakistan

Received 10rd April 2017 and Revised 26th December 2017

Abstract: Text Classification is a need of day, large text existing in the form of stories, news etc. Likewise, this system came into being along several techniques like, Support Vector Machine, Neural Networks and Decision Tree. Stories, newspapers are the page collection that belongs to text categorization. Various Sindhi newspapers are regularly published and Daily Kawish is one of them. People are facing difficulties during reading newspaper because there is no any specific option that will categorize particular news related to sports, technologies, crime, fashion and current affairs. For this purpose, a Text Categorization System (TCS) for Sindhi language is presented in this paper. Five classes are used and scanned each newspaper page inside a single class. It is too difficult to predict how many users will read newspaper simultaneously and for this, web performance is tested. Moreover, for the classification of the text from pages, precision, recall and f-measure are used to measure and achieved 67% of accuracy to classify the text from newspaper pages. It would be beneficial for those who want to save their precious time during reading newspaper.

Keywords: Online categorization; Support vector machine, Classification of text

1. INTRODUCTION

Sindhi language is a blend of many different cultural languages and lot of its words are carved in Arabic script (Odeh, 2014). Sindh is a well-known province of Pakistan which consists of several cultures and rich historical background relates to it. One of the most interesting facts about Sindhi language is that every word ends with a vowel, that's why, this pronunciation sounds more pleasant. Another interesting thing about Sindhi language is that it reflects many different cultures of Sindh very impressively. It proves that Sindhi language is spoken in Sindh from thousand years (Mutee, 2010). Many of the poets and saints wrote beautiful poetry in Sindhi language and so this language contains the treasure of literary works-poems novel, essays, poetry, and short stories. Sindhi language possesses a wide range of vocabulary that makes it colorful. Basically, Sindhi language has treasure from pre-historic age and plays a dynamic role in cultural inheritance of Sindh valley. Practically, all the people of this valley speak this language as their mother tongue and some people do not think that Sindhi language is our mother tongue and they avoid speaking it with other people.

The dialect was developed right around two thousand years back and now it has an extraordinary significance in the country with other provincial languages. The fundamental words are also called basic structure of Sindhi which consists of Sanskrit and Prakrit words and many other languages that make it an well-known language is spoken also by the people of Baluchistan. Several books are found in Sindhi language which prove it very ancient language of Sindh and one always loves to read it. The well-known saints and poets, who contributed Sindhi language in past times, which were Shah Abdul Latif Bhittai, Qalandar Lal Shahbaz and Sachal Sarmast (Chandio, 2016). The relation between Man and Allah has been perfectly clarified in their compositions. Many of the changes have been made in Sindhi with the passage of time; and now a days most of the people use Urdu words in Sindhi speaking but pure Sindhi language is spoken in rural areas of Sindh.

Text categorization of Sindhi text newspaper, SVM can efficiently work and give more than accurate results of online news corpus.

Moreover, such kind of approaches can be worked as symbolic labels and other things use a training set, which contains previous documents assigned to the target categories by experts. The disadvantage of these methodologies is that the classifier execution depends seriously on the extensive quantity of hand labeled records as they are the main source of information for taking in the classifier. Being a work focused on labor and tedious action, the manual work of documents to categorize is extremely expensive.

Furthermore, the text classification depends on the data that can be mined from the text documents which is related to set of text contents where content can be

*Department of Basic Sciences, Dawood University of Engineering and Technology, Karachi Pakistan[2]

**Department of Computer Science, Benazir Bhutto Shaheed University, Layari, Karachi[3]

recognized to some classification, so it can be related to personal decision of a human classifier. These techniques require a training set of pre-grouped records and it is frequently the case that an appropriate arrangement of well classified, normally by people, training contents is not accessible. This makes genuine restrictions for the convenience of the above learning methods in a few operational situations extending from the management of web-contents to the characterization of future news into classes, for example, business, sports, governmental issues, technologies, crimes, fashion and so on (Faraz, 2015). The importance of the Sindhi is one of the major reasons that driven us to do this research and provide an environment to news editors and readers respectively. Moreover, the primarily system came with several approaches in different languages but, no system in Sindhi language has been found.

The aim of this paper is to provide a mechanism of Sindhi Text categorization system that retrieves the news regarding concerned category. For this, Support Vector Machine is used as an approach. This system can be beneficial not only to editors, but impatient of readers as well. Hence, this system supports to all compatibility of the browsers and gives auto responsive interface.

2. SUPPORT VECTOR MACHINE

SVMs improved the quality performance of general dimension of feature vectors. SVMs could be taken learning abilities with their all functionalities without growing the computational difficulty by presenting the kernel function. Predictable algorithm cannot be controlled with these efficiently. Stochastic tagger can be required large number of annotated corpus. Stochastic taggers have better efficiency of 95% word-level accuracy for developing the German, English and European languages, for accessibility to large data. The main problems occur with Indian languages for privation of large annotated corpus (Fatima, 2017).

Support vector machines operate on numerical input vectors which means that biological data, be it sequences, structures, or other putatively important features, must be mapped to numerical values (Durgesh, 2010). This mapping plays an important role for the ultimate performance in the SVM. Hence, when SVMs are said to be deterministic, it should not be taken to imply that all SVMs produce the optimal classifier independent of the user's choices of several parameters of which the vector mapping is only one based on the input vectors and their assigned class membership, however, all SVMs try to find the maximum margin hyper plane the mathematical generalization of points in one dimension, lines in two dimensions, and planes in three dimensions that separates two classes (Sarkar, 2015). Simple working model of this approach is given in **(Fig.1)**.

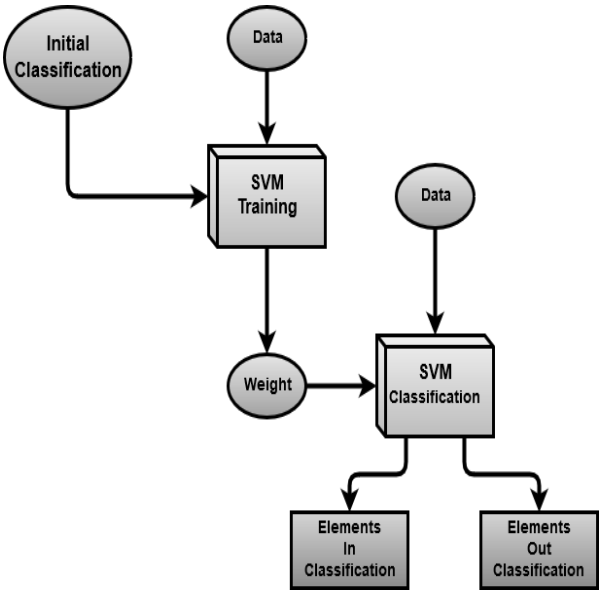

Fig.1 Support Vector Machine Model Execution

3. DEVELOPMENT OF SINDHI TCS

Development of software is the key bone for any developer that are being worked on different kinds of programming languages (Pandey, 2012). Similarly, in Pakistan dozen of software companies are available and working on either commercial project or natural language processing project for eliminating the issues of machine and tried to able them for understanding human language. So, in this chapter we develop an online Sindhi categorization text system that predict to category of the text belongs from newspaper. Likewise every software, the proposed work is also distributed into three major module including, User interface unit, server unit and data structure unit. Each description is given into sub sections.

User Interface (UI) Module

An interface of the application attracts to people eye and also gets attention to application due to its color and other mixed combination. It is said that user friendly environment is always admirable for users while they are working on it. So, we have also tried to give a user friendly interface to those who are access newspaper online. For this purpose, cascade style sheet is used to set the contents alignments and margins also with the placement of the object inside the class. These (Miraz, 2016) days, CSS3 is running successful with VS 2012 –2016 because most of the tags are used that only accepts CSS3 and thus, these are selected for accomplishing the task of designing.

Server Unit

At very first, internet information web server has installed to test the deployed web application and enable to investigate the incoming request from the un-reliable or reliable resource and purify them before giving desired results and provide sufficient information related to the newspaper category. It is centralized asset having all development tools like, C# language, style sheets and one major service database has been installed that is actually place where data would take place for manipulating and updating the records of the newspaper.

In this server, we have installed a programming language that supports to events and provides us built-in functions facility which enable us to use them, also, sessions are used to store temporary field of the visitors and displays information visited users to readers. Moreover, this session particular concern with the .aspx files and also these are used for log-in and out functions. Thus, several pages of the newspaper linked with to another page through the responsive redirect (Kanitkar, 2002) URL link and it helps to move the page from one to another. Furthermore, some additional methods are used to complete the task of the fetching from the database but with some limitations of the exception enforce us to add some ports before launching. The process of launching is then done by the visual studio. From the project properties, this is done along port information and server local host URL.

It is done via some functional methods providing by the VS and these methods have been used to design and developed TCS. List of the functional methods are given in **(Table-1)** showing along its description apparently. (**Table-1**) shows the methods that has used during development and execution of this system is given in **(Fig.2).**

Table-1 Used Methods with Description

Methods Name	Description
Is Not Post back	It redirects the pages from one to another.
Response redirect	It send request to server to move page from one to another.
Open	It helps to open database connection
Execute Non Query	It supports to execute query given with variable.
Close	This helps to close open database connection.
Command Execute	It supports to create new command for the query
Reader Execute	It helps to read record from the database
Command Execute Scalar	This helps to retrieve counter information that increment one by one.

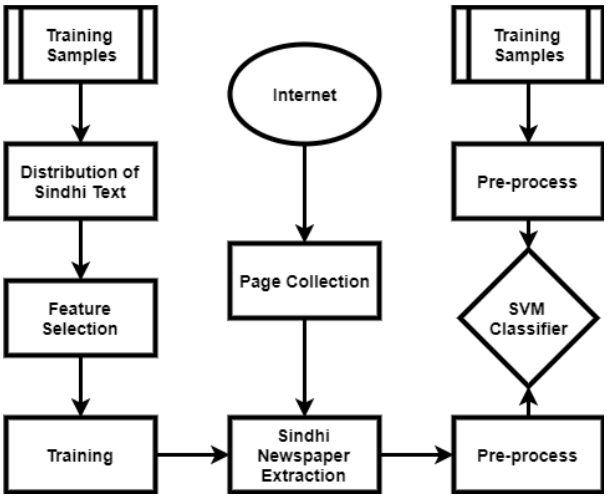

Fig.2 TCS Execution Process

4. RESULTS

To verify the system performance there is a need of results that shows the system ability and performance. To do this job, we have tested this system by using Support Vector Machine that classifies the newspaper category. Also, this online Sindhi text categorization system has implemented to achieve results as illustrative manner.

Data Sample

Data sampling is one of the best resource to get results and similarly, we have scanned 157 images of Kawish newspaper having different categories and chosen from different months. Among them, 34 images have selected from sports, 40 images from crime, 31 from fashion, 37 from current affairs and only 15 have selected from technology category. It is observed that Kawish newspaper has been publishing technology news not in a high amount.

These images have been captured through the digital camera. Digital camera based on 56 pixel and we believe it is feasible for better images quality. The concise summary of the newspaper or statistical are given in (**Table-2**).

Table-2 Statistical Information of Selected Samples

Categories	Selected samples
Crime	40
Current Affairs	37
Sports	34
Fashion	31
Technology	15
Total	**157**

(**Table-2**) shows the selected instances and their classification domains that have tested along SVM

approach. Through-out the precision, Recall and F-Measure, accuracy is measured along True (T) and False (F) parameters. Each parameter is given below;

Therefore, Precision, Recall, F-measure and accuracy are the measures to check the behavior of the tagged classified words based on 38000 and these measures are define as follows

a) Precision(P)=TP/(TP+FP) (1)

b) Recall(R)=TP/(TP+FN) (2)

c) F-measure=2(P*R)/(P+R) (3)*

Where true positive count as (TP) these are number of words tagged as tagboth in the test data, false positive count as (FP) words tagged as non tagi in the test set and as tagi by the marked, false negative as (FN) words tagged as tagi in the test set and as non tagi by the tagger and F-measure is the score that combine the two parameters.

The values of these measures lie between 0 and 1 as shown in **(Fig.3)**, we converted the values of obtained using equation 1 to 3 for this measure to percentage given below.

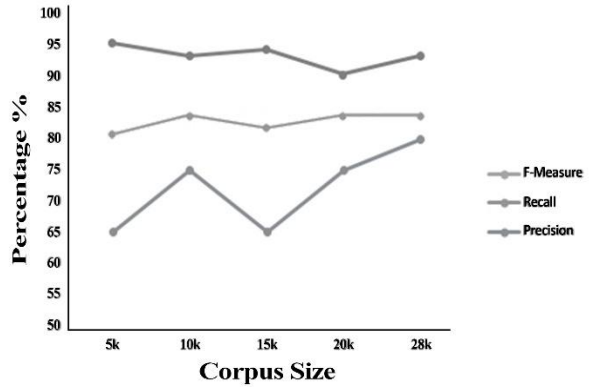

Fig.3. Calculated Precision, Recall and F-measure

Accuracy is the average number of correctly words which have tagged in the test data and the accuracy of the tagger is calculated with the help of the following equation.

*d) A=(N/T)*100 (4)*

Where A is the Accuracy and N is the number of words tagged correctly and T is the total number of words tagged. **(Fig.4)** shows the clear picture as we increase the corpus size the accuracy improves.

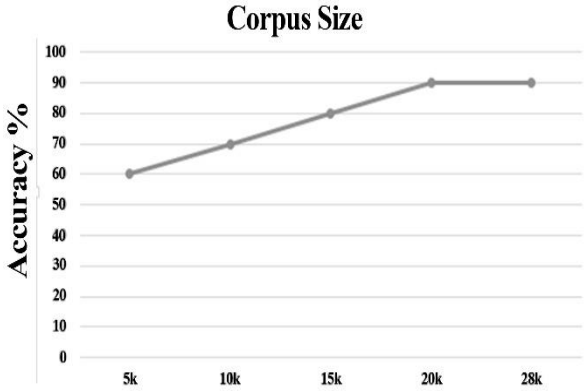

Fig.4. Achieved Accuracy with different Corpus Size

In this regard small size of all cases connects with every tag of all tagset which are not used and correctness of those related effects of the tagger. Due too many of the examples and training quantity of the tagger have ability to predict the right tag and overall performance of the tagger can be improves.

Cross validation can be measured the best performance which validates the estimation of the model with autonomous data set and make the prediction of correct model. Many of the cross validation techniques can be used like, K-fold, 2-fold, and leave one out cross validation etc. In this case we initialize the value of k as 5 i.e. the training set are divide into 5 smaller sets with equal sizes. In this method, performance of single subset as a validation data for testing the model and other K-1 are being used for training data. Whole procedure can be repeat K times with their subset of validate data. With the basis of predicted model give the average performance during k-iterations. During the testing proposed model gives 0.87 and 2.5 respectively.

5. CONCLUSION

In this paper, SVM is used as an approach to accomplish Sindh TCS under the development environment. It is a web application that provides facility to readers for retrieving news related to categories including, sports, fashion, crime, current affairs and technology. Moreover, for the measuring accuracy of this system three measuring parameters like, precision, recall and f-measure has used to figure out classified text accuracy 38000 words have collected as corpus and number of tagged correctly is depicted with T and negative considered as incorrectly depicted N. So, via the A=(N/T)*100 are the equation of the given system. This research would be highly beneficial for readers during classification of the news.

REFERENCES:

Chandio, A. A., M. Leghari, D. Hakro, S. Awan, and A. H. Jalbani, (2016). A Novel Approach for Online Sindhi Handwritten Word Recognition using Neural Network. Sindh University Research Journal (Science Series), 48(1), 213-216.

Durgesh, K. S. and B. Lekha, (2010). Data classification using support vector machine. Journal of Theoretical and Applied Information Technology, 12(1), 1-7.

Faraz, A. (2015). An Elaboration of Text Categorization and Automatic Text Classification through Mathematical and Graphical Modelling. Computer Science & Engineering: An International Journal (CSEIJ), 5(2/3), 1-11.

Fatima, S., and B. Srinivasu, (2017). Text Document categorization using support vector machine. International Research Journal of Engineering and Technology (IRJET), 4(2), 141-147.

Kanitkar, V., and A. Delis, (2002). Real-time processing in client-server databases. IEEE transactions on computers, 51(3), 269-288.

Leghari, M., M. U. Rahman, (2010). Towards Transliteration between Sindhi Scripts by using Roman Script. In Conference on Language and Technology, Lahore, Pakistan.

Miraz, M. H., P. S. Excell, and M. Ali, (2016). User interface (UI) design issues for multilingual users: a case study. Universal Access in the Information Society, 15(3), 431-444.

Odeh, A., A. Abu-Errub, and N. Turab, (2014). Arabic text categorization algorithm using vector evaluation method. International Journal of Computer Science & Information Technology, 6(6), 83-92.

Pandey, A., and S. Shrotriya, (2012). Development of Natural Language Processing Library in Nemerle using Dotnet Framework. International Journal of Scientific and Research Publications, 2(11), 1-6.

Rahman, M. U. (2010). Towards Sindhi corpus construction. In Conference on Language and Technology, Lahore, Pakistan.

Sarkar, A., S. Chatterjee, W. Das, and D. Datta, (2015). Text Classification using Support Vector Machine. International Journal of Engineering Science Invention, 4(3), 33-37.