Volume 7, Issue 4, April 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Identification of User's Behaviour on Web

Aishwarya Deep Rastogi[1], AakarshitAgarwal[1], Ajay Pratap Singh Tomar[1], Chirag Goyal[1], Nidhi Tyagi[2]
[1] Student, MIET-Meerut, Uttar Pradesh, India
[2] Professor, MIET-Meerut, Uttar Pradesh, India

Abstract: ***The exponential increase in the data on web has increased the need of classification techniques to identify user's behaviour on freely available user generated data available on different social sites. The paper discusses the methodology to identify the user's behaviour on Twitter. The different categories in which users are recognized are Politician, Food, Law, Health, Economy, Education, Banking, NGOs, Counsellors, Astrologist, etc. Implementation of the different techniques has used Machine Learning algorithms like Naïve Bayes Classification, and SVM (Support Vector Machine) along with natural language processing techniques for normalizing the raw data.***

Keywords: Naïve Bayes, SVM, Natural Language Processing

I. INTRODUCTION

User data classification is a process where user's data on web is classified into several categories to determine his behaviour or to simply classify his data into tags or known labels. For instance a user's twitter data or tweets can be used to classify his behaviour into known classes such as Politician, Food, Law, Health, Economy, Education etc.

Naïve Bayes [4] [5] Classifier-In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

SVM[7] -In machine learning , support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

II. LITERATURE REVIEW

AkshayRampuria, AnunayKulshrestha, AdityaRamakrishnan[1] investigated the problem of classifying user behaviour using freely available user generated data like tweets on Twitter, reviews on Amazon and eBay etc. Researchers follow the primary 25-label categorization that Amazon uses to bucket their products. By parsing a user's tweets, their algorithm attempts to predict which of the 25 categories the user is referring to.

Ashraf M. Kibriya, Eibe Frank, Bernhard Pfahringer, Geoffrey Holmes[2] shows how the performance of multinomial naive Bayes can be improved using locally weighted learning. However, the overall conclusion of their paper is that support vector machines are still the method of choice if the aim is to maximize accuracy.

Marcelo Maia, Jussara Almeida, and Virg´ılio Almeida[3] In this paper, researchers propose a methodology for characterizing and identifying user behaviours in online social networks. First, data is crawled from YouTube and using a clustering algorithm to group users that share similar behavioural pattern. Next, it is shown that attributes that stem from the user social interactions, in contrast to attributes relative to each individual user, are good discriminators and allow the identification of relevant user behaviour.

III. PROPOSED WORK

In the research work, classification algorithms for identifying internet user behaviour with classification techniques are implemented and compared. The entire process can be divided into three major steps, pre-processing, patterns discovery and finally prediction of test data labels [6].

Web data analysis includes the transformation and interpretation of web data, find out the information, patterns and knowledge discovery. But data is available in raw form so it becomes necessary to first normalize it using natural language processing and regular expressions. This pre-processed data is then fed into the classifier for training. The efficiency of the algorithm is analyzed by considering certain parameters.Fig.1 shows the flow diagram of the classification process.

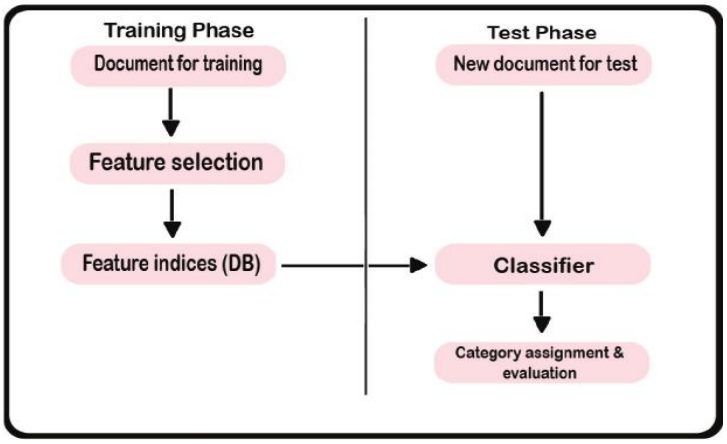

Fig.1: Flow Diagram of Classification Process

The purpose of this paper is to identify the User Behaviour using their tweets on the web. It provides a platform where organisation can classify Web Users in different categories, and utilize this information to predict his future behaviour.

ALGORITHM

Step1. Fetch web data i.e. label specific documents from various sources for instance ,UCI Machine Learning Repository.
Step2. Apply feature selection and extraction for extracting essential features such as tweet id, text and label.
Step3. Normalize the data removing all hash tags, handles,stop words etc.
Step4. The normalized data is partitioned in 3:2 ratio of trainingand testing dataset respectively.
Step5. The system is trained with training dataset,passing its text andclass labelinto the classifier (Naïve Bayes or SVM).
Step6. After the system is trained, test dataset is tested with the test dataset. The system is also evaluated for its accuracy.
Step7. Finally, user's label is predicted by entering the user's twitter handle.

Fig. 2: Classification Algorithm

IV. RESULTS

The algorithm is applied and implementation is done in Python [8]. The result based on the technique is shown in figures 3 and 4 respectively.

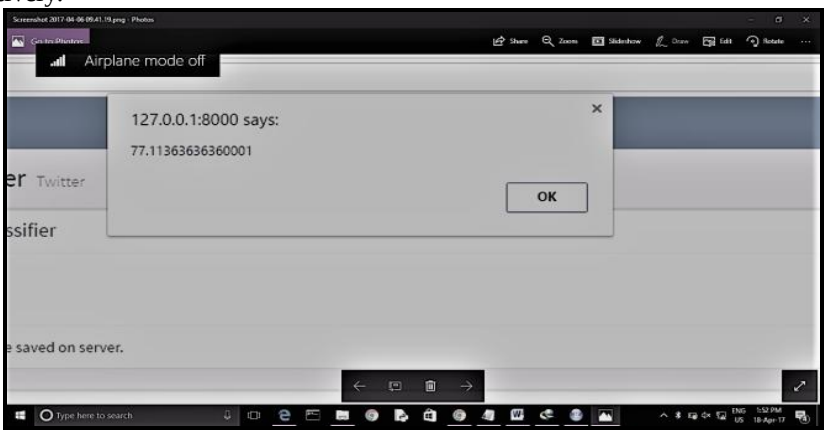

Fig.3: Accuracy of Naïve Bayes Classifier

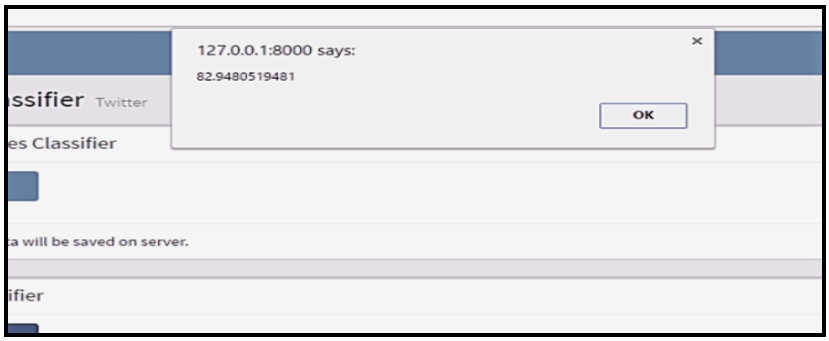

Fig.4: Accuracy of Support Vector Machine

On submitting the tweets related to Supreme Court, the predicted label is 'Law', as shown in figure 5

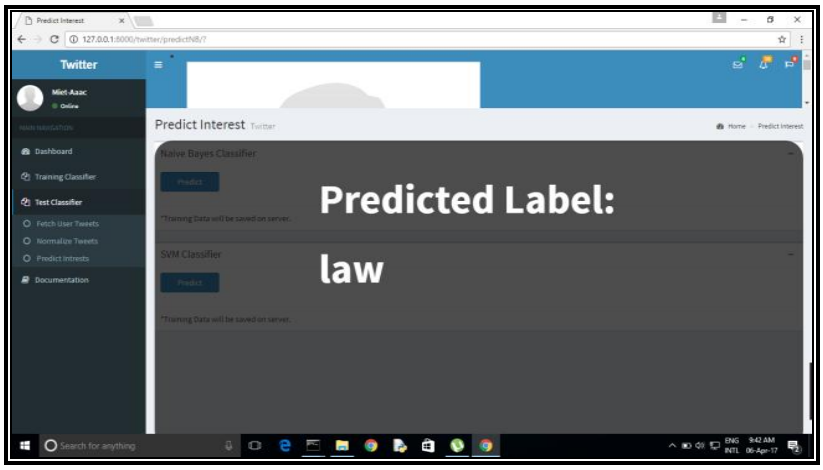

Fig.5: For user text supreme court the predicted label is law

V. CONCLUSION

The paper aims at classifying the web user's data in the 11 categories. This user classification information could be used by the organization for their benefit. Different data cleaning functions involving the use of regular expressions are utilised along with different types of classification methods have been used in this paper.

Prominently, naïve based classifier and SVM are used for the classification of the User in the different categories. The main focus of paper being the accurate classification of the user. The 11 categories in which user will be classified are Economy, Education, Entertainment, Food, Law, Health, Lifestyle, Nature, Politics, Sports and Technology.

The **naïve based classifier** gives **80.233%** accuracy and the **Support Vector Machine** gives **89.33%** accuracy for the classification of the Web User.

VI. FUTURE WORK

Many different adaptations, tests, and experiments have been left for the future due to lack of time. Future work concerns deeper analysis of particular mechanisms, new proposals to try different methods, or simply curiosity. There are some ideas that could be tested: (i)The number of classifiers could be increased.(ii)The categories of the Web User classification could be increased from 11.(iii)The accuracy of the classifiers could also be increased by increasing the size of training dataset.(iii)The algorithm can be modified for shopping sentiment analysis by retrieving data from online shopping sites like e-bay and amazon.

REFERENCES

[1] Akshay Rampuria, Anunay Kulshrestha, Aditya Ramakrishnan, "*Classifying Online User Behavior Using Contextual Data"*. Stanford University,2014.

[2] Kibriya, Ashraf M., et al., "*Multinomial naive bayes for text categorization revisited"*. Springer Berlin Heidelberg, pp488-499, 2005..

[3] Marcelo Maia, Jussara Almeida, and Virg´ılioAlmeida, " *Identifying user behavior in online social networks"pp*1--6. *(*2008*).*

[4] Rennie, Jason D., et al., "*Tackling the poor assumptions of naive bayes text classifiers*" Proceeding of International Conference on Machine Learning, Vol. 3, 2003.

[5] McCallum, A., Nigam, K., " *A comparison of event models for naive Bayes text classification"* Technical report, American Association for Artificial Intelligence Workshop on Learning for Text Categorization ,1998.

[6] Witten, I., Frank, E, " Data Mining: Practical machine learning tools and techniques with Java implementations." Morgan Kaufmann Series, San Francisco, 1999.

[7] Joachims, T, "*Text categorization with support vector machines: Learning with many relevant features",* In: Nédellec C., Rouveirol C. (eds) Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 1398. Springer, Berlin, Heidelberg

[8] Python libraries :scikit-learn, numpy, scipy, re.

