

Classification of a Persons Behavior by Using Web Content Mining

Ms.S.U.Revankar¹,Mr. S. S. Gurav²

Assistant Professor, E&TCE,SITCOE Yadrav,Ichalkaranji,India¹

Assistant Professor, E&TCE,SITCOE Yadrav,Ichalkaranji,India²

Abstract: In today's world, social media is used by every individual for expressing their feelings, opinion, experiences and emotions. People feel free to discuss and share their experiences on social media. We are intended to take people posts, likes, and comments from these social media and apply web content mining to the data collected for finding the person's behavior. Earlier there was a system which does web content mining on twitter data. But in our project we are going to take collective data from social media like Face book, Twitter, Google and so on.

Keywords: Java Netbeans IDE, MySQL, Xamp Server , Bayesian Equation.

I. INTRODUCTION

In today world we mostly use websites for Social Networking, Education, Marketing, Entertainment, Business, Shopping, and so many other things for making life easy. Now a day, social media craze is mounting to heights of success for every individual. People just comment, share, like and post their feelings on social media like Twitter, Face book and Google etc. These social media provides so many useful knowledge and information about the people emotions, feelings and experiences. We have designed a system which allows you to find behavior of person. To know person very well we need to spend lot of time with him thought not everyone is able to judge person behavior. Our software finds person behavior within seconds by using his social media comments. Accuracy of result is almost 70-80%.

II. LITERATURE SURVEY

Priyanka Patel, Ms. KhushaliMistry, [1], in existing system, only the prominent themes are identified with relatively large number of tweets in the data. There are a variety of other issues hidden in the others theme. Several of these issues may be of great interest to education researchers and practitioners. The fact that the most relevant data which are found on engineering students' learning experiences involve complaints, issues, and problems does not mean there are no positive sides in students' learning experiences'. This may entail that social media serve as a good venue for students to utter negative emotions and seek social support. In proposed work, New Label Good Things introduce and using the probability, the common keywords are considered for this label. Doing this, the proposed system is identifying and classifying the e-learning problems faced by student to improve their education quality with respect to their good and positive comments. Naive Bayes multilevel classifier is used for classification of experiences' by finding the probability of words in tweet for each category probability of each label contains how many users. Finally the tweets with new Label will be compared with the rest of the tweets with existing Labels.

Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, AnkitAgrawal, and AlokChoudhary [2], we construct word vectors with trending topic definition and tweets, and the commonly used tf-idf weights are used to classify the topics using a Naive Bayes Multinomial classifier. In network-based classification method, we identify top 5 similar topics for a given topic based on the number of common influential users. The categories of the similar topics and the number of common influential users between the given topic and its similar topics are used to classify the given topic using a C5.0 decision tree learner. Experiments on a database of randomly selected 768 trending topics (over 18 classes)show that classification accuracy of up to 65 % and 70% can be achieved using text-based and network-based classification modeling respectively.

Mr.Sachin Gurav [4] suspicious activity detection from surveillance video is the main objective of the work presented in this paper. The method developed consist of various stages of suspicious frame detection, and verifying the frame for suspicious activity related analysis of human movements within obtained set of suspicious frames. The method consist of GLCM feature extraction which constitutes the features such as energy, prominence, contrast, entropy, homogeneity type of features and matching using Euclidian distance along with descriptor features obtained by using Harris corner features and cosine similarity index estimation. The successful suspicious activity detection rate is analyzed which shows better performance and time saving method while analyzing large surveillance video dataset.

III. PROPOSED METHOD

Algorithm and Analysis of Core System Bayes Classification That was a visual intuition for a simple case of the Bayes classifier, also called:

Idiot Bayes , Naive Bayes ,Simple Bayes.

We are about to see some of the mathematical formalisms, and more examples, but keep in mind the basic idea. Find out the probability of the previously unseen instance belonging to each class, and then simply pick the most probable class. Bayesian classifiers use Bayes theorem, which says,

$$p(c_j | d) = p(d | c_j) p(c_j) p(d) \text{ ----- (1)}$$

$p(c_j | d)$ = probability of instance d being in class c_j . This is what we are trying to compute.

$p(d | c_j)$ = probability of generating instance d given class c_j , We can imagine that being in class c_j , causes you to have feature d with some probability.

$p(c_j)$ = probability of occurrence of class c_j , This is just how frequent the class c_j , is in our database.

$p(d)$ = probability of instance d occurring This can actually be ignored, since it is the same for all classes Assume that we have two classes $c_1 = \text{male}$, and $c_2 = \text{female}$.

We have a person whose sex we do not know, say drew or d . Classifying drew as male or female is equivalent to asking is it more probable that drew is male or female, i.e which is greater $p(\text{male} | \text{drew})$ or $p(\text{female} | \text{drew})$.

What is the probability of being called "drew" given that you are male?

What is the probability

$$\frac{p(\text{male} | \text{drew}) = p(\text{drew} | \text{male}) p(\text{male})}{p(\text{drew})}$$

What is the probability of being named "drew"?

$$\frac{p(c_j | d) = p(d | c_j) p(c_j)}{p(d)} \text{ -----(2)}$$

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(\text{male} | \text{drew}) = \frac{1/3 * 3/8}{3/83/8} = 0.125$$

$$p(\text{female} | \text{drew}) = \frac{2/5 * 5/8}{3/83/8} = 0.250$$

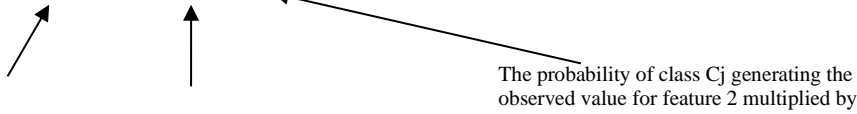
So far we have only considered Bayes Classification when we have one attribute

$p(c_j | d) = p(d | c_j) p(c_j) p(d)$ (the antennae length, or the name).

Name	Over 170cm	Eye	Hair Length	Sex
Drew	No	Blue	Short	Male
Claudia	Yes	Brown	Long	Female
Drew	No	Blue	Long	Female
Drew	No	Blue	Long	Female
Alberto	Yes	Brown	Short	Male
Karin	No	Blue	Long	Female
Nina	Yes	Brown	Short	Female
Sergio	Yes	Blue	Long	Male

To simplify the task, naive Bayesian classifiers assume attributes have independent distributions, and thereby estimate.

$$P(p|df) = p(d1|cf) * p(d2|cf) * \dots * p(dn|cf)$$



The probability of class cf generating instance d equals to

The probability of class Cj generating the observed value for feature 1 multiplied by

The probability of class Cj generating the observed value for feature 2 multiplied by

To simplify the task, naive Bayesian classifiers assume attributes have independent distributions, and thereby estimate

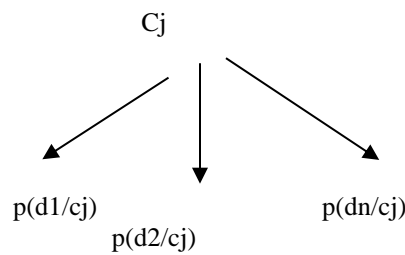
$$p(d|cj) = p(d1|cj) p(d2|cj) \dots p(dn|cj)$$

$$p(\text{officer drew} | cj) = p(\text{over } 170\text{cm} = \text{yes} | cj) p(\text{eye} = \text{blue} | cj)$$

$$p(\text{officer drew} | \text{Female}) = 2535$$

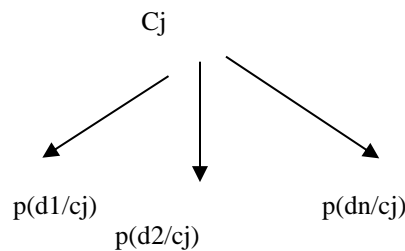
$$p(\text{officer drew} | \text{Male}) = 2323$$

The Naive Bayes classifiers are often represented as this type of graph. Note the direction of the arrows, which state that each class causes certain features, with a certain probability.



Naive Bayes is fast and space efficient

We can look up all the probabilities with a single scan of the database and store them in a (small) table.



Animal	Mass > 10kg	
Cat	Yes	0.15
	No	0.85
Dog	Yes	0.91
	No	0.09
Pig	Yes	0.99
	No	0.01

Animal	Color	
Cat	Black	0.33
	White	0.23
	Brown	0.44
Dog	Black	0.97
	White	0.03
	Brown	0.90
Pig	Black	0.04
	White	0.01
	Brown	0.03

Naive Bayes is NOT sensitive to irrelevant features. Suppose we are trying to classify a person's sex based on several features, including eye color. (Of course, eye color is completely irrelevant to a person's gender) .

$$\begin{aligned}
 p(\text{Jessica}|c_j) &= p(\text{eye}=\text{brown}|c_j) * p(\text{wears_dress}=\text{yes}|c_j) * \dots \\
 P(\text{Jessica}|\text{Female}) &= 9,000/10,000 * 9,975/10,000 * \dots \\
 P(\text{Jessica}|\text{Female}) &= 9,001/10,000 * 2/10,000 * \dots
 \end{aligned}$$

Almost same

Testing Methodology

Testing is dynamic method for verification and validation, where the system to be tested is executed and the behavior of the system is observed. Due to this, testing observed failures of the system, from which presence of fault can be detected.

Testing plays a critical role in quality assurance for software. Due to the limitations of the verification methods for the previous phases, design and requirement faults also appear in

The code. Testing is used to detect these errors, in addition to the errors introduced during

Coding phase.

There are three approaches to testing:

1. Black-box testing
2. White-box testing
3. Manual testing

Black-box testing:

In black-box system testing, the internal logic of the system under testing is not considered and the test cases are decided from the specifications or the requirements. It is often called as functional testing. Equivalence class partitioning, boundary value analysis and cause effect graphing are examples of methods for selecting test cases for black-box testing. State based testing is another approach in which the system is modeled as a state machine and then this model is used to select test cases using some transition or path based coverage criteria.

State-based testing can also be viewed as greybox testing in that it often requires more information than just the requirements.

Iterative Model:

Iterative Models any combination of both iterative design or iterative method and incremental build model for development. The combination is of long standing and has been widely suggested for large development efforts. During software development, more than one iteration of the software development cycle may be in progress at the same time and This process may be described as an 'evolutionary acquisition' or 'incremental build' approach. The relationship between iterations and increments is determined by the overall software development methodology and software development process. The exact number and nature of the particular incremental builds and what is iterated will be specific to each individual development effort.

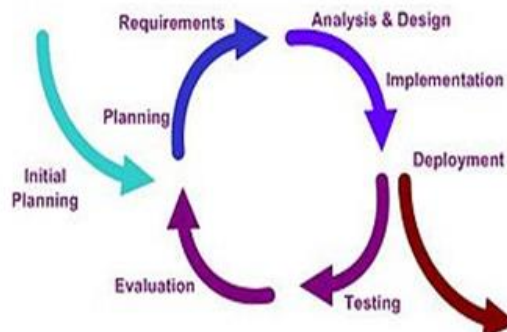


Fig. 1 Iterative Model

SRAS (Software Requirement Analysis Specification)

The implementation of proposed system is carried out using JAVA programming language and by using Netbeans editor.

1. Hashtable-

Hashtable was part of the original java.util and it implements the Map interface. Hashtable stores key or values pairs in table. When using a Hashtable , you specify an object that is used as a key, and the value you want to link that key. The key is then hashed, and the resulting hash code is used as the index at which the value is stored within the table.

Methods-

--Boolean contains (object value)

Return true if some value equal to value existing within the hash table. Returns false if the value does not found.

--String to String ()

Return the string equivalent of hash table.

2. String Tokenizer

String Tokenizer class contain in java.util package. This class is allowing you to break a string into tokens. It is simple way to break string.

Methods-

_ Boolean has More Tokens ()

Checks if there is more tokens available.

_ String next Token ()

Returns the next token from the String Tokenizer object.

3. Enumeration

This class contain in a java.util package. The Enumeration interface defines the method by which you can enumerate the elements in a collection of objects.

Methods-

---Boolean has More Elements ()

When implemented, it must return true while there are still more elements to extract, and false when all th elements have been enumerated.

--Object next Element ()

This returns the next object in the enumeration as a generic object reference.

4. J File Chooser

This class contain in a swing package. It extends from Component and implements Accessible .The class J File Chooser is a component which provides a simple mechanism for the user to choose a file.

Methods-

-- Int show Open Dialog (Component parent)

Pops up an 'Open File' file chooser dialog.

--File get Selected File ()

Returns the selected file.

5. Basic Stroke

This class contain in a java.awt package. It extends from object and implements Stroke. The Basic Stroke class states colors in the default RGB color space or colors in arbitrary color spaces identified by a Color Space.

Methods-

--Int hash Code ()

Returns the hashcode for this stroke.

--Shape create Stroked Shape(Shape s) Returns a Shape whose interior defines the stroked outline of a specified Shape.

IV. RESULTS AND ANALYSIS

1. GUI

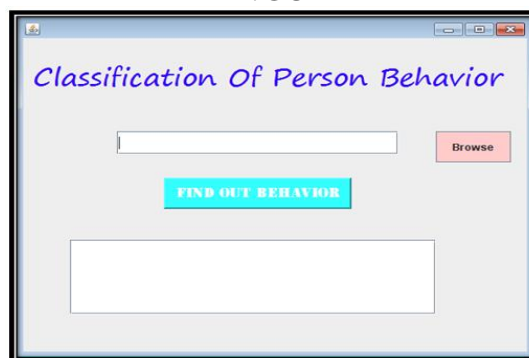


Fig.2 GUI

2. Browse the file of tweets you have downloaded through API

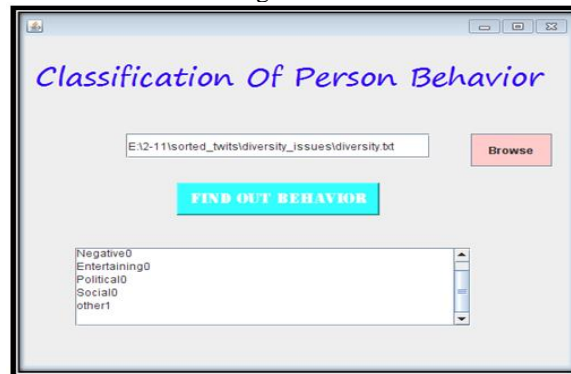


Fig.3 Browse the file of tweets you have downloaded through API.

3. Output will display the appropriate category to which person belongs. For better understanding result is shown in the form of graph. If user does not belong to any of the specific category then software will put that person in other category.

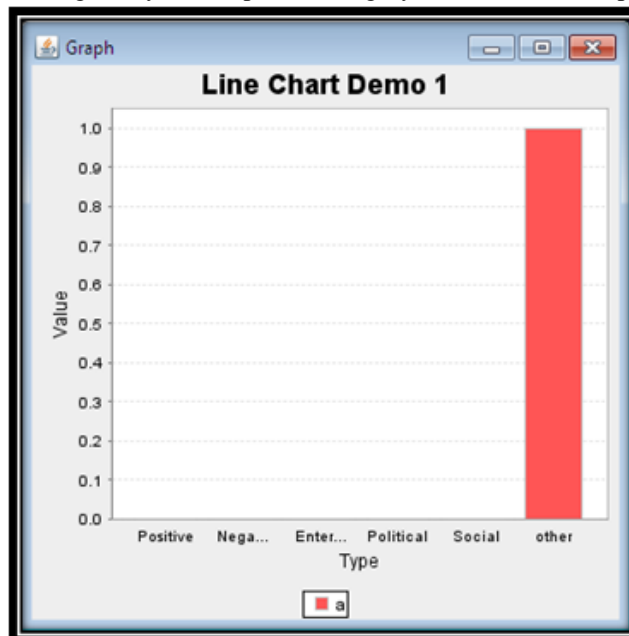


Fig.4 Output result

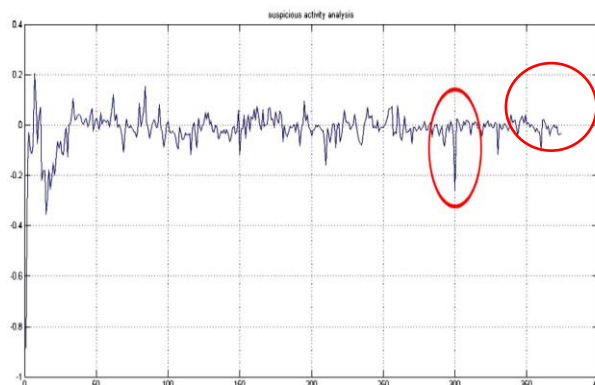


Fig.5 Graph of above table indicating classification of person behavior.

V. CONCLUSION

From Different papers only knowing the sentiments of person shared tweets do not provide much actionable knowledge in sentiment analysis tweets can be classified positive or negative or neutral but cannot be judge that is belongs to which categories so in our proposed system multilabel classification model is used which will make comments into multi categories at same time.

REFERENCES

- [1] Classification of Students E-Learning Experiences in Social Media via Text Mining Ms.Priyanka Patel1, Ms. Khushali Mistry (Department of CSE, PIET, Vadodara, India,(Department of CSE, PIET, Vadodara, India)
- [2]Twitter Trending Topic Classification- Kathy Lee, Diana Palsetia, Ramanathan Narayanan,Md. Mostofa Ali Patwary, AnkitAgrawal, and AlokChoudhary
- [3]Ajay Kumar Pal, Saurabh Pal, 'Classification model of Predication for Placement of Students,' in I.J. modern Education and Computer Science, November 2013.
- [4] Mr.Sachin Gurav,Dr.B.B.Godbole. "An Efficient Method for Suspicious Activity Detection" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-11, September 2019