I.J. Information Technology and Computer Science, **2016, 2, 69-77**
Published Online February 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2016.02.09

Web Video Mining: Metadata Predictive Analysis using Classification Techniques

Siddu P. Algur
Department of Computer Science, Rani Channamma University, Belagavi-591156, Karnataka, India
E-mail: siddu_p_algur@hotmail.com

***Prashant Bhat**
Department of Computer Science, Rani Channamma University, Belagavi-591156, Karnataka, India
E-mail: prashantrcu@gmail.com

***Abstract*—**Now a days, the Data Engineering becoming emerging trend to discover knowledge from web audio-visual data such as- YouTube videos, Yahoo Screen, Face Book videos etc. Different categories of web video are being shared on such social websites and are being used by the billions of users all over the world. The uploaded web videos will have different kind of metadata as attribute information of the video data. The metadata attributes defines the contents and features/characteristics of the web videos conceptually. Hence, accomplishing web video mining by extracting features of web videos in terms of metadata is a challenging task. In this work, effective attempts are made to classify and predict the metadata features of web videos such as length of the web videos, number of comments of the web videos, ratings information and view counts of the web videos using data mining algorithms such as Decision tree J48 and navie Bayesian algorithms as a part of web video mining. The results of Decision tree J48 and navie Bayesian classification models are analyzed and compared as a step in the process of knowledge discovery from web videos**.**

***Index Terms*—**Web Videos, Classification, Web Video Classification, J48 Classification, navie Bayesian Classification.

I. INTRODUCTION

The absurd rapid escalation of videos on the social websites such as YouTube, Yahoo Screen, and Face Book etc, it is becoming complex and challenging task for data engineers and researchers to mine knowledge from web videos. The web video mining can be accomplished by two major techniques- through image/signal processing and through metadata feature analysis of web videos. As a part of knowledge discovery from web videos, the classification of web videos is an increasingly outstanding area of research, growing with the quantity of videos shared through social sites such as YouTube, Yahoo Screen etc. As much as its importance, web based video classification poses serious challenges to computer vision researchers [1].

In this information age, web video data are very essential resource of information [2]. Different categories of web videos (e.g.- YouTube) are shared on social websites and used by the billions of users all over the world [3]. The shared web videos will have different kind of web metadata such as category, comment information, rating information, and view counts etc [4]. These different kind of metadata are referred as the features of the web videos. The classification/prediction and analysis of web videos in terms of such different kind of web metadata is a complex and challenging task. Many classification models/algorithms and data mining and machine learning tools are developed in recent years. Using different data mining algorithms and machine learning tools, it is possible to classify the web videos based on their features/metadata as a research problem. In this work, the web video metadata/features are extracted and effective attempts are made to classify/predict each of them using data mining algorithms such as J48 and navie Bayesian classification methods. Also in this work, we proposed a novel framework for metadata predictive analysis for web videos.

YouTube is one of the most popular and largest video sharing websites (with social networking features) on the Internet [4][5]. In this experiment, the YouTube video metadata are used for predictive analysis. According to official declaration by the YouTube authority [3], it has more than 1 billion users. Every day, people are spending hundreds of millions of hours on YouTube videos. More than 300 hours of video are uploaded to YouTube every minute, and also YouTube is available in 75 countries with 61 languages. Approximately 300 hours of video are uploaded to YouTube every minute [3]. This statistics shows that, how YouTube is popular day by day in a rapid increasing way. Hence, the social media researchers' are attracting towards YouTube video which contains huge unstructured complex data.

The rest of the paper is organized as follows: The section 2 represents related works on the classification of web videos, section 3 represents proposed web video classification methodology, section 4 represents performance evaluation analysis of classification models and comparison of efficiency of classification models, and finally section 5 represents conclusion and future enhancements.

II. Related Works

This section represents some related previous works which are implemented to classify web videos using metadata. The authors Amjad Mahmood, Tianrui Li, Yan Yang, Hongjun Wang and Mehtab Afzal [1], worked on categorization of web videos based on textual metadata. The proposed techniques to categorize web videos are based on Semi-supervised Evolutionary Ensemble (SS-EE) framework. For example, web video categorization using low cost textual features such as title, tags, descriptions etc, intrinsic relations and extrinsic web support. In order to implement the proposed technique, the traditional Vector Space Model(VSM) extended to Semantic VSM (S-VSM). Finally, Experiments on real world social-Web data (YouTube) have been performed to validate the SS-EE framework.

The authors Siddu P. Algur, Prashant and Suraj Jain[6], described implication of web video descriptive metadata, offered an effective and efficient method for construction and extraction of web video descriptive metadata. The proposed method established the effectiveness of constructing the descriptive metadata with timeline for a domain specific web video. The paper also suggested the construction of event specific and objects specific metadata and which are considered to be very useful. Using proposed descriptive metadata model, users may process the video contents effectively and efficiently.

The authors Polyxeni Katsiouli, Vassileios Tsetsos and Stathes Hadjiefthymiades [7], explored an unsupervised technique to classify video content by analyzing the matching subtitles. The proposed system is based on the WordNet lexical catalog and the WordNet domains and applies natural language processing techniques on video subtitles. The proposed method includes subtitle text preprocessing, a keyword extraction method, a word sense disambiguation technique and Subsequently, the WordNet domains that correspond to the correct word senses are identified. The final step allots category labels to the video content based on the extracted domains. The experimental result with documentary videos shows that the proposed method is effective in predicting the correct category for each web videos.

The authors Anil Kale and D.G. Wakde [8], proposed an automated video classification technique. This technique presents a model that provides automation of video classification and video annotation. The videos are classified and annotated on the keywords.

Bin Cui Ce Zhang and Gao Cong [9] proposed a novel video classification system which is able to utilize both content and text features for video classification while avoiding the pricey estimation of extracting content features at classification time. The proposed approach utilizes the content features extracted from training data to improve/enrich the text based semantic kernels, yielding content-enriched semantic kernels. The content-improved/enriched semantic kernels enable to utilize both content and text features for classifying new videos without extracting their content features. The experimental results show that the proposed technique significantly performs well on the state-of-the-art video classification methods.

Automatic categorization of videos in a Web-scale unconstrained collection such as YouTube is a challenging task. A key issue is how to build an effective training set in the presence of missing, sparse or noisy labels. In this regard, the authors Zheshen Wang, Ming Zhao, Yang Song, Sanjiv Kumar, and Baoxin Li [10], proposed to achieve this by first manually creating a small labeled set and then extending it using additional sources such as related videos, searched videos, and text-based web pages. The data from such disparate sources has different properties and labeling quality, and thus fusing them in a coherent fashion is another practical challenge. The proposed approach [10] uses a fusion framework in which each data source is first combined with the manually-labeled set independently. Then, using the hierarchical taxonomy of the categories, a Conditional Random Field (CRF) based fusion strategy has been designed. Based on the final fused classification model, category labels are predicted for the new videos. Extensive experiments on about 80K videos from 29 most frequent categories in YouTube show the effectiveness of the proposed method for categorizing large-scale wild Web videos.

The authors, Chunneng Huang, Tianjun Fu and Hsinchun Chen [11] proposed a text-based methodology for video content classification/categorization of internet-video sharing Web sites. Different kinds of user-generated data (for example- titles, descriptions, and comments) were used as proxies for internet videos, and three types of text features (syntactic, lexical and content-specific features) were extracted. Three feature-based classification techniques (Na ïve Bayes, C4.5 and Support Vector Machine) were used to classify internet videos.

III. Proposed Methodology

This section represents novel methodology of the proposed work. The web metadata of online videos are extracted using Info Extractor tool [12]. This metadata includes uploader information, category, comments, ratings, length of the video, descriptions about content of the video etc. We propose a novel and effective methodology to classify/predict the web video metadata features such as length, view counts, number of comments and rating information by applying data mining techniques. For experimental purpose, the metadata features are discretized and transformed to nominal values by ‘Equal Width Partitioning’ method, and out of the total metadata dataset, 60% are used for training and remaining 40% are used for testing the classification model built using Decision Tree J48 and navie Bayesian classification methods. The classification/prediction results of each considered metadata features are analyzed and the efficiency of the proposed method has been demonstrated. The system model of the proposed system is represented in Fig. 1, and it consists of the following components:

A) Web video metadata extraction
B) Metadata refinement
C) Classification models
D) Classification analysis

A) Web Video Metadata Extraction

The different categories of web videos such as 'Entertainment', 'News and Politics', 'Sports', 'People and Blogs' are randomly selected and given to the InfoExtractor tool, to extract different types of web metadata such as length, rating information, category, etc. The extracted raw metadata will be in the form of text and these metadata are then stored in a disk [13] with ARFF or CSV file format.

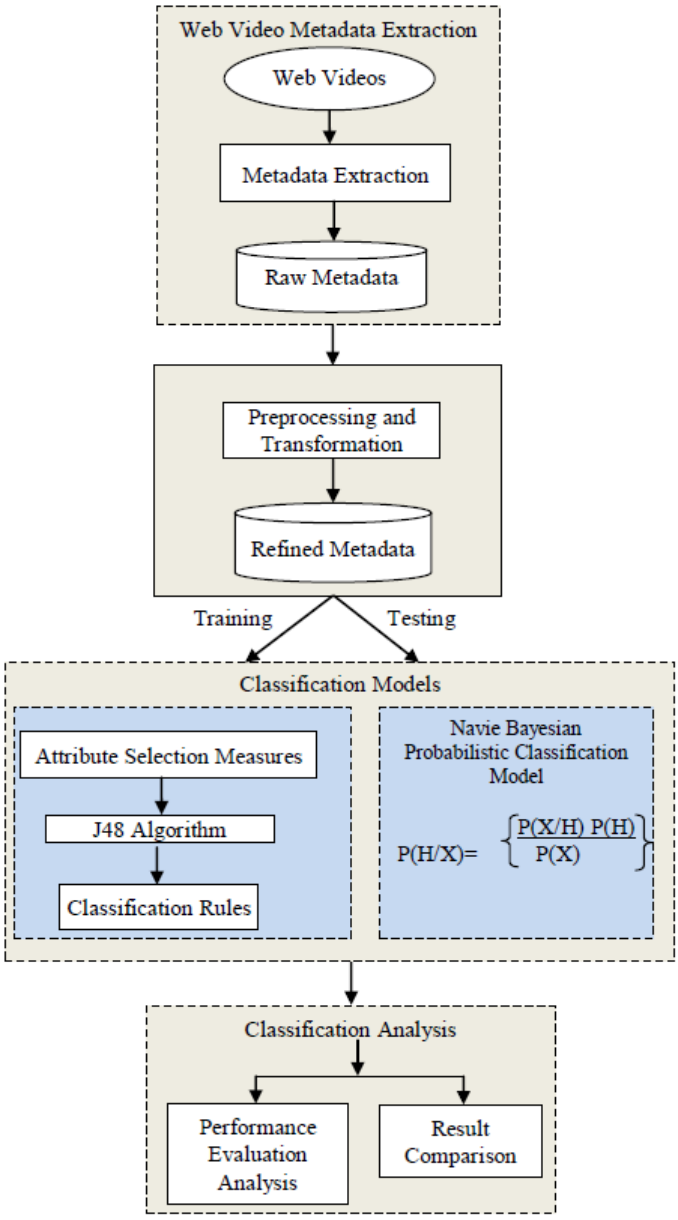


Fig.1. System Model of the Proposed Methodology

B) Metadata Refinement

The input to this component is raw metadata extracted from the web videos. This raw metadata has to be preprocessed for the refinement such as file format conversion and to identify the unimportant metadata. The WEKA tool is used for file preprocessing and to build classification model. The extracted raw metadata are converted to ARFF or CSV format from the text format for effective classification. Some web videos might have less metadata information, whereas some web videos might have more metadata information. Through observations, it is found that, all web videos contains minimum metadata information such as- length, view counts, ratings, average ratings and number of comments, author information and URL. Among these minimum metadata information of web videos, only the numeric and nominal attributes-length, view counts, ratings, average ratings, category and number of comments are considered for classification and structure of the dataset considered is represented in Table 1 [14].

Table 1. Structure of Web Video Metadata Dataset

Category	*Length*	*Views*	*Rate*	*Ratings*	*Comments*
People & Blogs	217	1157	3.6	5	3
Comedy	426	667	4	4	4
Entertainment	237	1063	4.8	30	10
Sports	294	274	1	1	2
……..	……	……	……	…….	…..
……..	……	……	……	…….	…..

The missing values are replaced by mean of each numeric attribute and the missing values of the attribute 'category' are replaced by most repeated nominal values. Since, the metadata attributes length, comments, rate, views, and ratings are extracted in the form of continuous/numeric values and hence, to improve the classification and prediction accuracy, data transformations are needed to transfer from continuous numeric values to nominal values. The considered numeric attributes are then discretized and transformed to nominal values by 'equal width partitioning' method. The Table 2 represents typical structure of transformed web video metadata dataset. The transformed metadata dataset are stored in a database [12] for classification.

Table 2. Structure of Transformed Web Video Metadata Dataset

Category	*Length*	*Views*	*Rate*	*Ratings*	*Comment*
People & Blogs	Medium	High	Medium	Low	Low
Comedy	Medium	Medium	High	Low	Low
Entertainment	Medium	High	Low	Low	Low
Sports	Medium	Low	Very High	Low	High
……..	……	……	……	….	…..

C) Classification Models

This component has two sub components:

i) J48 classification model
ii) The navie Bayesian probabilistic classification model.

The functionality of each subcomponent is discussed as follows.

- **Building the J48 classification model :**

The proposed J48 classification model consists of three major steps such as,

i) Attribute selection measures
ii) J48 algorithm
iii) Classification rules.

The efficiency of the classification result is largely depends on the classification model itself. Hence, construction of robust classification model plays important role in classification. The classification model construction for web videos are discussed in the following subsections.

i) Attribute Selection Measures:

The attribute selection measures provide a splitting criteria for each attribute describing the given tuples. The attribute selection measures for web video metadata are:

Information needed to identify the category of an element of a metadata tuple, Information gain of each attributes and Splitting criteria.

As discussed in the section 3.B, only five attributes with nominal class labels (outcomes) are considered for our study and are listed in Table 3.

Table 3. Attribute Selection for Classification

Sl.No	*Attributes*	*Descriptions*	*Class Labels (Outcomes)*
1	Length	Length unit of the web videos	Low Medium High Very High
2	Rate	Ratings given by the users	Low Medium High Very High
3	Views	View counts of the web videos	Low Medium High Very High
4	Ratings	No. of ratings of the web videos	Low Medium High Very High
5	Comments	No. of comments given by the users	Low Medium High Very High

The procedure to measure attribute selections for the web video metadata are discussed in our previous work [14]. According to attribute selection procedure discussed in [14], the information gain of each attribute will be calculated and the attribute which has highest information gain will be labeled as splitting node. The splitting criterion decides which attribute to test at each node by determining the best way to split or partition the tuples into different categories. Each selected attributes for the experiment can have maximum four outcomes at any given time as shown in Fig. 2.

- **Navie Bayesian Probabilistic Classification Model**

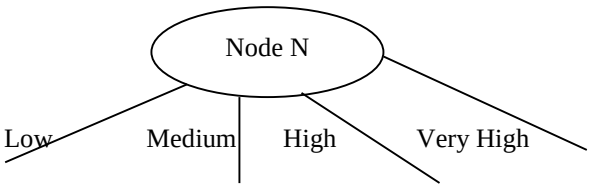


Fig.2. Labeling Root Node of the Tree

ii) J48 classification algorithm:

J48 is bespoke version of C4.5 classification algorithm. The J48 algorithm generates a classification-decision tree

for the web video metadata data-set by recursive partitioning the tuples. The decision tree is grown using depth-first strategy. The algorithm considers all the possible tests that can split the metadata data set and selects a test that gives the best information gain. For each metadata features of the web videos such as 'Length', 'Ratings', 'Comments' etc, binary tests involving every distinct values of the attribute are considered. In order to gather the information gain of all these binary tests efficiently, the information gain of the binary partition point based on each distinct values are calculated and sub trees are formed accordingly. This process is repeated for each attributes considered for classification.

iii) Classification Rules

A part or segment of the Decision Tree structure of J48 classification model for the dataset chosen is represented in Fig. 3, where the leaves node represents class labels of the attribute 'Length'.

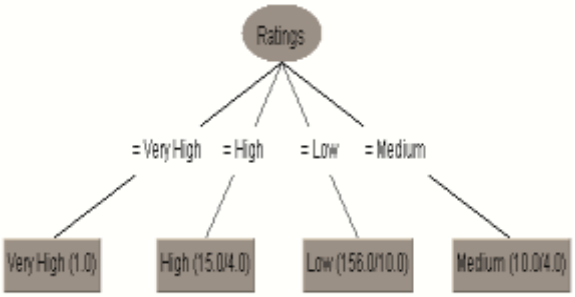


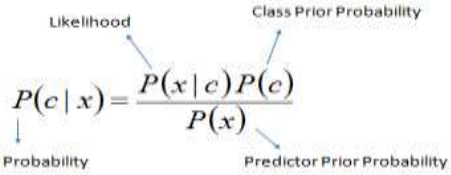
Fig.3. Tree Structure Result of J48 Classification Model

The above tree can be converted to classification rules by traversing the path from root node to each leaf node in the tree. The j48 classification rules extracted from Fig. 2 and are represented in the form of 'pruned tree' as follows:

```
J48 pruned tree
------------------
Ratings = Very High: Very High (1.0)
Ratings = High: High (15.0/4.0)
Ratings = Low: Low (156.0/10.0)
Ratings = Medium: Medium (10.0/4.0)
Number of Leaves:  4
Size of the tree:  5
```

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$, where, P(c|x) is the posterior probability of class (target) predictor, P(c) is the prior probability of class, P(x|c) is the likelihood which is the probability of predictor given class, and P(x) is the prior probability of predictor.

Naive Bayes classifier assumes that the effect of the value of a predictor x on a given class c is independent of the values of other predictors. This assumption is called class conditional independence. The Navie Bayesian probabilistic classification model can be described as-



$$P(c \mid \mathbf{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Using the described probabilistic model, the class labels of the test dataset will be predicted

D) Classification Analysis

In this section, performance evaluation measures such as TP, FP, precision, recall and F-Measure will be calculated to measure classification efficiency of J48 and navie Bayesian classification model. Also the classification/prediction efficiency of each attribute considered for the experiment will be compared with respect to J48 and navie Bayesian classification models.

IV. Experimental Results and Discussion

A) Classification using J48 and Navie Bayesian Classification Model

To test the efficiency of the classification models built using J48 and navie Bayesian algorithms, the dataset is downloaded from the website [15] which consists of 182 web video metadata instances. The performance of the both the models is measured in terms of number of correctly classified instances, number of incorrectly classified instances, TP rate, FP rate, precision, recall and F-score. The Table 4 represents classification performance evaluation metrics obtained by the J48 and navie Bayesian classification models. The performance evaluation of J48 and navie Bayesian (NB) classification models on the class label prediction of each attributes is discussed in the following subsections.

B) Class Label Prediction of 'Length' Attributes

In the prediction of the class labels of the 'Length' attribute, 165 metadata instances are correctly classified and 17 instances are incorrectly classified by the J48 classification model, whereas, 156 metadata instances are correctly classified and 26 instances are incorrectly classified by the navie Bayesian classification model. This statistics shows that, the J48 classifier predicted class labels more correctly as compared to navie Bayesian classification model. However, the classification efficiency of the J48 classification model is 86.2%, and efficiency of the navie Bayesian classification model is 86.8%. This is because, the J48 classification model incorrectly classified all the instances which are belongs to the class labels 'Medium' and 'High'.

Table 4. Classification Result of J48 and Navie Bayesian Classification Models

l.No	*Prediction Attribute*	*Classification Model*	*Total Instances: 182*		*TP*	*FP*	*Precision*	*Recall*	*F-Score*
			Correctly Classified	*Incorrectly Classified*					
1	Length	J48	165	17	0.907	0.907	0.822	0.907	0.862
		navie Bayesian	156	26	0.857	0.485	0.885	0.857	0.868
2	Views	J48	155	27	0.852	0.069	0.805	0.852	0.818
		navie Bayesian	149	33	0.819	0.181	0.795	0.819	0.797
3	Rate	J48	153	29	0.841	0.121	0.75	0.841	0. 79
		navie Bayesian	156	26	0.857	0.1	0.81	0.857	0.827
4	Ratings	J48	165	17	0.907	0.431	0.856	0.907	0.88
		navie Bayesian	165	17	0.907	0.17	0.912	0.907	0.909
5	Comments	J48	164	18	0.901	0.262	0.89	0.901	0.893
		navie Bayesian	166	16	0.912	0.084	0.916	0.912	0.913

Table 5. Confusion Matrix for 'Length' Prediction

Confusion Matrix

J48 Classification Model ===Confusion Matrix===				NB Classification Model ===Confusion Matrix===			
a	**b**	**c**	**←Classified as**	**a**	**b**	**c**	**←Classified as**
149	10	6	a=Low	165	0	0	a= Low
9	5	1	b=Medium	15	0	0	b=Medium
0	0	2	c=High	2	0	0	c=High

Hence, the FP rate of J48 classification model became high which will lead to reduce in the classification efficiency even though it has more number of correctly classified instances as compared to navie Bayesian (NB) classification model. The comparative analysis of classification result is represented in the form of confusion matrix as shown in the Table 5:

In this confusion matrix, the column 'a' and row 'a' corresponds to the class label 'Low', column 'b' and row 'b' corresponds to the class label 'Medium' and so on. The J48 classifier correctly predicted all the web video tuples which are belongs to the class label 'Low' and incorrectly classified all the tuples of the class labels 'Medium' and 'High'. The all the tuples of the class 'High' and 5 tuples of the class 'Medium' are correctly classified by NB classifier. This analysis shows that, the NB classification model is more efficient to predict the class label of 'Length' for web videos.

C) Class Label Prediction of 'View' Attributes

In the prediction of the class labels of the 'View' attribute, 155 metadata instances are correctly classified and 27 instances are incorrectly classified by the J48 classification model, whereas, 149 metadata instances are correctly classified and 33 instances are incorrectly classified by the navie Bayesian classification model. This statistics shows that, the J48 classifier predicted class labels more correctly as compared to navie Bayesian classification model. Also the classification efficiency of the J48 classification model is 81.8%, and efficiency of the navie Bayesian classification model is 79.7%. The comparative analysis of classification result is represented in the form of confusion matrix as shown in the Table 6:

All the web video tuples which are belongs to the class label 'Very High' are correctly predicted by the J48 classifier, and 14 web video tuples are correctly predicted by the NB classifier. And both classifiers incorrectly classified all the web video tuples of the class label 'Medium' and predicted as 'High' and 'Very High'. This analysis shows that, the J48 classification model is more efficient to predict the class label of 'View' for web videos as compare to NB classification model.

D) Class Label Prediction of 'Rate' Attributes

In the prediction of the class labels of the 'Rate' attribute, 153 metadata instances are correctly classified and 29 instances are incorrectly classified by the J48 classification model, whereas, 156 metadata instances are correctly classified and 26 instances are incorrectly classified by the navie Bayesian (NB) classification model. This statistics shows that, the NB classifier predicted class labels more correctly as compared to J48 classification model. Also the classification efficiency of the J48 classification model is 79.3%, and efficiency of the navie Bayesian classification model is 82.7%. The comparative analysis of classification result is represented in the form of confusion matrix as shown in the Table 7.

All the 97 web video tuples which are belongs to the class label 'Low' are correctly classified by the both classifiers as shown in the confusion matrix of Table 7. Also, the both J48 and NB classification models incorrectly classified all the web video tuples which are belongs to the class label 'High' and wrongly predicted as 'Very High'. Out of 65 web video tuples which are

belongs to the class labels 'Very High', 56 tuples are correctly classified by the both classifiers. Also few web video tuples of class label 'Medium' are correctly classified by the NB classification model, whereas, all the tuples of 'Medium' are incorrectly classified by the J48 classifier. This analysis shows that, the NB classier is more efficient to predict the class label of 'Rate' of web videos.

E) Class Label Prediction of 'Ratings' Attributes

In the prediction of the class labels of the 'Rate' attribute, 165 metadata instances are correctly classified and 17 instances are incorrectly classified by the both J48 classification model and navie Bayesian (NB) classification model. This statistics shows that, the number of correctly classified and incorrectly classified web video tuples are same for the both considered models. However, the classification efficiency of the J48 classification model is found 88.0%, and efficiency of the navie Bayesian classification (NB) model is found 90.9%. This is because, the FP rate of J48 classification model is high as compared to FP rate of NB classification model. Also, the precision of NB classifier is high as compared to J48 classifier. Hence, the efficiency of NB classification model found with good accuracy. The comparative analysis of classification result is represented in the form of confusion matrix as shown in the Table 8.

From the result of confusion matrix presented in the Table 8, all the web video tuples which belongs to 'Very High' class label are correctly classified by the J48 classification model, where as all the web video tuples which are belongs to 'Very High' are incorrectly classified by NB classification model. With respect to the class label 'Medium', all the tuples are incorrectly classified by the J48 classifier, and few are correctly classified by the NB classifier. This will lead to increase in the FP rate of J48 classifier and decrease in the FP rate of NB classifier. This analysis shows that, the NB classier is more efficient to predict the class label of 'Ratings' of web videos.

F) Class Label Prediction of 'Comments' Attributes

In the prediction of the class labels of the 'Comments' attribute, 164 metadata instances are correctly classified and 18 instances are incorrectly classified by the J48 classification model, whereas, 166 metadata instances are correctly classified and 16 instances are incorrectly classified by the navie Bayesian (NB) classification model. This statistics shows that, the NB classifier predicted class labels more correctly as compared to J48 classification model. Also the classification efficiency of the J48 classification model is 89.3%, and efficiency of the navie Bayesian classification model is 91.3%. The comparative analysis of classification result is represented in the form of confusion matrix as shown in the Table 9. The performance evaluation metrics for the classifiers J48 and NB are found nearly same. However, the precision, recall and F-score of the NB classifier is found with good accuracy as compared to J48 classifier. This analysis shows that, the NB classier is more efficient to predict the class label of number of 'Comments' of web videos.

Table 6. Confusion Matrix for 'View Count' Prediction

Confusion Matrix											
J48 Classification Model ===Confusion Matrix===						NB Classification Model ===Confusion Matrix===					
a	**b**	**c**	**d**		**←Classified as**	**a**	**b**	**c**	**d**		**←Classified as**
58	0	0	0	|	a=Very High	44	13	1	0	|	a= Very High
10	97	0	0	|	b=Low	5	102	0	0	|	b=Low
15	0	0	0	|	c=High	6	6	3	0	|	c=High
2	0	0	0	|	d=Medium	1	0	1	0	|	d=Medium

Table 7. Confusion Matrix for 'Rate' prediction

Confusion Matrix											
J48 Classification Model ===Confusion Matrix===						NB Classification Model ===Confusion Matrix===					
a	**b**	**c**	**d**		**←Classified as**	**a**	**b**	**c**	**d**		**←Classified as**
56	0	0	9	|	a=Very High	56	2	0	7	|	a= Very High
11	0	0	1	|	b=Medium	8	3	0	1	|	b=Medium
8	0	0	0	|	c=High	8	0	0	0	|	c=High
0	0	0	97	|	d=Low	0	0	0	97	|	d=Low

Table 8. Confusion Matrix for 'Ratings' Prediction

Confusion Matrix											
J48 Classification Model ===Confusion Matrix===						NB Classification Model ===Confusion Matrix===					
a	**b**	**c**	**d**		**←Classified as**	**a**	**b**	**c**	**d**		**←Classified as**
1	0	0	0	|	a=Very High	0	1	0	0	|	a= Very High
0	11	4	0	|	b=High	0	12	2	1	|	b=High
0	3	153	0	|	c=Low	0	3	147	6	|	c=Low
0	1	9	0	|	d=Medium	0	1	3	6	|	d=Medium

Table 9. Confusion Matrix for 'Comments' Prediction

Confusion Matrix											
J48 Classification Model ===Confusion Matrix===						NB Classification Model ===Confusion Matrix===					
a	**b**	**c**	**d**		**←Classified as**	**a**	**b**	**c**	**d**		**←Classified as**
1	0	0	0	|	a=Very High	0	1	0	0	|	a= Very High
0	11	1	3	|	b=High	0	12	2	1	|	b=High
0	3	6	7	|	c=Low	0	3	11	2	|	c=Low
0	1	3	146	|	d=Medium	0	1	6	143	|	d=Medium

G) Classification Efficiency Comparison of J48 and NB Classification Models

In this section, the classification result comparative analysis of efficiency of J48 and NB classification models are represented for the attributes 'Length', 'Views', 'Rate', 'Ratings', and 'Comments'. The J48 classification model is found with good accuracy for the prediction of class labels of the attribute 'Views'. For the remaining attributes the NB classification model is found with highest efficiency as compared to J48 classification model. Hence, by considering the overall experimental results, the NB classification model with predictive analysis is found with highest efficiency for the classification of web video metadata attributes/features. The comparative analysis of efficiency of J48 and NB classification model is represented in the Fig. 4.

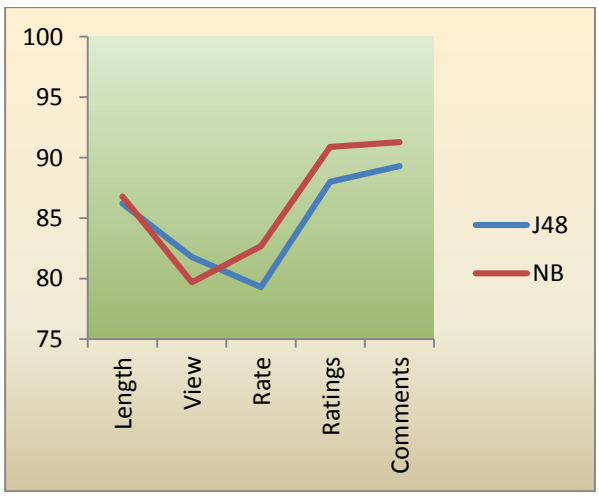


Fig.4. Comparative Analysis of Efficiency of J48 and NB Classification Models

V. Conclusion and Future Work

In this work, we classified web videos based on their metadata attributes/features such as- length, view counts, rate, ratings, and number of comments as a part of knowledge discovery from web videos. The web video metadata are extracted from standard website and stored in a database for classification. The J48 and navie Bayesian (NB) classification algorithms are chosen to classify/predict the class labels of different attributes chosen. The classification results of J48 and NB classification models are compared and found NB classification model with predictive analysis is more efficient for classify web videos using metadata. The future work is to classify the objects present in the different categories of web videos using NB classification model.

References

[1] Amjad Mahmood, Tianrui Li, Yan Yang, Hongjun Wang and Mehtab Afzal, "Semi-supervised evolutionary ensembles for Web video categorization", Elsevier-Knowledge-Based Systems 76 (2015) 53–66.

[2] J.Slimi, A.B.Ammar, and A.M.Alimi, "Video data Visualization System:Semantic Classiffication and Personalization System", International Journal of Computer Graphics and Animation, July 2012.

[3] https://www.youtube.com/yt/press/index.html

[4] C.F-Hsu, James C., and E.Khabiri, "Hierarchical Comment Based Clustering", ACM 978-1-4503-0113-8/11/03, March 2011.

[5] Aggarwal N, Agrawal, S. and Sureka, A., "Mining YouTube Metadata for Detecting privacy Invading Harassment and Misdemeanor Videos", Privacy, Security and Trust (PST), 2014 IEEE Twelfth Annual International Conference on , vol., no., pp.84,93, 23-24 July 2014.

[6] Siddu P. Algur, Prashant Bhat, Suraj Jain, "Metadata Construction Model for Web Videos: A Domain Specific Approach", International Journal of Engineering and Computer Science, December 2014.
[7] Polyxeni Katsiouli, Vassileios Tsetsos and Stathes Hadjiefthymiades, "Semantic Video Classification Based on Subtitles an Domain Terminologies", http://ceur-ws.org/Vol-253/paper05.pdf.
[8] Anil Kale and D.G. Wakde, "An Automated Video Classification and Annotation Using Embedded Audio for Content Based Retrieval", Journal of Industrial and Intelligent Information, December 2013.
[9] Bin Cui Ce Zhang and Gao Cong, "Content Enriched Classifier for Web Video Classification", 2010 ACM 978-1-60558-896-4/10/07.
[10] Zheshen Wang, Ming Zhao, Yang Song, Sanjiv Kumar, and Baoxin Li, "YouTubeCat: Learning to Categorize Wild Web Videos", Google Research.
[11] Chunneng Huang, Tianjun Fu and Hsinchun Chen, "Text-based video content classification for online video-sharing sites", Journal of the American Society for Information Science and Technology, May 2010.
[12] Chirag Shah, Charles File, "Infoextractor – A Tool for Social Media Data Mining", JITP 2011.
[13] Siddu P. Algur, Prashant Bhat, Suraj Jain, "The Role of Metadata in Web Video Mining: Issues and Perspectives", International Journal of Engineering Sciences & Research Technology, February-2015.
[14] Siddu P. Algur, Prashant Bhat, "Metadata Based Classification and Analysis of Large Scale Web Videos", International Journal of Emerging Trends and Technologies in Computer Science, May-June 2015.
[15] Dataset for "Statistics and Social Network of YouTube Videos", http://netsg.cs.sfu.ca/youtubedata/.

Authors' Profiles

Dr. Siddu P. Algur is working as Professor, Dept. of Computer Science, Rani Channamma University, Belagavi, Karnataka, India. He received B.E. degree in Electrical and Electronics from Mysore University, Karnataka, India, in 1986. He received his M.E. degree in from NIT, Allahabad, India, in 1991.He obtained Ph.D. degree from the Department of P.G. Studies and Research in Computer Science at Gulbarga University, Gulbarga.

He worked as Lecturer at KLE Society's College of Engineering and Technology and worked as Assistant Professor in the Department of Computer Science and Engineering at SDM College of Engineering and Technology, Dharwad. He was Professor, Dept. of Information Science and Engineering, BVBCET, Hubli, before holding the present position. He was also Director, School of Mathematics and Computing Sciences, RCU, Belagavi. He was also Director, PG Programmes, RCU, Belagavi. His research interest includes Data Mining, Web Mining, Big Data and Information Retrieval from the web and Knowledge discovery techniques. He published more than 45 research papers in peer reviewed International Journals and chaired the sessions in many International conferences.

Mr. Prashant Bhat is pursuing Ph.D programme in Computer Science at Rani Channamma University Belagavi, Karnataka, India. He received B.Sc and M.Sc (Computer Science) degrees from Karnatak University, Dharwad, Karnataka, India, in 2010 and 2012 respectively. His research interest includes Data Mining, Web Mining, web multimedia mining and Information Retrieval from the web and Knowledge discovery techniques, and published 8 research papers in International Journals. Also he has attended and participated in International and National Conferences and Workshops in his research field.

How to cite this paper: Siddu P. Algur, Prashant Bhat,"Web Video Mining: Metadata Predictive Analysis using Classification Techniques", International Journal of Information Technology and Computer Science(IJITCS), Vol.8, No.2, pp.69-77, 2016. DOI: 10.5815/ijitcs.2016.02.09