

## Enhancing multi-class web video categorization model using machine and deep learning approaches

Wael M. S. Yafooz<sup>1</sup>, Abdullah Alsaeedi<sup>1</sup>, Reyadh Alluhaibi<sup>1</sup>, Abdel-Hamid Mohamed Emara<sup>1,2</sup>

<sup>1</sup>Department of Computer Science, College of Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia

<sup>2</sup>Computers and Systems Engineering Department, Faculty of Engineering, Al-Azhar University, Cairo, Egypt

---

### Article Info

#### Article history:

Received Jul 19, 2021

Revised Jan 19, 2022

Accepted Feb 8, 2022

---

#### Keywords:

Classification

Deep learning

Machine learning

Web video categorization

YouTube

---

### ABSTRACT

With today's digital revolution, many people communicate and collaborate in cyberspace. Users rely on social media platforms, such as Facebook, YouTube and Twitter, all of which exert a considerable impact on human lives. In particular, watching videos has become more preferable than simply browsing the internet because of many reasons. However, difficulties arise when searching for specific videos accurately in the same domains, such as entertainment, politics, education, video and TV shows. This problem can be solved through web video categorization (WVC) approaches that utilize video textual information, visual features, or audio approaches. However, retrieving or obtaining videos with similar content with high accuracy is challenging. Therefore, this paper proposes a novel mode for enhancing WVC that is based on user comments and weighted features from video descriptions. Specifically, this model uses supervised learning, along with machine learning classifiers (MLCs) and deep learning (DL) models. Two experiments are conducted on the proposed balanced dataset on the basis of the two proposed algorithms based on multi-classes, namely, education, politics, health and sports. The model achieves high accuracy rates of 97% and 99% by using MLCs and DL models that are based on artificial neural network (ANN) and long short-term memory (LSTM), respectively.

*This is an open access article under the [CC BY-SA](#) license.*



---

### Corresponding Author:

Wael M. S. Yafooz

Department of Computer Science College of Computer Science and Engineering, Taibah University

Madinah, Saudi Arabia

Email: waelmohammed@hotmail.com

---

## 1. INTRODUCTION

The convenient accessibility and speed of the internet has made it a staple tool for many people. The most noticeable and rapidly growing spheres in the context of videos are Daily motion and YouTube. YouTube is known as the largest repository of videos and is widely used for video sharing by billions of users [1]–[4]. However, given the massive number of videos on the web, users face difficulties in accurately retrieving and obtaining the videos they need [5], [6]. The best method to examine, extract and classify web videos on the basis of content similarity is web video categorization (WVC) [7]–[10]. As the number of videos on the web has increased exponentially, the traditional way of manually processing video categorization has become time consuming and thus requires much effort [11]. Along with software applications for categorization purposes, human intervention is sometimes necessary for refining categorization. Therefore, extensive effort is spent on areas of WVC with automatic concepts that can help improve video retrieval accuracy retrieve videos with high content similarity. The similarity of user queries is also used to increase user satisfaction with viewing relevant and required videos.

Existing research has focused on WVC that uses classification [7], [12]–[17] or clustering techniques [18]–[21] and surveys [9], [10], [14], [22], [23]. Categorizing web videos is generally based on visual, audio or textual information. In visual categorization, the main focus is to extract video frames whilst dealing with them as images. The features extracted, such as faces, objects, colors and shapes, are used to compare and classify processes. Audio-based features are extracted from videos, those features are the signals from sounds such as music, loudness and pitch that represent the values used in the classification process. For example, the sound of music is different from the sound of speech. Moreover, a male voice is different from that of a female. Perceptual features, such as music, and violent words differ from each other. Finally, in textual information, authors use the textual information of video titles or video descriptions or their metadata. The combination of visual-based and audio-based categorization can result in satisfactory improvement. However, WVC is a massive challenge in computer vision and machine learning [20], [21].

People share their thoughts, ideas, beliefs, daily activities, experiences, entertainment, feelings and academic knowledge in the form of comments [24], [25]. Users comment on and like and dislike videos to express their ideologies. These comments are considered unstructured data that can be relevant or irrelevant for video content [26]. Such relevant data can be useful for further processing in WVC, particularly in platforms such as YouTube. Therefore, the current work explores the existing methods and techniques for WVC. In addition, this study proposes a novel model called the enhanced multiclass web video categorization model (EMVC). The proposed EMVC enhances the way in which WVC is conducted by utilizing and extracting user comments and weighted features from video descriptions using machine and deep learning (DL) approaches as a form of supervised learning. In addition, this work examines the machine learning classifiers (MLCs) and DL models for the proposed algorithms by using the proposed dataset. The dataset was collected from four types of YouTube videos, namely, sports, health, education and politics, as predefined classes. A total of 86 videos with 42,668 user comments and video descriptions were used. The dataset called Arabic multi-classification dataset (AMCD), publicly available in [27]. AMCD was subjected to several steps, including annotation, noise removal, data cleaning and data pre-processing, model building and model evaluation. After the completion of the pre-processing steps, the dataset was reduced to 8,046 user comments and was thus considered balanced. The two distinct experiments were conducted using MLCs and DL models on the basis of two proposed algorithms. These algorithms utilized the textual information extracted from user comments and video descriptions to extract informative features. These are given weights based on term frequency-inverse document frequency (TD-IFD) and the average and maximum weights of term frequency-inverse document frequency (TF-IDF) of user comments to the video description. The model showed good accuracies of 97% and 99% using MLCs and DL models that were based on artificial neural network (ANN) and long short-term memory (LSTM), respectively.

The main contributions of this work: i) it explores the existing techniques for WVC and highlights the importance of using user comments and video metadata to enhance WVC; ii) it proposes the EMVC that is based on video descriptions and user comments to enhance WVC through MLCs and DL models; iii) it introduces a new dataset (AMCD) that is based on the Arabic dialect collected from 86 YouTube videos with 8,046 user comments; iv) it proposes a novel mathematical equation for improving WVC through two scenarios and by using two proposed algorithms that utilizes the average and maximum TF-IDF weights of user comments to the video descriptions; and v) it examines the importance of using user comments and video descriptions in video classification.

The rest of the paper is organized: in section 3 explains proposed methods, model architecture and design. Section 3 presents the proposed mathematical equation while the experiments, results and discussion are presented in section 4. Finally, the conclusion of this paper is described in section 5.

## 2. PROPOSED METHOD

This section demonstrates the methods and system architecture of the propose model for WVC as shown in Figure 1. The system architecture consists of six main interrelated phases namely; data acquisition, pre-processing, term extraction and word representation, term weighting, classification methods, and model evaluation.

### 2.1. Data acquisition phase

The first level in the model is known as the input phase. In this phase, data is collected from YouTube videos to use in the video categorization process. According to the core objective of this research, the enhanced video categorization is based upon four predefined classes; including health, sport, politics, and education. The determining criteria are the video description and video comments required to extract. The Arabic videos and their associated Arabic comments will be used in the experiments on the condition that the video publication date ranged from 2015 till 2020 and obtained more than 2000 comments. Python 3.6 and YouTube are used for data collection. During the extraction process of the video description and user

comments, the required information is extracted into a single file for each video. In addition, some of the attributes in the file are removed. The output of this phase is used as the input in the pre-processing phase.

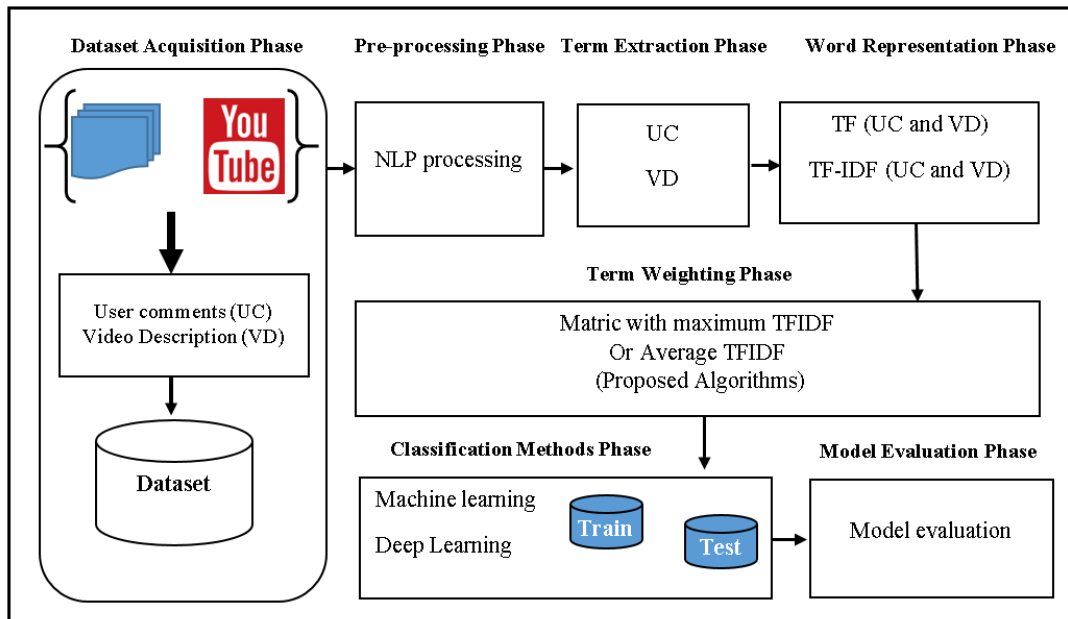


Figure 1. EMVC architecture

## 2.2. Pre-processing phase

There are three steps in the pre-processing phase. They are data cleaning, annotation process, and data pre-processing. In data cleaning, the initial process is to remove the duplicate records and remove any English comments, numbers, or tags. The data in each file that belongs to one video is cleaned. The annotation process is started in the second phase with help of three Native Arabic annotators in computer science. All three annotators are scholars with PhDs. In this process, if two annotators agreed on one classified video that belongs to one of the four classes, the decision is taken that the video belongs to a specific predefined class. Otherwise, the comments are removed if they are not clear or ambiguous. During the annotation process, the class labelling is given for health is “1”, for education “2”, for politics “3” and sport is “4”. After the annotation process, the files are collected in one single file. In the data pre-processing step, Python 3.6 is utilized to perform automatic pre-processing for the dataset. Several steps such as the removal of any HTML tags, numbers, English characters, character extensions, and repeated characters using regular expressions are performed. The porter stemming was used to obtain the root of the words.

## 2.3. Term extraction and word representation phase

In this phase, the terms are extracted from both the video description (VD) and user comments (UC) as described in definition 1 and definition 2. For each comment, a set of extracted words with videos description is called word representation. These comments are transformed into vector representation using “TfidfTransformer” in the “sklearn”. For each video, the set of comments is called vector representation and denoted by VR. The set of vector representations is called a data collection, denoted by DC, containing sets of comments for all videos. For each video, the combination of terms presented in users comments and video description is called word representation (WR).

- Definition 1. Given a video  $v \in V$ , the set of user comment  $UC^v$  and video description  $VD^v$  for  $v$  is defined:

$$\{UC^v + VD^v\}$$

- Definition 2. Given  $n$  videos, the set of word representation (WR) is defined:

$$WR = \{\{UC^1 + VD^1\}, \{UC^2 + VD^2\}, \{UC^3 + VD^3\} \dots \dots \{UC^n + VD^n\}\}$$

- Definition 3. Given n videos, the set of comments is called vector representation (VR) is defined:

$$VR = \{WR^1, WR^2, WR^3, \dots \dots WR^n\}$$

- Definition 4. Given n videos, the set of vector representation is called data collection (DC) is defined:

$$DC = \{VR^1, VR^2, VR^3, \dots \dots VR^n\}$$

#### 2.4. Term weighting phase

In the term weighting phase, the proposed mathematical formula has been employed based on the TF-IDF. The TF-IDF has been extracted from user comments and YouTube metadata, particularly on the video description only, the mathematical formula as shown in (1):

$$W(w, C) = TF(w) C \text{ Log } \frac{N}{CF(T)} \quad (1)$$

where, TF(w) C is denotes number of word (w) in comment (C). CF(T) is denotes number of comments containing word (w). N is denotes is the total number of comments in dataset.

#### 2.5. Classification methods phase

In this phase, two types of classification methods were used are; classical machine learning classifiers (MLC) and deep learning. In classical machine learning classifiers, k-nearest neighbours (KNN), naive Bayes (NB), decision trees (DT), random forest (RF), support vector machine (SVM) and regular regression.

##### 2.5.1. Naive Bayes (NB)

Naive Bayes (NB) classifiers, known as a parametric classifier which is based on some parameters. It is a simple probabilistic classifier based on concepts of the Bayes theorem in statistics. It is used to solve the classification problem with assigned data points to class label with an independence concept. The next mathematical has been used in the proposed model:

$$P\left(\frac{B}{A}\right) = \frac{P\left(\frac{A}{B}\right) * P(B)}{P(A)} \quad (2)$$

where, B is the collection of text in specific class/classes, Let B={Education, Health, Sport and Politics}. A is the word or comments, Let A={User comments and Video Descriptions}. P(A/B) is probability of that word or comment B is belong to class A. P(B/A) is Probability of that the word or comment (A) in the specific class (B).

##### 2.5.2. K-nearest neighbours (KNN)

KNN is non-parametric classification algorithm. It is classifying dataset based on the distance between data points using similarity measure using distance function such as Euclidean distance, Manhattan distance, cosine similarity, chi-square and correlation. KNN classify data points to its close neighbours so the more close distance is assigned to the same category. The k represents to which group the data point is assigned known as nearest neighbours, if the K is odd the voting will be considered and the majority will be considered.

##### 2.5.3. Decision tree (DT)

Decision tree is non-parameter machine learning classifier. It uses a concept of tree structure which consist of root, children/internal nodes and tree leaf nodes. In this way, the dataset into split based on threshold and some conditions from the tree root until reach tree leaves. The tree internal node represents the testing process on the features while tree leaf represent the decisions and class labels as shown in Figure 2. In the decision tree classifier, we need to start with one root of the extracted features in DC in order to do this, we are required to know the highest information gain of the extracted features. This can be calculated using the (3) which is based on calculating the entropy using the (4). Where, Pi is the probability of that features in the data collection (Call) belong to class i.

$$Info(C_{all}) = - \sum_{i=1}^m P_i \log_2(P_i) \quad (3)$$

$$Information\ Gain\ (feature) = (Info(C_{all}) - Info_{feature}(C_{all})) \quad (4)$$

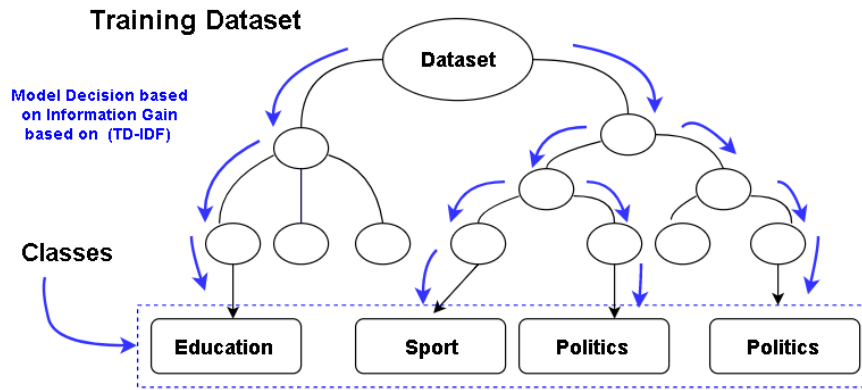


Figure 2. Decision tree (DT)

**2.5.4. Random forest (RF)**

Random forest (RF) is conation several DTs as shown in Figure 3. The dataset is divided randomly into all the DT and also can be duplicate to DT. The final results of the model are based on the majority vote of outcomes of DTs model. In addition, a large number for DT increase the model performance accuracy.

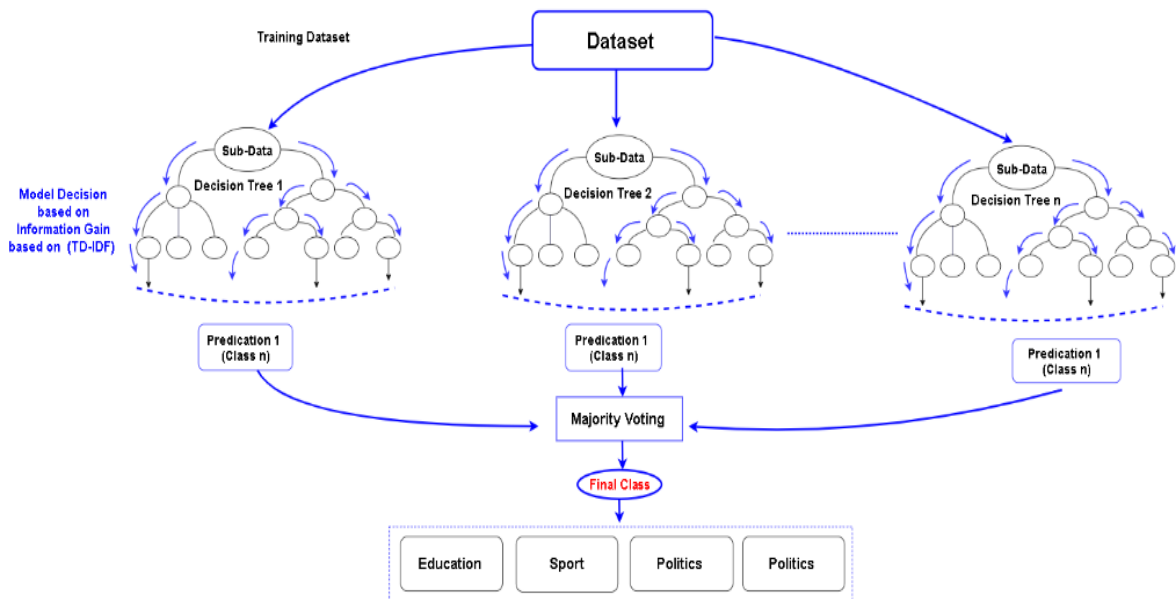


Figure 3. Random forest (RF)

**2.5.5. Support vector machine (SVM)**

SVM is mainly used for binary classification problem. It is divided the data points in the multidimensional space into two classes based on the supports vectors which are closest to the hyperplane. In this case which is multi-classification, SVM breaks down the problem into binary classification problem based on two main approaches are one-to-one or one-to-rest.

**2.5.6. Deep learning model**

In deep learning, this study used an ANN model which known as multilayer perceptron (MLP) [28] in order to classify the proposed data and to examine the model performance. Generally, the deep learning model consists of three layers namely; hidden and output players. In this study, the input layer received it from the maximum or average of the TF-IDF. The output layers consist of four neurons for the four classes (education, health, politics and sport). The ANN as shown in Figure 4.

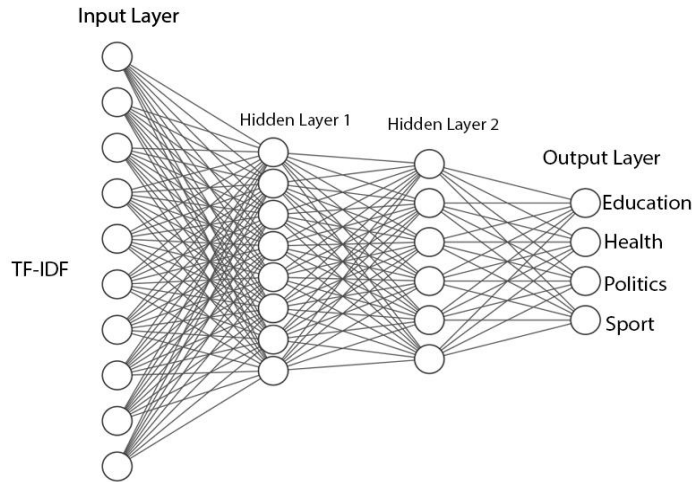


Figure 4. Artificial neural network architecture

**2.6. Model evaluation**

In order to evaluate the model performance, the most popular methods have been used the confusion matrix as shown in Figure 5 and the cross-validation process. The confusion matrix has been used to evaluate the model performance in the accuracy. In the confusion matrix, the recall, precision, F-score and accuracy have been utilized based on next the mathematical formulas (5)-(8):

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

$$F1 - score = \frac{2*(Precision*Recall)}{Precision+Recall} \tag{7}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

where, *TP* is true positive, *TN* is true negative, *FP* is false positive and *FN* is false negative.

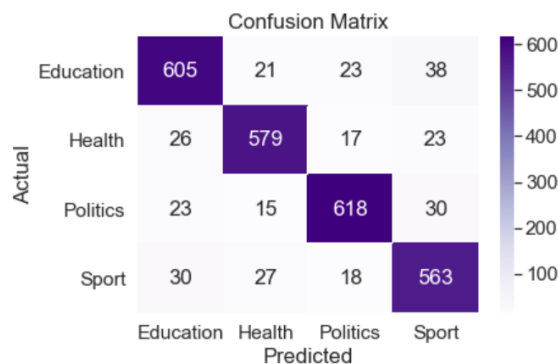


Figure 5. Confusion matrix

In addition, cross-validation had performed on the proposed dataset in order to examine the mode on the introduced dataset. The process was carried into three types 3, 5 and 10 folds, in all the experiments the results show that the difference between the validation and training data is between 1-3% only which indicates no overfitting or underfitting issues.

### 3. MATHEMATICAL FORMULA AND PROPOSED ALGORITHMS

This section explains the mathematical equations and the proposed algorithms used in WVC based on the introduced dataset. The dataset is a collection of unstructured data which consist of video descriptions and user comments. The comments collection denoted by  $C_{all}$  and  $A_{all}$  represents a collection of video descriptions that needs to be classified into one of the four classifications. Both  $C_{all}$  and  $A_{all}$  are used to extract the feature representations. The proposed feature representation can be represented by the following terms.

- Definition 1. There exists user comment (UC) and a collection of user comments ( $C_{all}$ ) where  $C_{all}$  represent all features in C. Therefore,  $C_{all}=UC^1, UC^2, UC^3, \dots, UC^n$  where n is the total, can be defined:

$$\exists UC \in C_{all}$$

- Definition 2. There exists a video description ( $VD_{all}$ ) where  $A_{all}$  consists of one or several video. Therefore,  $VD_{all}=VD^1, VD^2, VD^3, \dots, VD^m$ , where m is the total video, can be defined:

$$\exists VD \in VD_{all}$$

- Definition 3. There also exists a set of extracted features ( $FC_{all}$ ) from (UC) in  $C_{all}$  that can be represented as  $FC_{all}=FC^1, FC^2, FC^3, \dots, FC^r$ , where FC is word/term and r is the number of features, can be defined:

$$FC_{all}=\{\exists FC \in FC_{all} \wedge FC_{all} \in C \wedge C \in C_{all}\}$$

- Definition 4. There also exists a set of extracted features ( $FA_{all}$ ) from (VD) in  $A_{all}$  that can be represented as  $FA_{all}=FA^1, FA^2, FA^3, \dots, FA^s$ , where FA is word/term and s is the number of features, can be defined:

$$FA_{all} = \{\exists FA \in FA_{all} | FA_{all} \in VD \wedge VD \in VD_{all}\}$$

- Definition 5. The maximum value of TF-IDF of extracted features ( $FC_{all\_TF-IDF}$ ) from collection of comments  $C_{all}$ , denoted by MaxFC, as defined in (9).

$$MaxFC = \{\forall FC \in FC_{all} | (FC_{all} (\max(TF-IDF)))\} \quad (9)$$

- Definition 6. The maximum value TF-IDF of extracted features ( $FC_{all}$ ) assigned to  $FA_{all}$ , if it greater than TF-IDF ( $FA_{all}$ ), denoted by  $FA_{all\_TF-IDF}$  as defined in (10).

$$FA_{all\_TFIDF} = \max(TFIDF (FA_{all}) \vee MaxFC) \quad (10)$$

- Definition 7. The average value TF-IDF of extracted features ( $FC_{all}$ ) assigned to  $FA_{all}$ , if it greater than TF-IDF ( $FA_{all}$ ), denoted by  $FA_{all\_TF-IDF}$ , as defined in (11).

$$FA_{all\_TF-IDF} = \text{Average} (TF - IDF (FA_{all}) \vee MaxFC) \quad (11)$$

- Definition 8. Based on obtain the max and assigned to  $FA_{all\_TFIDF}$ , as defined in (12).

$$\text{Matrix Representation} = FC_{all\_TFIDF} \cup FA_{all\_TFIDF} \quad (12)$$

Based on the aforementioned equations, there are two scenarios, the first scenario has applied the algorithm 1 which is considered the average TF-IDF of the user comments to the extracted features (terms) of video description while the second scenario is algorithm 2 which consider the maximum TF-IDF of user comments to extract features (terms) of the video description.

### 4. EXPERIMENTS, RESULTS AND DISCUSSION

This section demonstrates the experiment results of the proposed models and algorithms using Classical MLCs and DL Models. The dataset description is included in the subsequent section.

#### 4.1. Dataset

In this section, the dataset used in the experiments is textual data collected from YouTube videos that contain metadata and user comments. It consists of four classes are health, education, politics and sport, the total number of comments is 8,046 after pre-processing steps. The maximum length of a comment is 1235 in political class while the minimum length is one in all three classes. The detailed description of the dataset is shown in Table 1.

Algorithm 1: Matrix representation for TF-IDF for user comments and average weighted feature of video description

```

0 INPUT:
1 User Comments (UC)
2 Video Description (VD)
3 Call denoted collection of comments
4 Aall denoted collection of video description
5 OUTPUT: Matrix Representation (TF-IDF (Call and Aall))
6 BEGIN
7 INT FCall_TF-IDF, FAall_TF-IDF;
8 CHAR Call, Aall, UC, VD;
9 While true Do
10 Call=UC++;
11 Aall=VD++;
12 End;
13 While true Do
14 FCall_TF-IDF=TF-IDF(Call);
15 FAall_TF-IDF=TF-IDF(Aall);
16 END;
17 While true Do
18 If Average (FCall_TF-IDF)>FAall_TF-IDF Then
19 FAall_TF-IDF=Average (Call_TF-IDF);
20 End IF;
21 End;
22 Matrix_Representation=FCall_TF-IDF+FAall_TF-IDF;
23 End;

```

Algorithm 2: Matrix representation for TF-IDF for user comments and maximum weighted feature of video description

```

0 INPUT:
1 User Comments (UC)
2 Video Description (VD)
3 Call denoted collection of comments
4 Aall denoted collection of video description
5 OUTPUT: Matrix Representation (TF-IDF (Call and Aall))
6 BEGIN
7 INT FCall_TF-IDF, FAall_TF-IDF;
8 CHAR Call, Aall, UC, VD;
9 While true Do
10 Call=UC++;
11 Aall=VD++;
12 End;
13 While true Do
14 FCall_TF-IDF=TF-IDF(Call);
15 FAall_TF-IDF=TF-IDF(Aall);
16 End;
17 While true Do
18 If Max(FCall_TF-IDF)>FAall_TF-IDF Then
19 FAall_TF-IDF=Max(Call_TF-IDF);
20 End IF;
21 End;
22 Matrix_Representation=FCall_TF-IDF+FAall_TF-IDF;
23 END;

```

Table 1. Dataset description

Description/Item	Class	Val.	Comments Number	Percentage	Max. Length	Min. Length
1	Education	1	2001	24.8%	99	1
2	Health	2	2021	25.2%	970	2
3	Politics	3	2017	25.1%	1235	1
4	Sport	4	2007	24.9%	119	1
Total			8046	100%		



#### 4.2. MLCs experiments

In this section, the experiment based on classical machine learning classifiers was carried out using the most common classifiers; KNN, SVM, NB, LR, DT, SGD, and RF. This experiment was performed using Python 3.6 with the aforementioned pre-processing steps. The model performance with/out proposed algorithms were measured using the confusion matrix in precision, recall, f-score, and accuracy.

There were four types of experiments carried out. These are as follows; experiment based on MCLs, experiment MCLs with applied algorithms 1 and 2 with 30 features, experiment MCLs with applied algorithms 1 and 2 with 40 features, and experiment based on MCLs with applied algorithms 1 and 2 with 50 features. In the first experiment, the experiment based on MCLs was performed using N-grams in form of bigrams and trigrams without applied proposed algorithms on the proposed dataset. In addition, the user comments have been utilized only in this experiment. This is aimed to examine the model performance before using the proposed algorithms, the results are as shown in Table 2. Table 2 shows the comparative analysis of the results on model performance based on the first experiment between MCLs. The model accuracy using LR and SGD reached 87% and 88%, respectively with bigram. However, no improvement was recorded using trigram for all MLCs, practically, the experiments were repeated several times. Consequently, all the experiments were carried out only using the bigram.

Table 2. Results of MLC without algorithm 1 and 2

MLCs	Class	Bigram				MLCs	Class	Trigram			
		Precision	Recall	F-score	Accuracy			Precision	Recall	F-score	Accuracy
KNN	1	74%	61%	66%	62%	KNN	1	73%	63%	67%	63 %
	2	55%	87%	67%			2	55%	88%	68%	
	3	47%	71%	56%			3	44%	85%	58%	
	4	73%	49%	58%			4	80%	47%	59%	
SVM	1	85%	88%	86%	87%	SVM	1	85%	87%	86%	87%
	2	90%	93%	91%			2	90%	93%	91%	
	3	87%	90%	89%			3	87%	91%	89%	
	4	86%	78%	82%			4	86%	78%	82%	
NB	1	75%	93%	83%	83%	NB	1	73%	94%	82%	83%
	2	87%	87%	87%			2	87%	87%	87%	
	3	80%	89%	84%			3	80%	89%	84%	
	4	90%	67%	77%			4	90%	67%	77%	
LR	1	87%	88%	87%	88%	LR	1	87%	88%	87%	88%
	2	91%	92%	92%			2	91%	92%	92%	
	3	88%	91%	89%			3	88%	91%	90%	
	4	85%	81%	83%			4	86%	81%	83%	
DT	1	38%	97%	54%	56%	DT	1	38%	97%	54%	56%
	2	54%	96%	70%			2	54%	96%	69%	
	3	36%	94%	52%			3	36%	94%	52%	
	4	99%	36%	53%			4	99%	36%	53%	
SGD	1	87%	87%	87%	87%	SGD	1	88%	86%	87%	88%
	2	91%	93%	92%			2	91%	93%	92%	
	3	88%	89%	89%			3	88%	91%	89%	
	4	86%	82%	84%			4	85%	82%	84%	
RF	1	84%	87%	85%	86%	RF	1	85%	86%	85%	86%
	2	89%	91%	90%			2	87%	91%	89%	
	3	86%	88%	87%			3	86%	88%	87%	
	4	85%	78%	81%			4	85%	79%	82%	

In the second experiment, 30 features were extracted from the video description and applied MCLs with applied algorithm 1 and algorithm 2. The outcome of this experiment has been compared with the results of the first experiment in order to measure the improvement of the model performance using the proposed algorithms. In this experiment, the algorithm 1 was applied that used the average of TF-IDF which outperformed model performance in term of accuracy in the first experiment as shown in Figure 6. Additionally, the results of this experiment show that algorithm 2 had been recorded a significant improvement compared to algorithm 1 shown in Figure 7.

In the third experiment, the extracted features for MCLs with applied algorithms 1 and 2 were increased to 40. Therefore, the model performance recorded the highest accuracy compared to the 30 features shown in Figure 8. The highest accuracy recorded using RF reached 97, whereas using KNN recorded the worst. Thus, the model performance using algorithm 2 outperformed algorithm 1.

The fourth experiment was to examine the performance of the 50 model-based features extracted from the video description or MCLs with applied algorithms (1) and (2). All the MCLs achieved the highest accuracy compared to all aforementioned experiments. The RF, NB, SGD reached 99% in terms of accuracy

as shown in Figure 9. Overall, based on the results of the four experiments using the MCLs with applied the proposed algorithms, the highest accuracy has been attained using TF-IDF with algorithm 2 with 50 features. The experiments were repeated several times with more features however, the accuracy not improved.

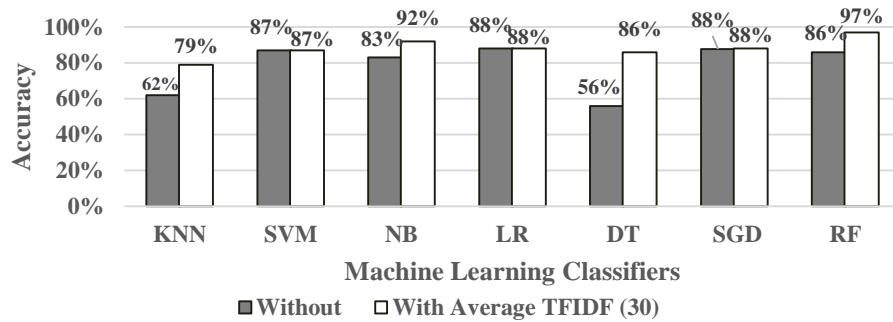


Figure 6. The model performance between normal and algorithm 2

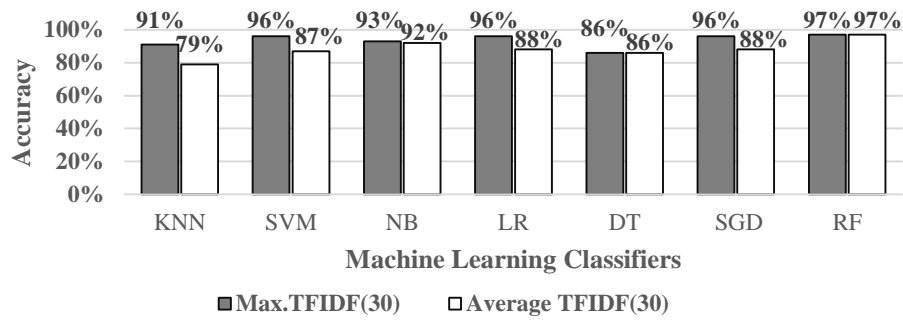


Figure 7. Accuracy of MCLs (max. and average TF-IDF of 30 features)

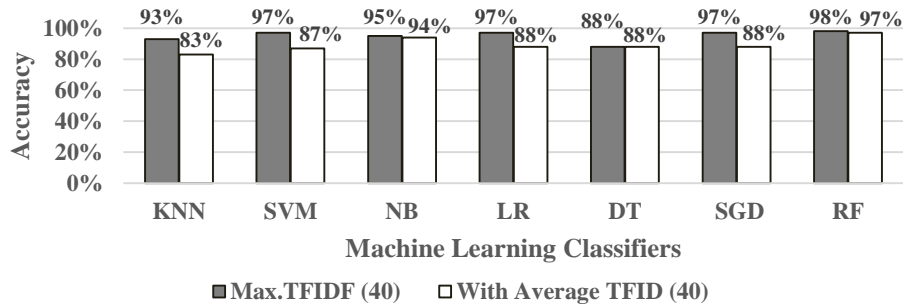


Figure 8. Accuracy of MCLs (max. and average TF-IDF of 40 features)

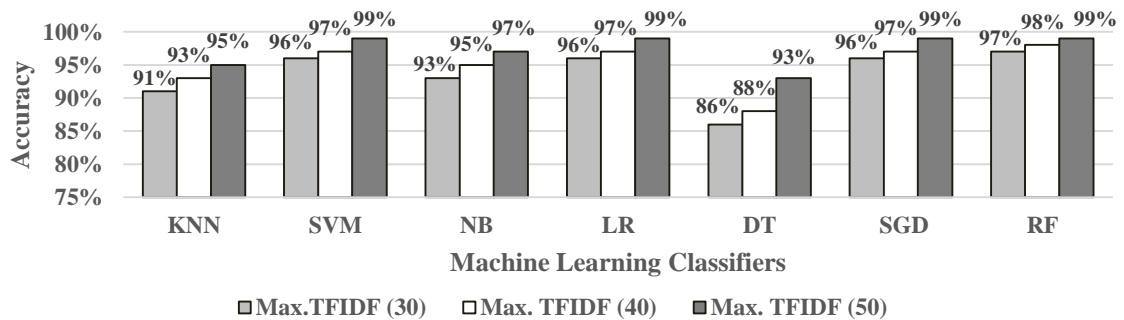


Figure 9. Accuracy of MCLs (Max. TF-IDF of 50 features)

### 4.3. Deep learning models

This section explains the second type of experiment that is conducted using the deep neural network models based MPL on the proposed dataset. This is to measure the proposed model and algorithms using deep learning models and to examine the results and model performance compared with MCLs. Two model architectures have been applied are ANN and LSTM using the two proposed algorithms.

In the ANN experiment, the model builds from 4 layers using Keras. The hyper parameter of the first input layer uses the input dimension of 1000 and an output of 128 neurons the activation function is “relu”. The second and the third layers are hidden. Their output shape consists of 64 neurons and 32 neurons with dropout (0.5), each using “relu” as an activation function. The output layer consists of four neurons using “softmax” as an activation function. The optimizer used is ‘Adam’ with a learning rate of 0.001. The loss function is ‘sparse\_categorical\_crossentropy’ and the accuracy is the training performance. A model training with 70% of the dataset that includes 20 epochs of 64 size batches is used in the training phase. Both proposed algorithms were applied, the validation and training accuracy loss has been decreased as shown in Figures 10 and 11 for algorithm 1 and Figures 12 and 13 for algorithm 2. The model has achieved high performance in the validation and training process in terms of accuracy. In the test phase that uses 30% of the dataset, the model performance has been achieved is approximately 93% and 99% in terms of testing accuracy using algorithm 1 and algorithm 2, respectively. Besides, the experiments with the same configurations were repeated with different learning rates as shown in Table 3.

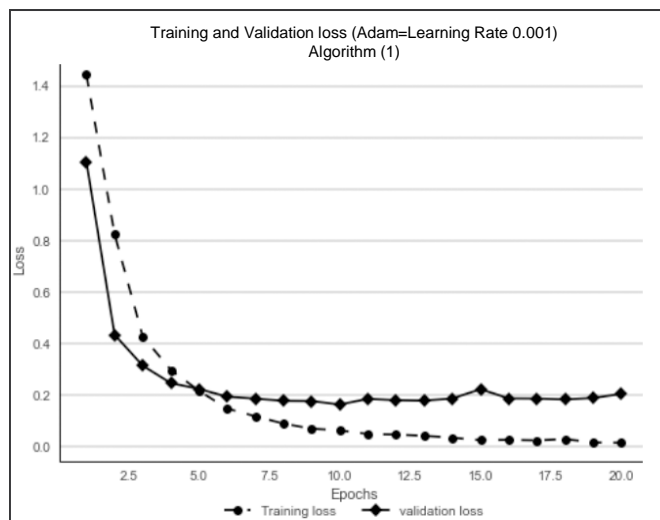


Figure 10. Training and validation loss-algorithm 1

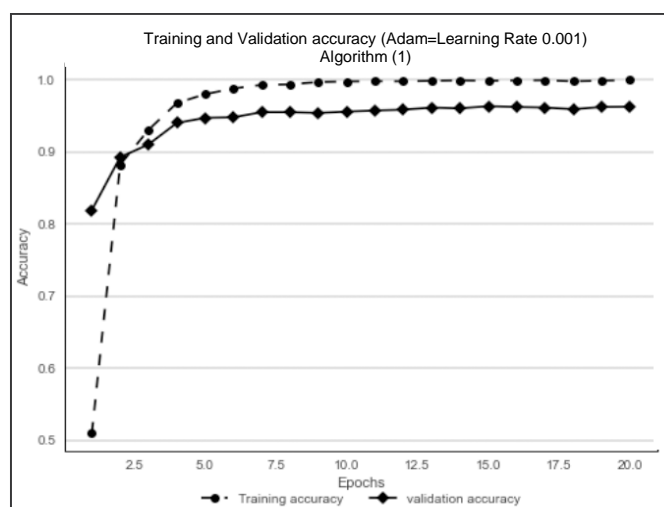


Figure 11. Training and validation accuracy-algorithm 1

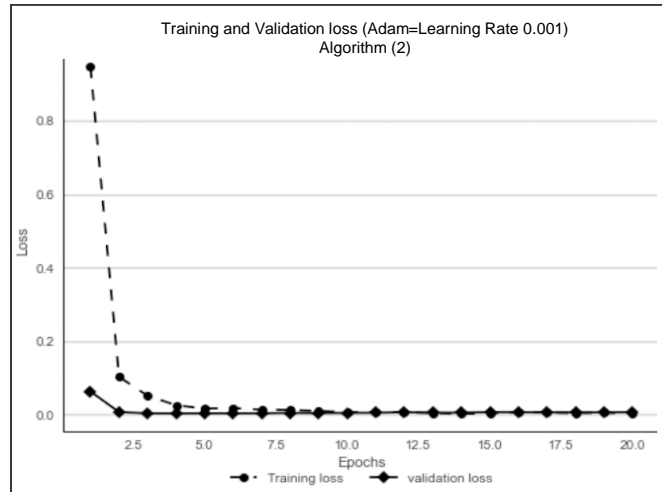


Figure 12. Training and validation loss-algorithm 2

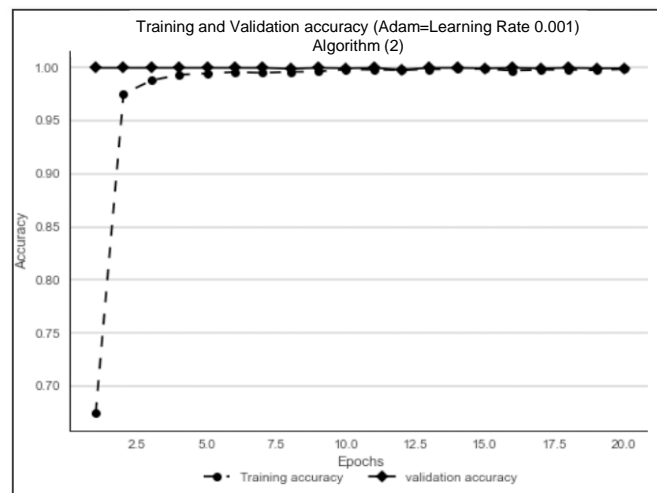


Figure 13. Training and validation accuracy-algorithm 2

Table 3. Experiment results of Adam optimizer with different learning rate

Learning Rate	Algorithm	Loss	Accuracy		
			Training	Validation	Testing
0.01	1	0.0430	0.98	0.9172	0.9171
	2	0.0100	0.9972	0.9962	0.9962
0.001	1	0.0351	0.9878	0.9307	0.9307
	2	0.0065	0.9981	0.9974	0.9973
0.0001	1	0.4638	0.8474	0.8584	0.8584
	2	0.0812	0.9798	0.9962	0.9962

In LSTM experiment, the same dataset is used with a different model architecture. The LSTM model architecture consists of a stack of layers of three LSTM layers. In the first layer, the shape of the output includes 128 LSTM units with an input dimension of 5,392 user comments and 500 features and 50 features of video description. The second layer contains the same hyperparameters and 64 LSTM units while the third 32 LSTM units. The out layer with four units and the activation function is ‘softmax’. The input shape of the batch size is 64 and the number of training iterations is 20 epochs. The optimizer used is ‘Adam’ with a learning rate of 0.001. The loss function is ‘sparse\_categorical\_crossentropy’ and the accuracy as the training performance. The validation and training accuracy loss has been decreased with stability of improvement with 20 epochs as shown in Figures 14 and 15 for algorithm 1 and Figure 16 and 17 for algorithm 2.

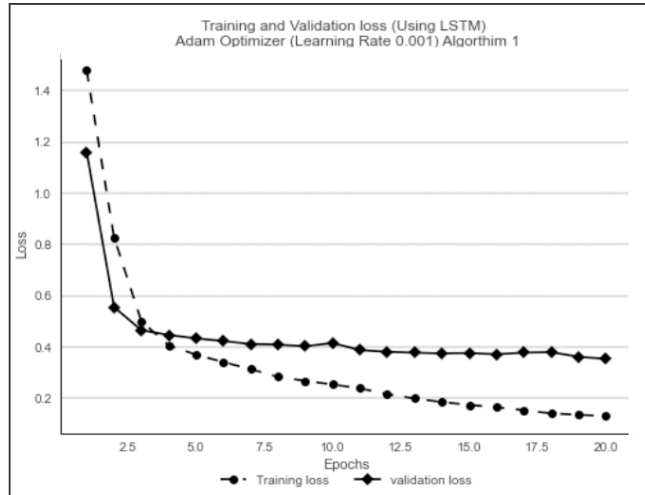


Figure 14. Training and validation loss (algorithm 1)

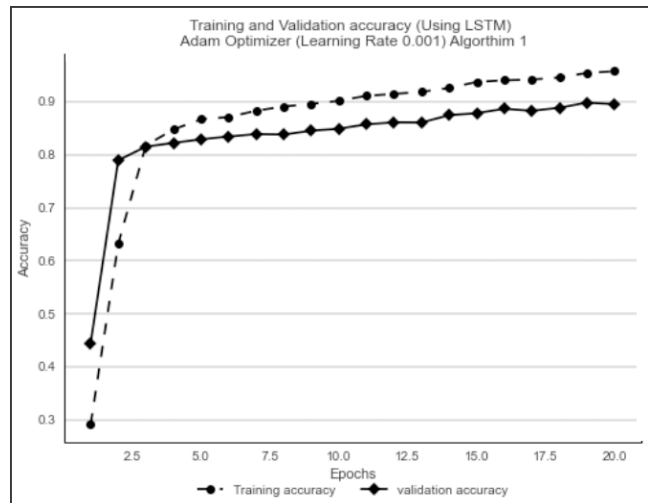


Figure 15. Training and validation accuracy (algorithm 1)

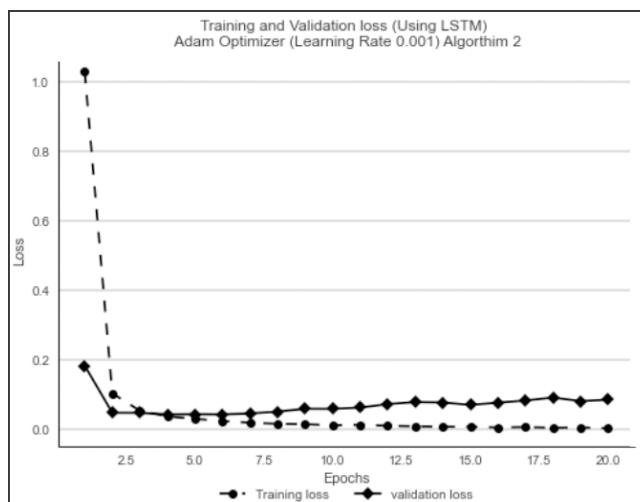


Figure 16. Training and validation loss (algorithm 2)

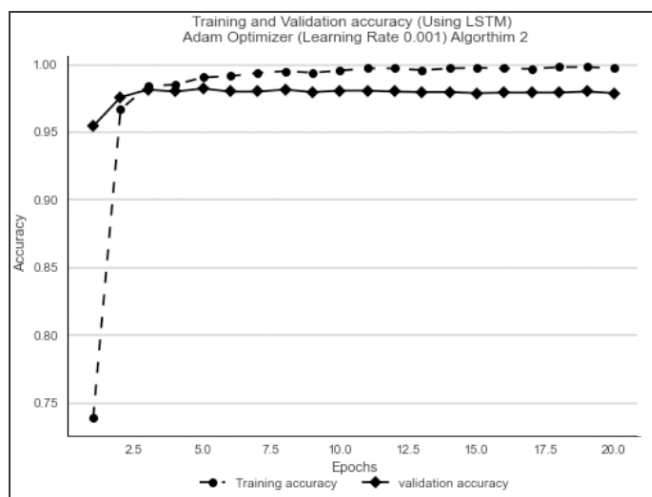


Figure 17. Training and validation accuracy (algorithm 2)

#### 4.4. Results discussion

The majority of the related studies focus on the WVC based on video and audio methodologies. The textual information plays a significant role in achieving a high accuracy rate in enhancing WVC. This is due to many people using social media as a platform to express their opinions by commenting on videos. In addition, the majority of the used dataset as benchmarks such as MCG-WEBV and YouTube-8m is based on the English language that contains an image, video, audio, and meta-data. This study focuses on enhancing the multi-class WVC based on combined textual information that are user comments and video description through proposing two algorithms that utilize the user comments and weight TF-IDF through average or maximum to be assigned to the extracted features from video description. In order to evaluate the model performance, we have conducted experiments based on machine learning and deep learning methods. The experiment results using proposed algorithm 1 and algorithm 2 outperform existing methods, this is found when the experiment results of the model performance in terms of accuracy with more closed classifications approaches are compared. The comparison conducted on the existing approaches are mainly based on textual information. Meanwhile, the results were compared with the approaches with literature. In [29], the accuracy reached 80% through a combination of three approaches that utilize video and audio. In [30], model performance reached 64% for the visual information extracted from frames using a VSM classifier. The work in [31] also highlighted the importance of using sentiment analysis in retrieving data, with the model achieving 75.43% accuracy.

In our experiment, we use a multi-class WVC for four classes ( $M=4$ ), namely, sports, economics, health and education and the introduced dataset based on Arabic language. For the machine learning classifier, the model achieves the highest accuracies of 97%, 93% and 96% when algorithm 2 was applied to 30 features for the RF, NB and stochastic gradient descent (SGD) classifiers, respectively. As for algorithm 1, the accuracies reach 98%, 95% and 97% given 40 features for the RF, NB and SGD classifiers. KNN achieves the worst accuracies of 79% and 83% when algorithm 1 was applied to 30 and 40 features. In the machine learning experiments, the RF, NB and SGD classifiers always outperform the other classifiers. On the other hand, in the deep learning experiments, ANN and LSTM are applied. Both models produce the highest accuracy of up to 99% given a few numbers of layers, neurons and iterations (epochs). The loss function also decreases with a few iterations, with the learning rate being 0.001, which is better than 0.01 and 0.0001. Generally, this result reflects the importance of utilizing user comments and video descriptions as informative features for enhancing WVC. Specifically, the average or maximum TF-IDF weights of user comments to be assigned to video descriptions' extracted features are calculated using the proposed algorithms. In addition, 30, 40 and 50 extracted features of video descriptions are used in this study, with the category comprising 50 features achieving the best performance.

#### 5. CONCLUSION

This paper focuses on enhanced video categorization based on user comments and video descriptions. There are various methods of WVC. They are visual-based, Audio-based, and textual-based. Hybrid methods such as video-based and Audio-based are given more attention in scholarly articles. This

proposed model utilizes the user comments and YouTube video metadata, specifically on video description in enhancing the WVC. Four experiments are carried out using MLCs and DL models with the proposed datasets, and two algorithms. TF-IDF extracted from the video description are used in three categories 30, 40, and 50. The results of these experiments emphasize the usage of the hyper user comments and video descriptions outperform the Standard methods that focus purely on comments. The usage of the third category with the 50 extracted features recorded the highest model performance in terms of accuracy.




## REFERENCES

- [1] W. M. S. Yafooz and A. Alsaeedi, "Sentimental analysis on health-related information with improving model performance using machine learning," *Journal of Computer Science*, vol. 17, no. 2, pp. 112–122, Feb. 2021, doi: 10.3844/jcssp.2021.112.122.
- [2] R. F. Alhujaili and W. M. S. Yafooz, "Sentiment analysis for YouTube videos with user comments: review," in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Mar. 2021, pp. 814–820, doi: 10.1109/ICAIS50930.2021.9396049.
- [3] H. Eksi Ozsoy, "Evaluation of YouTube videos about smile design using the DISCERN tool and Journal of the American Medical Association benchmarks," *The Journal of Prosthetic Dentistry*, vol. 125, no. 1, pp. 151–154, Jan. 2021, doi: 10.1016/j.prosdent.2019.12.016.
- [4] N. N. Moon *et al.*, "Natural language processing based advanced method of unnecessary video detection," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 6, pp. 5411–5419, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5411-5419.
- [5] Z. A. A. Ibrahim, S. Haidar, and I. Sbeity, "Large-scale text-based video classification using contextual features," *European Journal of Electrical Engineering and Computer Science*, vol. 3, no. 2, Apr. 2019, doi: 10.24018/ejece.2019.3.2.68.
- [6] J. Jacob, M. S. Elayidom, and V. P. Devassia, "Video content analysis and retrieval system using video storytelling and indexing techniques," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 6, pp. 6019–6025, Dec. 2020, doi: 10.11591/ijece.v10i6.pp6019-6025.
- [7] Y.-L. Chen, C.-L. Chang, and C.-S. Yeh, "Emotion classification of YouTube videos," *Decision Support Systems*, vol. 101, pp. 40–50, Sep. 2017, doi: 10.1016/j.dss.2017.05.014.
- [8] P. D. P. Woogue, G. A. A. Pineda, and C. V. Maderazo, "Automatic web page categorization using machine learning and educational-based corpus," *International Journal of Computer Theory and Engineering*, vol. 9, no. 6, pp. 427–432, 2017, doi: 10.7763/IJCTE.2017.V9.I180.
- [9] M. Tavakolian and A. Hadid, "Deep discriminative model for video classification," in *Computer Vision textendash ECCV 2018*, Springer International Publishing, 2018, pp. 401–418.
- [10] P. Rani, J. Kaur, and S. Kaswan, "Automatic video classification: a review," *EAI Endorsed Transactions on Creative Technologies*, vol. 7, no. 24, Art. no. 163996, Jun. 2020, doi: 10.4108/eai.13-7-2018.163996.
- [11] M. Afzal, N. Shah, and T. Muhammad, "Web video classification with visual and contextual semantics," *International Journal of Communication Systems*, vol. 32, no. 13, Art. no. e3994, Sep. 2019, doi: 10.1002/dac.3994.
- [12] C. Huang, T. Fu, and H. Chen, "Text-based video content classification for online video-sharing sites," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 5, pp. 891–906, May 2010, doi: 10.1002/asi.21291.
- [13] L. Bahatti, O. Bouattane, M. Elhoussine, and M. Hicham, "An efficient audio classification approach based on support vector machines," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 5, 2016, doi: 10.14569/IJACSA.2016.070530.
- [14] Z. Wu, T. Yao, Y. Fu, and Y.-G. Jiang, "Deep learning for video classification and captioning," in *Frontiers of Multimedia Research*, ACM, 2017, pp. 3–29.
- [15] Y. Yang, D. Krompass, and V. Tresp, "Tensor-train recurrent neural networks for video classification," *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, PMLR 70, 2017.
- [16] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 131–135, doi: 10.1109/ICASSP.2017.7952132.
- [17] G. S. Kalra, R. S. Kathuria, and A. Kumar, "YouTube video classification based on title and description text," in *2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, Oct. 2019, pp. 74–79, doi: 10.1109/ICCCIS48478.2019.8974514.
- [18] P. Q. Nguyen, A.-T. Nguyen-Thi, T. D. Ngo, and T.-A. H. Nguyen, "Using textual semantic similarity to improve clustering quality of web video search results," in *2015 Seventh International Conference on Knowledge and Systems Engineering (KSE)*, Oct. 2015, pp. 156–161, doi: 10.1109/KSE.2015.47.
- [19] D. Saravanan, "Efficient video indexing and retrieval using hierarchical clustering technique," in *Proceedings of the Second International Conference on Computational Intelligence and Informatics*, Springer Singapore, 2018, pp. 1–8.
- [20] X. Long, C. Gan, G. de Melo, J. Wu, X. Liu, and S. Wen, "Attention clusters: purely attention based local feature integration for video classification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7834–7843, doi: 10.1109/CVPR.2018.00817.
- [21] V. Mekthanavanh, T. Li, J. Hu, and Y. Yang, "Social web video clustering based on multi-modal and clustering ensemble," *Neurocomputing*, vol. 366, pp. 234–247, Nov. 2019, doi: 10.1016/j.neucom.2019.07.097.
- [22] J. Wu, S. Zhong, J. Jiang, and Y. Yang, "A novel clustering method for static video summarization," *Multimedia Tools and Applications*, vol. 76, no. 7, pp. 9625–9641, Apr. 2017, doi: 10.1007/s11042-016-3569-x.
- [23] D. Brezeale and D. J. Cook, "Automatic video classification: a survey of the literature," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 3, pp. 416–430, May 2008, doi: 10.1109/TSMCC.2008.919173.
- [24] O. J. Ying, M. M. A. Zabidi, N. Ramli, and U. U. Sheikh, "Sentiment analysis of informal Malay tweets with deep learning," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 9, no. 2, pp. 212–220, Jun. 2020, doi: 10.11591/ijai.v9.i2.pp212-220.
- [25] N. S. A. Abu Bakar, R. Aziehan Rahmat, and U. Faruq Othman, "Polarity classification tool for sentiment analysis in Malay language," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 8, no. 3, pp. 259–263, Dec. 2019, doi: 10.11591/ijai.v8.i3.pp259-263.
- [26] S. Choi and A. Segev, "Finding informative comments for video viewing," in *2016 IEEE International Conference on Big Data (Big Data)*, Dec. 2016, pp. 2457–2465, doi: 10.1109/BigData.2016.7840882.




- [27] "Arabic multi classification dataset-AMCD." Github. <https://github.com/waelyafooz/Arabic-Multi-Classification-Dataset-AMCD> (accessed Aug. 10, 2021).
- [28] F. Albu, A. Mateescu, and N. Dumitriu, "Architecture selection for a multilayer feedforward network," *International Conference on Microelectronics and Computer Science*, pp. 131–134, 1997.
- [29] H. M. Al Amin, M. S. Arefin, and P. K. Dhar, "A method for video categorization by analyzing text, audio, and frames," *International Journal of Information Technology*, vol. 12, no. 3, pp. 889–898, Sep. 2020, doi: 10.1007/s41870-019-00338-2.
- [30] C. Ortega-León, P. A. Marín-Reyes, J. Lorenzo-Navarro, M. Castrillón-Santana, and E. Sánchez-Nielsen, "Video categorisation mimicking text mining," in *Advances in Computational Intelligence*, Springer International Publishing, 2019, pp. 292–301.
- [31] H. Bhuiyan, J. Ara, R. Bardhan, and M. R. Islam, "Retrieving YouTube video by sentiment analysis on user comment," in *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, Sep. 2017, pp. 474–478, doi: 10.1109/ICSIPA.2017.8120658.

## BIOGRAPHIES OF AUTHORS






**Wael M. S. Yafooz**    is an associate professor in the computer Science Department, Taibah University, Saudi Arabia. He received his bachelor degree in the area of computer science from Egypt in 2002 while a master of science in computer Science from the University of MARA Technology (UiTM)- Malaysia 2010 as well as a PhD in Computer Science in 2014 from UiTM. He was awarded many Gold and Silver Medals for his contribution to a local and international expo of innovation and invention in the area of computer science. Besides, he was awarded the Excellent Research Award from UiTM. He served as a member of various committees in many international conferences. Additionally, he chaired IEEE international conferences in Malaysia and China. Besides, He is a volunteer reviewer with different peer-review journals. Moreover, he supervised number of students at the master and PhD levels. Furthermore, He delivered and conducted many workshops in the research area and practical courses in data management, visualization and curriculum design in area of computer science. He was invited as a speaker in many international conferences held in Bangladesh, Thailand, India, China and Russia. His research interest includes, Data Mining, Machine Learning, Deep Learning, Natural Language Processing, Social Network Analytics and Data Management. He can be contacted at email: [waelmohammed@hotmail.com](mailto:waelmohammed@hotmail.com)/[wyafooz@taibahu.edu.sa](mailto:wyafooz@taibahu.edu.sa).






**Abdullah Alsaeedi**    received the B.Sc. degree in computer science from the College of computer science and engineering, Taibah University, Madinah, Saudi Arabia, in 2008, M.Sc. degree in Advanced software engineering, The University of Sheffield, department of computer science, Sheffield, UK, in 2011, and the Ph.D. degree in computer science from the University of Sheffield, UK, in 2016. He is currently an Assistant Professor at the Computer Science Department, Taibah University, Madinah, Saudi Arabia. His research interests include software engineering, software model inference, grammar inference, machine learning. He can be contacted at email: [aasaeedi@taibahu.edu.sa](mailto:aasaeedi@taibahu.edu.sa).



**Reyadh Alluhaibi**    received the B.E degree from Taibah University in 2005. He received M.Sc. degree from Tulsa University in 2009. After working as a lecturer (from 2009 to 2012) in the Dept. of Computer Science, Taibah University. He received the PhD degree from Manchester University in 2017. After working as an assistant professor (from 2017) in the Dept. of Computer Science, Taibah University. His research interest includes Machine Learning, Natural Language Processing, Computational Linguistics, Computational Semantics, Knowledge Representation, and Temporal Logics. He can be contacted at email: [rluhaibi@taibahu.edu.sa](mailto:rluhaibi@taibahu.edu.sa).



**Abdel-Hamid Emara**    is an associate professor in the Department of Computers and Systems Engineering, Faculty of Engineering, Al-Azhar University, Cairo, Egypt He received his BSc, MSc, and PhD in computers and Systems engineering from Al-Azhar university in 1992, 2000, 2006, respectively. He supervised number of students at the master and PhD levels. Furthermore, He delivered and conducted many workshops in the research area He served as a member of various committees in many international conferences. He is a volunteer reviewer with different peer-review journals. He is currently an Assistant Professor in Computer Science Department at University of Taibah at Al-Madinah Al Monawarah, KSA. His research interest includes, Data Mining, Machine Learning, Deep Learning, Natural Language Processing, Social Network Analytics and Data Management. He published several research papers and participated in several in International journal and local/international conferences. He can be contacted at email [aemara@taibahu.edu.sa](mailto:aemara@taibahu.edu.sa).