On Demand Video Retrieval Based on Arabic TEXT

[1]Reem A.K. Aljorani; [2]Boshra F. Zopon Al_Bayaty
[1,2]Mustansiriyah University, Department of Computer Science, Iraq
reem.kareem91@gmail.com; bushraalbayaty123@gmail.com

Article History: Received: 10 November 2020; Revised 12 January 2021 Accepted: 27 January 2021; Published online: 5 April 2021

Abstract: Duo to fast digital technology development in recent years and the evolution of digital cameras and social media, which depends mainly on videos, the interest towards video mining and manipulation field, was increasing because of the rich content of the video. Although Arabic videos are not so common to be studied because of the Arabic language complexity and since education in the meanwhile is facing difficulties duo to the medical situation effecting the whole world. A system was designed and implemented to make the electronic learning process easier for students that study in Arabic language. A model that serves the Iraqi students and the ministry of education was pro-posed. The system provides a search engine for students to reach out a specific topic they need to understand and provide a convenient environment to search for their favorite subject. A dataset of educational videos uploaded by the Iraqi educational television on their channel on YouTube platform. Audio feature was extracted from the videos, and then transcripts were generated by converting Arabic audio into text documents. A Stochastic Gradient Decent technique, which is a machine learning technique, was used to classify each of the videos and figure out to which category or subject the video belongs to. A search technique was applied to enable the student search through different categories of Arabic subjects. The results showed high classification accuracy for SGD in compare to other models.

Key words: Video Mining, Text Classification, Machine learning, Agile Model, Software engineering

1. Introduction

As computer technologies evolves over the years, digital information spread out enormously. In particular, digital videos increased as people prefer to share videos in social media and other web platforms duo to its rich content. Therefore, the process of video retrieval has become more difficult and the development of retrieval systems has become required [1].Video digital contents consist of three main data formats combined to form the video. These contents are (Images in the form of frames, Audio sounds, and Text) which characterize the features of the video. Video features must be extracted in order to retrieve the videos. A video retrieval process is the process of providing the user with a specific video from the database, which is related to a query given as the user needs [2]. One of the applications for video retrieval process is on demand video search engines. Searching for the requested video is not an easy task, which takes a lot of time and resources. In addition, most of the searching techniques on the web depends on detecting the hashtags, the keywords in the title of the video or any other information provided manually by the uploader of the video. In order to retrieve and obtain the requested videos, an automatic and accurate video classification process must be performed first, which is the process of attaching the right labels automatically to a video to facilitate the retrieval process also enhance which videos are stored depending on the digital content of the video [3]. In recent years, many researches in text-based video classification and retrieval were available which deals with different aspects such as movies and music classification, poetry classification, article classification and educational MOOC classification and retrieval. These aspects were mainly videos spoken in English language. The aim towards Arabic text classification started recently. Because of the Arabic grammar complexity, it was not easy to deal with Arabic text [4].

Video classification and retrieval is a broad research field that many researchers dive through it. However, there is not many of them that Retrieve videos based on Arabic Audio and text. Kastrati, Z. Imran and et al. [5] proposed a CNN system based on text transcripts that were extracted from the

MOOC Coursera educational platform and classify the transcripts using embedding's learned from the MOOC dataset and transfer learnings. While L. H. Medida and K. Ramani [6] gathered different e-lecture videos from various internet sources then converted to audio by using (FFmpeg). Audio tracks was translated into text transcripts using the Google Speech Recognition (GSR) library. Then a Support Vector Machine, Naive Bayes and Logistic Regression algorithms was applied. A search technique was performed based on the user query. H. Chatbri [7] suggested an automatic classification for Massive open online database of educational videos based on convolutional neural networks (CNN) by first extracting text transcripts from the video then forming an images with the extracted text. These papers were the most related to our work with an educational dataset of videos spoken in English. In the meanwhile, S. Boukil et al., [8] proposed a novel method for Arabic text classification by using a stemmer algorithm to extract features and choose the most relevant ones to this work then reduce the dimensionality of the feature vectors. Each feature was assign to a weight by using Tf-Idf algorithm. Convolutional Neural Networks algorithm was implemented to classify Arabic text. K. Sundus et al. [9] presented a deep learning model based on feed forwarding for Arabic text classification.

2. The Materials and Methods

A novel work was proposed in this paper to design Arabic educational platform that serves the Iraqi students and the ministry of education by creating a new training dataset based on the subjects studied by secondary sixth grade students, which is the most important stage in the educational journey in Iraq. The new training dataset was used in the system to train and classify six categories of Arabic subjects in the form of educational videos operative in Arabic language then allow the students to navigate through these subjects and search for a specific lessons. The main purpose of this work was to create an educational system that provide rich materials for teaching which enables the students to improve their abilities and help themselves. Figure 1 illustrate the educational system architecture.

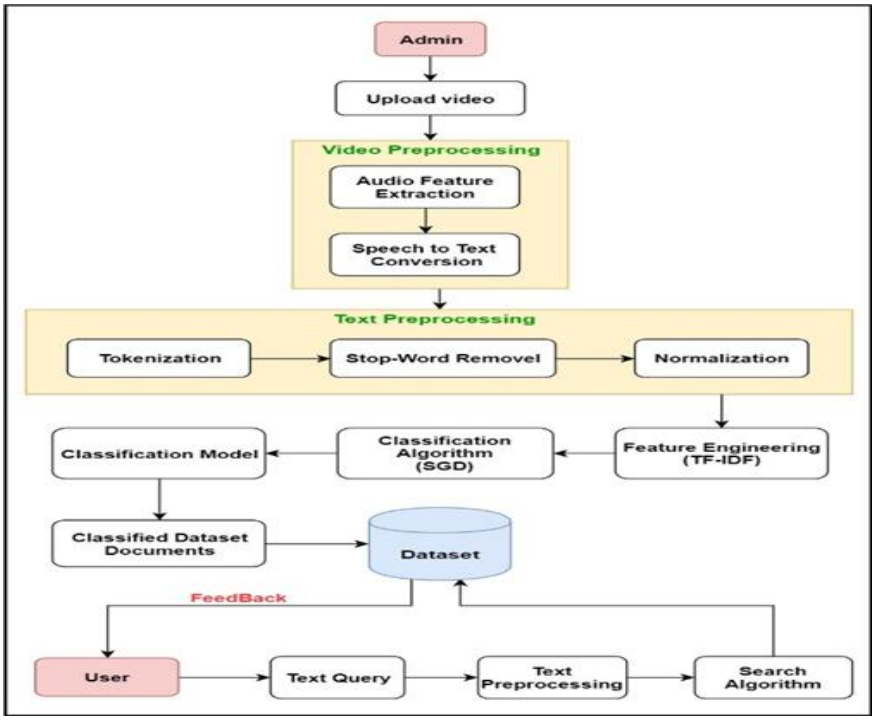

Figure 1. The Educational System Architecture

The proposed system was built by studying the functional and non-functional requirements first. These requirements gave a general overview of what the user and the administrator functions are. Also gave a better insight of how the system works. Figure 2 shows the use case diagram in which the relationship between the admin and the user was illustrated.

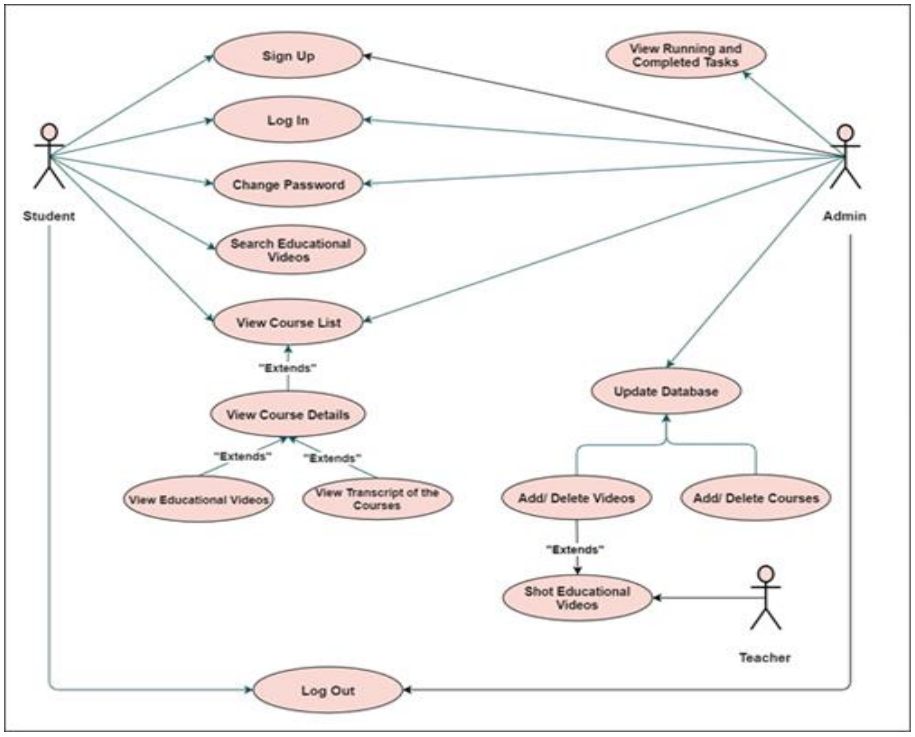

Figure 2. Use Case Diagram of the Educational System Tasks

1.1. Functional Requirements

(1) The Admin should have a log of all users registered in the platform.
(2) The system admin should be able to create user groups to share content, participate and communicate among these groups.
(3) When a task is running in the system to add a new course or video, a background task manager is enabled to ensure that the process is completed.
(4) Background task manager should be refreshed each 10 minutes to check for new tasks such as (add/ Delete / Edit) a video.
(5) When an error is occurred while running a task. The background task manager occurrence must show an error and starts a new process to complete the task.
(6) The system should only allow system admin to make changes to the database.

1.2. Non-Functional Requirements

(1) Accuracy: The system provides an accurate informative teaching material to the students.
(2) Availability: The user can access all the courses and videos available on the educational platform therefore; the system should be available to students all the time.
(3) Usability: The system must be easy to access and easy to deal with.
(4) Security: only secondary sixth grade students who have a user name and password registered in the platform can access the system information.
(5) Reliability: the videos uploaded in the system must contains a reliable information collected from the official sources.
(6) Performance Efficiency: The system must have reasonable speed to be used and accessed by many of users at the same time. In addition, system database must be updated immediately while uploading a new video.
(7) Simplicity: The system design must be simple to use by the students.

1.3. Software Requirements

Python programming language (JetBrains PyCharm Community Edition 2019.2.3 x64), is used to implement the educational system, Django web framework to design the interface as well as Azure Cognitive Services to convert Arabic Audio to a text format.

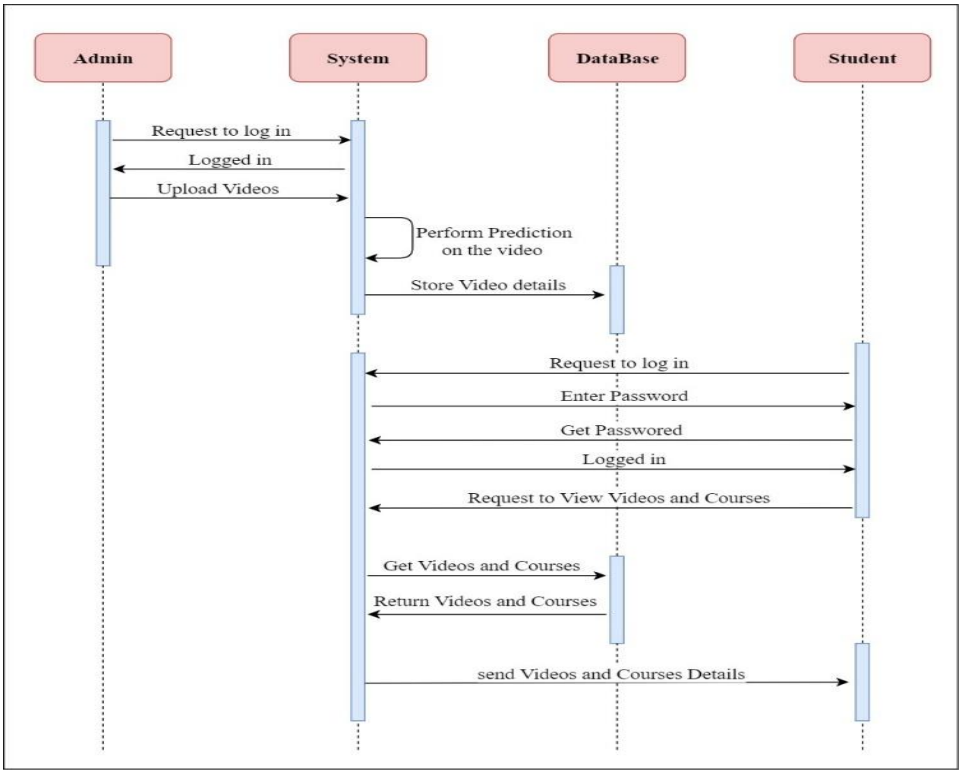

Figure 3. The Sequence Diagram of the Proposed Educational System

Figure 3 gives the sequence of each interaction between the Participants of the system represented by the admin and the student with the proposed educational system.

To check running and completed tasks of the system a background tasks bench was used for the admin to observe whether a video was classified properly or an error occurred during the task of classification. Also the time in which a video was uploaded and a new task was created. Background tasks work simultaneously with the system and check for new video uploads every 10 minutes. Once a video is uploaded by the admin, background task manager gives a notice to the system to start the classification stage.

3. Dataset Gathering and Preparation

In this work, a training dataset was created manually by collecting educational books for secondary sixth grade along with other related books and explanatory documents. The entire set of documents were in Arabic language. The collected documents had been divided into six categories: Math, Arabic Grammar & Literature, Chemistry, Physics, Biology, and Religion. The new training dataset was used to train and prepare the model to predict Educational videos operative in Arabic.

In Addition, a set of 70 videos were gathered from the Iraqi educational TV channel (10 videos for each category) which were used for prediction. A 70% of the dataset have been processed and used for training the system while a 30% of the dataset have been used to evaluate the system. Note that we did not include English language in our classification categories because our system is dealing with subjects written and spoken in Arabic language only. Table 1. Shows the size of the training dataset.

Table 1. The Size of the Dataset

No.	Categories	No. of Documents	No. of Words
1	Biology	218	483,841
2	Chemistry	252	538,837
3	Physics	213	469,105
4	Arabic Grammar & Literature	40	203,093
5	Religion	133	390,942
6	Math	146	390,904

4. Experimental Setup of the Arabic Educational System

The proposed educational system consist of five main stages that make the system integrated. These stages show the flow of data in the system to be in a form that is readable and easy to access by the students.

3.1. Pre-Processing Stage

One of the most important steps involved in building the educational system is the pre-processing stage. Documents and videos must be pre-processed first to be converted to a readable form to be prepared for the next stage.

4.1.1. Video pre-processing stage

The first step in our system building is to manipulate videos in a way to be readable by the system. This step demands several processes to be implemented on the videos.

(1) Audio feature extraction: Since our dataset is educational, which deals with video lectures and the visual content of such videos cannot express the knowledge discussed in the video clearly. Therefore, a distinct special feature of the video was extracted. This feature is the audio feature, that contains all the knowledge as the teacher, explain the rich information in the form of a speech. Audio extraction was implemented for all videos in the dataset using a python module called MoviePy.editor, which provides a feature for converting videos with MP4 format to an audio of WAV format. These WAV audio files generated were stored and organized according to the subject categories.

(2) Speech to text conversion: After converting our dataset of educational videos to audio streaming files with a WAV format, a speech recognition and analysis step was implemented in order to convert Arabic audio files to text to be prepared for the classification and evaluation stage. In the audio to text step, Azure service was used which a service is provided by Microsoft to generate text transcripts. We used Azure Microsoft python library for two reasons:

- Facilitate the converting process from audio files into text files.
- Support several languages including Arabic language, which is the bases of our dataset used in the system.

4.1.2. Text Preprocessing Stages

Over the last few decades, studying text classification problems had been excessive in many applications. In particular, in the field of Natural Language Processing (NLP) and with text mining. Therefore, many researchers head towards developing applications that elevate the effectiveness of text classification methods. The first step in text classification and document categorization is the pre-processing stage which is an important stage in which text was being prepared for the feature engineering stage. In this section, methods were introduced for cleaning text datasets and removing unwanted noise [10]. Three processes were applied to the documents, these are:

(1) Tokenization Process, extract words from the text data files by converting text into tokens. A total of 2,676,722 words was found in the dataset. See Table 1 that shows the generated text files with their corresponding No. of words.
(2) Stop-word removal process was applied on the text files to achieve accuracy, remove punctuation, extraction of redundant words and filter out words with low or no semantic meaning. Stop-words list has been developed for languages like English, Chinese, Arabic, Hindi, etc.
(3) Normalization Process where Arabic words or tokens were converted to its conventional orthodox form by extracting the source abbreviations of a word depending on Arabic dictionary. The normalization process in Arabic language is much more difficult than in English because of the complicated Arabic grammar in compared to English grammar. Two main techniques were used to normalize text document. These are stemmer and Lemmatization techniques. Lemmatization technique was used in this work, which is a NLP process that replace or remove the suffix of a word to get the original basic form of the word. Stanza model was used for this purpose.

3.2. Feature Engineering Stage

In this step, a Term Frequency-Inverse Document Frequency (TF-IDF) is used for feature engineering that assigns a higher weight to words with either high or low frequencies term in the document. The purpose of this method is to convert text documents to a numerical format in order to be used by the classification algorithm. A feature vector is created which contains values calculated by the following equation:

$$W(d,t) = TF(d,t) * \log\left(\frac{N}{df(t)}\right) \quad (1)$$

3.3. Training and Prediction Stage using SGD Algorithm

Gradient Decent GD is an optimization algorithm aims to estimate the best co-efficient of (f) function that minimize the cost by learning a set of classifiers. The cost is evaluated by passing several coefficients to the evaluation function to compute the predicted cost for each sample of the training dataset then compare the predicted cost with the actual value to choose the lowest cost. The algorithm then pass a new different coefficient to the function. The process is completed when the cost approaches zero value [11]. However, the implementation of gradient decent can be exhaustively slow duo to the huge computations by going through all the iterations to compute all the samples in the training dataset. Therefore, a Stochastic Gradient decent SGD algorithm is used to reduce computations and implement large datasets [12]. It is a modified technique to the Gradient Decent GD algorithm. SGD is a supervised machine learning optimization technique that works mainly by selecting random samples in each iteration instead of computing the entire data samples [13]. It is often used in text classification and Natural language processing NLP. Mathematically, SGD have four main steps in the algorithm to ensure reaching the minimum value of the cost function these steps are: Compute Predictions, Compute Loss, Compute Gradients and Update the Parameters.

A linear Support Vector Machine (SVM) algorithm was used along with SGD optimization technique for its effectiveness in this field to implement a multi class classification. The implementation is performed using scikit-learn Python library [14].

Algorithm 1. Stochastic Gradient Decent Algorithm

Input: Training data samples (Xi, Yi) where Xi represents [n-documents, n- Features]
Yi= [n-labels], Alpha α = 0.0001, Learning Rate η = 0.1
Output: X-train = (n-documents, n-Labels)
Begin

1. While stopping criterion is not reached:
2. For Xi in training examples (Xi, Yi)
3. Pick a random document Xi
4. approximates the true gradient by minimize

the regularized training error:

5. $E(w,b) = \frac{1}{n}\sum_{i=1} L\left(y_i, f(x_i)\right) + \alpha\, R(w)$
6. Update model parameters:
7. $w = w - \eta\left[\alpha \frac{\partial\, R(w)}{\partial\, w} + \frac{\partial\, L\left(w^T x_i + b, y_i\right)}{\partial\, w}\right]$
8. End For

End

After training the system, our goal in this stage is to upload a new video to the system and predict what category the video belongs to. This is done by first, upload the video by the admin to the system then apply pre-processing stage by extracting the audio feature from the video, convert audio feature extracted to a textual form then apply text pre-processing processes. Perform TF-IDF to extract features from the document. The stochastic gradient decent algorithm will predict the video according to the trained model built in previously. See algorithm 2 that illustrate the steps of the prediction stage. See algorithm 2 that illustrate the steps of the prediction stage.

Algorithm 2. Prediction Stage of SGD Algorithm

Input: Video
Output: (*Document, Predicted Label*)
Begin

1. Upload the video.
2. Extract Audio Feature.
3. Convert Audio to text document.
4. Perform pre-processing Processes on the text document.
5. Transform the document into word-vector using **TfidfVectorizer and TfidfTransform.**
6. Load the trained model.
7. Predict the document using **Predict(Text)** method.

End

3.4. Retrieval Stage

A simple yet powerful algorithm was implemented in this stage provided by Django python library to help the student navigate through the system to search for the videos that serves the student queries. A search engine was placed in the educational platform that made the access process easier to students.

Students enter the query keywords in the search box first, then the searching algorithm send a message to the database where all videos are stored and classified according to the previous training and prediction stages. Our search algorithm depends essentially on the audio feature represented by the voice of the teacher in the video. The audio feature is being pre-processed as mentioned in the pre-processing section and converted to a text form. These text documents are stored in the database along with the videos. Keywords provided by the student are being searched through the entire text document of the video rather than searching for only the keywords presented in the title of the video or hashtags written manually by the uploader of the videos.

3.5. Evaluation Stage

The evaluation process uses Standard measures to evaluate the performance of the model. The evaluation of text classification depends on four standard measures. These four standard measures are classification accuracy (CA), precision (P), recall (R), and F1-score.

Accuracy is the rate of the correct predicted positive samples to the total number of samples.

$$Accurcy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{Correct\ predicted\ samples}{Total\ Samples}$$

(2)

Precision is the rate of the correct predicted positive samples to the total number of predicted positive samples.

$$Precision = \frac{TP}{TP+FP} = \frac{True\ positive\ samples}{Total\ predicted\ positive\ Samples}$$

(3)

Recall is the rate of the correct predicted positive samples to all samples in actual class.

$$Recall = \frac{TP}{TP+FN} = \frac{True\ positive\ samples}{Total\ actual\ positive\ Samples}$$

(4)

The F1- score is a function used when a balance between precision and recall is needed.

$$F1 - Score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \quad (5)$$

5. Results and Discussions

In this section, our experiment aims to compare the performance of SGD, Neural Network, KNN and Logistic Regression classifiers to classify the dataset into six categories in order to facilitate the retrieval process for the students. A total of 70 educational videos have been preprocessed and converted to a text documents to be prepared for the prediction stage where videos were classified.

Testing dataset, which represent a 30% of the total dataset, was used to evaluate the performance of the system by using the evaluation matrices and cross validation method. In this experiment, the Scores of precision, recall, and F1 metrics were calculated with a 5-fold cross validation and compared to different machine learning models such as Neural Networks (NN), Logistic Regression LR, and k-nearest neighbor's algorithm KNN. SGD model scores the highest F1-Score with 96.6% as shown in Table 2.

Table 2. Evaluation Results of the Performance of the Classification Models

Model	CA	F1	Precision	Recall
SGD	0.966	0.966	0.968	0.966
Neural Network	0.963	0.964	0.965	0.963
Logistic Regression	0.950	0.951	0.956	0.950
kNN	0.447	0.503	0.878	0.447

The evaluation Results showed the highest classification accuracy with 96.6 % for SGD algorithm, which makes the SGD algorithm, is the most suiTable for the presented work. In addition, Confusion matrix was calculated where the actual documents are the correctly classified documents, whereas the predicted documents are the misclassified ones. As an example, the confusion matrix shows that class Chemistry, has (76/82) of its documents correctly classified which is equivalent to (92.7 %) of the documents. While six documents were misclassified (6/82) equivalent to (7.3 %) and had been predicted as Physics. The result of the Confusion matrix and Proportion of the actual classified documents are shown in Table 3 in addition to Table 4. Also Figure 4 and Figure 5 Analyze the performance of the algorithm.

Table 3. Confusion matrix of SGD Classification Model for the Proposed Dataset

		Predicted						
Actual		**Chemistry**	**Math**	**Religion**	**Physics**	**Biology**	**Arabic Grammar & Literature**	∑
	Chemistry	76	0	0	6	0	0	82
	Math	0	45	0	1	0	0	46
	Religion	0	0	35	0	0	0	35
	Physics	0	1	0	67	1	0	69
	Biology	0	0	0	1	58	0	59
	Arabic Grammar & Literature	0	0	0	0	0	9	9
	∑	76	46	35	75	59	9	300

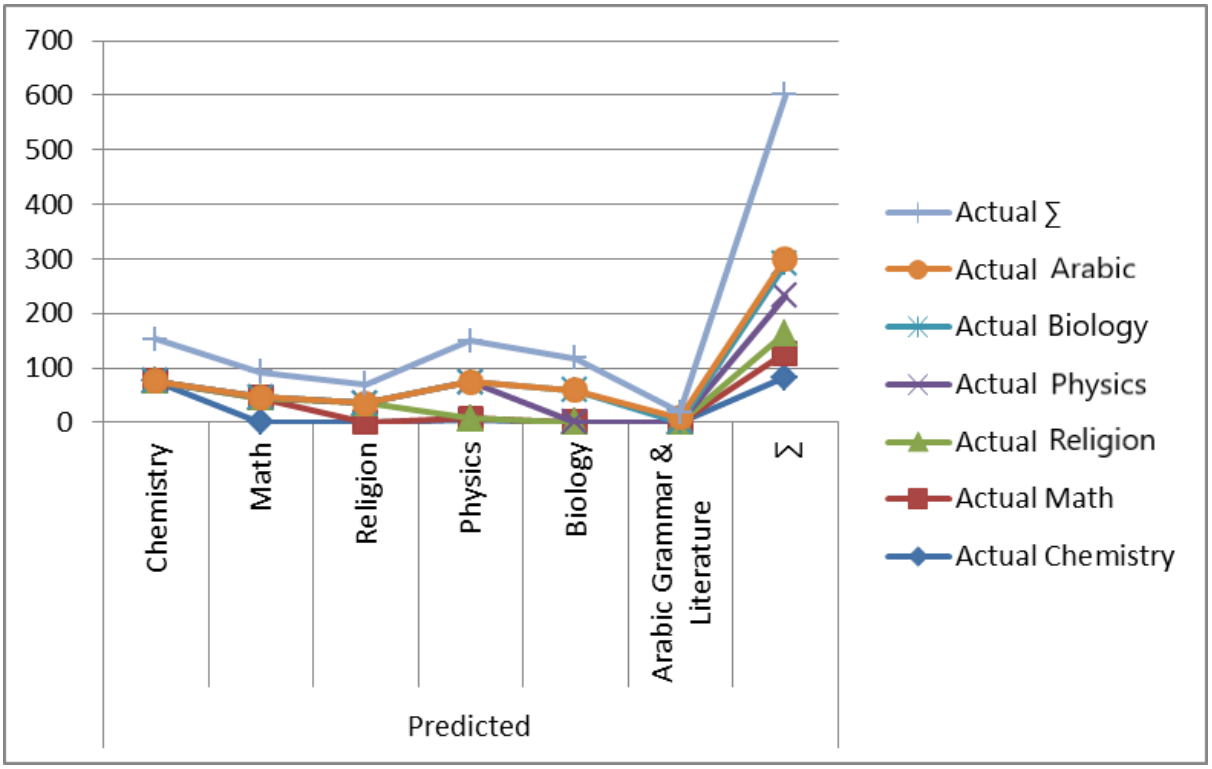

Figure 4. Analysis of the Confusion Matrix of SGD Classification Model

Table 4. Proportion of the Actual Classified Documents

		Predicted						
Actual	**Category**	**Chemistry**	**Math**	**Religion**	**Physics**	**Biology**	**Arabic Grammar & Literature**	**∑**
	Chemistry	100.00%	0.00 %	0.00%	8.00%	0.00%	0.00%	**82**
	Math	0.00%	97.8 %	0.00%	1.30%	0.00%	0.00%	**46**
	Religion	0.00%	0.00 %	100.00 %	0.00%	0.00%	0.00%	**35**
	Physics	0.00%	2.20 %	0.00%	89.30 %	1.70%	0.00%	**69**
	Biology	0.00%	0.00 %	0.00%	1.30%	98.30 %	0.00%	**59**
	Arabic Grammar & Literature	0.00%	0.00 %	0.00%	0.00%	0.00%	100.00%	**9**
	∑	**76**	**46**	**35**	**75**	**59**	**9**	**300**

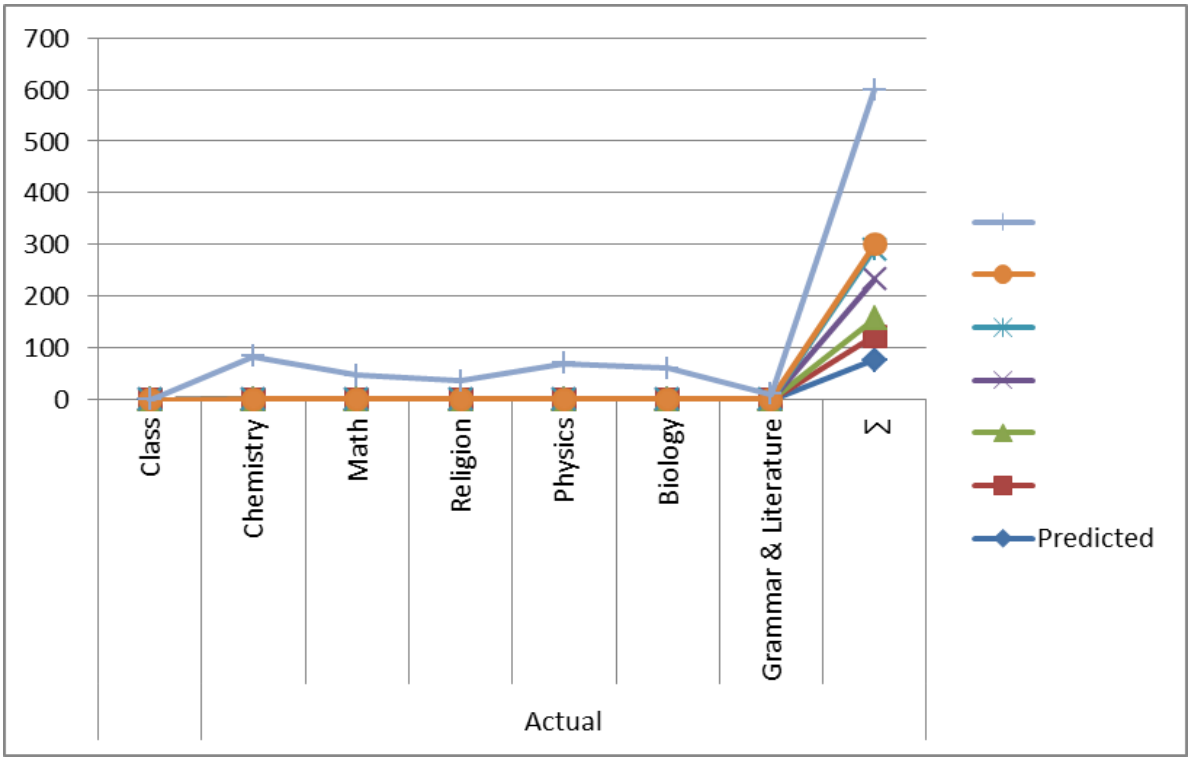

Figure 5. Analysis of the Proportion of the Actual Classified Documents

6. Conclusions

In this paper, an Arabic videos classification and retrieval system based on different machine learning techniques was proposed. An e-learning platform was introduced to make subjects easily available to all students. A contribution was introduced to classify Arabic educational videos based on SGD technique then retrieve these videos based on the students query. Limitations and challenges were

discussed which led to a conclusion that the distinctive feature of educational videos spoken in Arabic language is the Knowledge spoken by the teacher. Speech gives a rich information about educational videos that is why audio feature was extracted from the video then converted to Arabic text. SGD technique was applied on the extracted text in which accurate results were obtained.

Acknowledgements

The authors would like to thank the MUSTANSIRIYAH UNIVERSITY (www.uomustansiriyah.edu.iq) Baghdad-Iraq for affording this honorable opportunity and for providing us with the necessary resources that made the study more prod

References

1. Vijayakumar, V., & Nedunchezhian, R. (2012). A study on video data mining. *International Journal of Multimedia Information Retrieval*, 1(3), 153–172. https://doi.org/10.1007/s13735-012-0016-2.
2. Toriah, S. T. M., Ghalwash, A. Z., & Youssif, A. A. A. (2018). Semantic-Based Video Retrieval Survey. Journal of Computer and Communications, 06(08), 28–44. https://doi.org/10.4236/jcc.2018.68003
3. Ibrahim, Z. A. A., Haidar, S., & Sbeity, I. (2019). Large-scale Text-based Video Classification using Contextual Features. European Journal of Electrical Engineering and Computer Science, 3(2). https://doi.org/10.24018/ejece.2019.3.2.68
4. Zhang, F., Liu, D., & Liu, C. (2020). MOOC Video Personalized Classification Based on Cluster Analysis and Process Mining. Sustainability, 12(7), 3066. https://doi.org/10.3390/su12073066
5. Kastrati, Z., Imran, A. S., & Kurti, A. (2019). Transfer Learning to Timed Text Based Video Classification Using CNN. Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics - WIMS2019. https://doi.org/10.1145/3326467.3326483
6. Medida, L. H. & Ramani, K. (2019). An Optimized E-Lecture Video Retrieval based on Machine Learning Classification. International Journal of Engineering and Advanced Technology, 8(6), 4820–4827. https://doi.org/10.35940/ijeat.f9114.088619
7. Chatbri, H., McGuinness, K., Little, S., Zhou, J., Kameyama, K., Kwan, P., & O'Connor, N. E. (2017). Automatic MOOC Video Classification using Transcript Features and Convolutional Neural Networks. Proceedings of the 2017 ACM Workshop on Multimedia-Based Educational and Knowledge Technologies for Personalized and Social Online Training, 21–26. https://doi.org/10.1145/3132390.3132393
8. Boukil, S., Biniz, M., Adnani, F. E., Cherrat, L., & Moutaouakkil, A. E. E. (2018). Arabic Text Classification Using Deep Learning Technics. International Journal of Grid and Distributed Computing, 11(9), 103–114. https://doi.org/10.14257/ijgdc.2018.11.9.09
9. Sundus, K., Al-Haj, F., & Hammo, B. (2019). A Deep Learning Approach for Arabic Text Classification. 2019 2nd International Conference on New Trends in Computing Sciences (ICTCS), 1–7. https://doi.org/10.1109/ictcs.2019.8923083
10. Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, & Brown. (2019). Text Classification Algorithms: A Survey. Information, 10(4), 150. https://doi.org/10.3390/info10040150
11. Diab, S. (2019). Optimizing Stochastic Gradient Descent in Text Classification Based on Fine-Tuning Hyper-Parameters Approach. A Case Study on Automatic Classification of Global Terrorist Attacks, 16(12), 1–6. [Online]. Available: https://arxiv.org/abs/1902.06542
12. Kaoudi, Z., Quiane-Ruiz, J.-A., Thirumuruganathan, S., Chawla, S., & Agrawal, D. (2017). A Cost-based Optimizer for Gradient Descent Optimization. Proceedings of the 2017 ACM International Conference on Management of Data. https://doi.org/10.1145/3035918.3064042
13. Sun, S., Cao, Z., Zhu, H., & Zhao, J. (2020). A Survey of Optimization Methods From a Machine Learning Perspective. IEEE Transactions on Cybernetics, 50(8), 3668–3681. https://doi.org/10.1109/tcyb.2019.2950779
14. Kosmajac, D., & Keselj, V. (2017). DalTeam@INLI-FIRE-2017: Native Language Identification using SVM with SGD training. CEUR Workshop Processing 2017. 2036, 118–1.