

## Sentiment analysis of YouTube videos comments for children using machine learning and deep learning

Amal Alrehaili, Abdullah Alsaeedi, Wael M.S. Yafouz

Department of Computer Science, College of Computer Science and Engineering, Taibah University, Medina, Saudi Arabia

### Article Info

#### Article history:

Received Oct 19, 2024

Revised Jun 15, 2025

Accepted Jul 3, 2025

#### Keywords:

Deep learning  
Machine learning  
Sentiment analysis  
User comments  
YouTube

### ABSTRACT

Nowadays, online connectivity is increasing with the rapid growth of the world wide web. Consequently, content shared across numerous platforms varies in appropriateness. It is necessary to ensure the suitability of the content since children are among the consumers of online content. A lot of children watch videos on YouTube these days, and such platforms can contain useful content. However, such videos can also have a negative impact on children. The suitability of these videos can be determined through sentiment analysis to refine the content for children on YouTube, by classifying the posted comments as either positive or negative. Therefore, this study utilizes natural language processing methods, machine learning classifiers (MLCs) and deep learning models (DLMs) to detect and classify negative user comments using the proposed dataset. Different MLCs such as random forest (RF), logistic regression (LR), multinomial Naïve Bayes (MNB), decision tree (DT), K-nearest neighbour (KNN), AdaBoost, and support vector machine (SVM) have been used. Additionally, DLMs were also used such as artificial neural network (ANN), convolutional neural network (CNN) and long short-term memory (LSTM). Overall, the experimental results showed that the LR, RF, AdaBoost, ANN and LSTM classifiers outperformed all the other classifiers in terms of accuracy.

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



### Corresponding Author:

Amal Alrehaili

Department of Computer Science, College of Computer Science and Engineering, Taibah University  
Medina, Saudi Arabia.

Email: amal.alrehaily@gmail.com

## 1. INTRODUCTION

In this day and age, we are living in an era where social media is an essential part of our day-to-day life. This era of social media contains several prominent social media networks, on which users are able to express their opinions and emotions constantly in the form of microblogging. An example of these various social media networks would be platforms such as Facebook, YouTube, Twitter and Instagram. Such posts and interactions with other posts can be utilised to make relevant recommendations to users, therefore being useful to the user on a daily basis. Due to this, research based on users' feelings gained wide attention lately using Sentiment Analysis. Additionally, the process of detecting a user feeling has been proposed in the NLP area in order to study the attitude of the user and the overall feelings of the user, this is commonly referred to as opinion mining/sentiment analysis, this area first picked up traction in the 2000s. However, it depends on natural language processing (NLP) which was first developed started in the 1950s [1]. Sentiment analysis has become a main research field in the NLP community; therefore, it sports a wide spectrum of practical applications that include opinion mining and emotion extraction in social media trend predictions [2], [3].

Sentiment analysis is defined as a field of study which detects and analyzes peoples' positive and negative sentiments or opinions about a particular entity. This process uses textual data from different

sources to be used automatically for analyzation through algorithms. This concept is applied in various affairs such as YouTube videos [1]. Sentiment classification techniques can be divided into three major categories: machine learning (ML), ensemble learning, and deep learning (DL), as shown in Figure 1. Importantly, sentiment analysis is mainly utilized to identify and categorize peoples' opinions on a specific topic to express their feelings, perceptions, and opinions as well as to discover experiences and opinions from others [4]. The utilization of a sentiment model is important in analyzing user comments, opinions, and other forms of rating such as movie or store reviews for example. All forms of ratings also include feature extraction from users' opinion, and it plays an important role in our daily decision-making about a product or a movie [5]-[7].

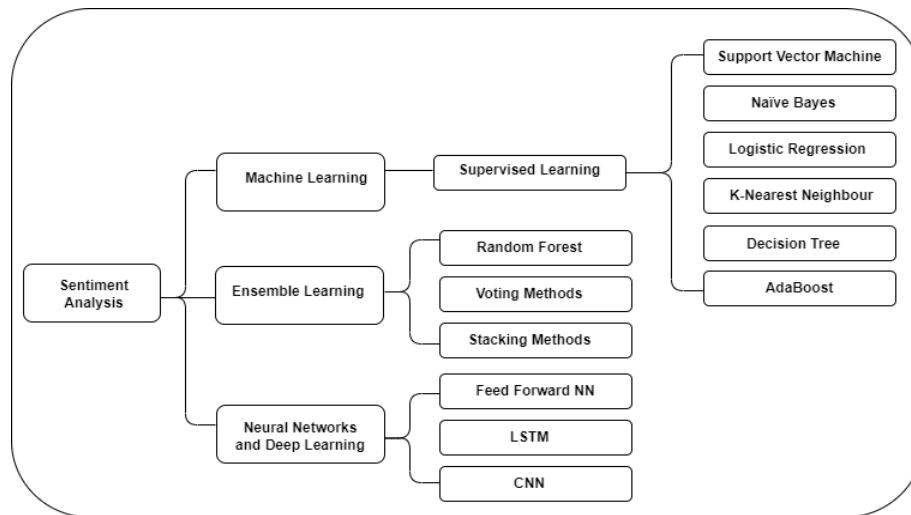


Figure 1. Sentiment classification techniques

YouTube is a platform for sharing videos and allows users to upload, view, comment and share videos [8]. In addition, YouTube includes movies, animation and educational videos. YouTube ranks these types of content based on the number of likes, dislikes and comments [9]-[11]. Also, user comments tend to provide feedback about the quality of the YouTube videos. All these points make Youtube an exceptional platform for SA. YouTube is one of the platforms targeted by children due to its popularity. Children are among online content consumers, and it is necessary to ensure that they access suitable content [12], [13]. Notably, YouTube introduced the YouTube Kids application, which enables parents with the ability to control what their children should and can watch on YouTube. In spite of YouTube's efforts to reduce the spreading of videos unsuitable to children, the disturbing videos still appear to them [14]-[16], and it is essential to ensure that they access suitable content [17]. There are studies that have given attention to this issue to find out whether or not the video is appropriate on YouTube, but the issue continues to exist and requires a more precise solution and improved framework. Currently, the study of SA on English YouTube videos for children is limited. Consequently, there is a need for new methods of inappropriate video content detecting for children using SA techniques.

Recently, several literary studies in the field of sentiment analysis for social media networks were presented utilising supervised ML and DL approaches: Supervised ML, DL, ensemble, and other approaches. Various ML and DL algorithms, such as SVM, DT, RF, LR, and CNN which are available have been used for social media sentiment analysis. Sentiment analysis is still an evolving field and needs to be researched deeply. Though researchers have come up with ways of sentiment analysis with algorithms, the number of these studies is little in the area of social media and especially on YouTube. To sum up, most of the literature reviews have compared the performance of DL algorithms and ML techniques. It consists of data collection, pre-processing data, feature extraction, and classification. Some researchers, for example: have conducted experiments of SA about comments on YouTube using the SVM method [2]. Moreover, compared different algorithms on performance of ML for YouTube data classification but NB has achieved the highest accuracy [18]. Took part in introduction of ML based dual language SA model for YouTube videos ranking in Asia, but LR classifier has achieved 87% accuracy [19]. Ensemble approaches are used to obtain more precise and accurate predictions by combining multiple classifiers. According to Xia *et al.* [20] they studied the impact of utilizing ensemble learners for sentiment classification purposes. While Da Silva *et al.* [21] proposed an

ensemble classifier that relied upon four base classifiers, SVM have achieved the highest accuracy. On the other hand [22] introduced an ensemble system to categorize tweets based on the majority voting of three classifiers and showed various experiments that were conducted to examine the effect of information gain on the accuracy of the classifier and the collected results showed clear improvements in classification accuracy after utilizing information gain for reducing feature vector dimensionality. Alam *et al.* [23] proposed a mobile apps that can help paraent in monitoring their children. Furthermore, DL is used in classification tasks to bring better results used to perform SA of data. According to Ramadhani and Goo [24] their research applied the deep neural network (DNN) to analyze the tweets based on SA, classifies SA using DL method and it is combined with CNN\_LSTM have achieved a higher of 89.2% accuracy [25], while proposed an access model using DL for the development of Kid-Friendly YouTube.

This study evaluates the performance of various MLCs and DLMs for detecting and classifying YouTube video comments into positive and negative categories. Several experiments have been conducted based on a novel proposed Arabic dataset using two stemmers: Snowball and Porter. In these experiments the number of features (2500, 5000, and 10,000) alongside the n-gram (unigram, bigram, and trigram) methods were utilized. Particularly, The ML experiment's results demonstrated that LR, RF, and AdaBoost achieved the highest accuracies of 90% through unigrams when being tested with 5000 features alongside both the stemmers. Moreover, in the DL experiments, three models were put to the test: ANN, CNN and LSTM. CNN and LSTM were tested through word embeddings, while ANN was tested through n-gram methods. The experiment portrayed that the highest accuracies were reached by the ANN model through unigrams and the LSTM model in word embeddings which was 90%, specifically being achieved through 5,000 features alongside the snowball stemmer. The research contributes to detecting the suitability of YouTube videos to children based on video comments, title, and the number of likes through SA to provide safety to children. The main contributions of this research can be summarized as follows:

- Introduced a novel dataset collected from YouTube videos targeted for children.
- Applied the most popular classical machine learning classifiers and deep learning models to investigate the best performance.
- Compared the performance of machine learning classifiers with deep learning models based on the proposed dataset.
- Examined models' performance using porters and snowball stemmer using different features size of extracted features from the proposed dataset.

Therefore, this paper provides a study to the sentiment analysis of children regarding YouTube videos based on user comments. This is done to help parents determine what is and what is not considered a suitable video for their children. Therefore, there are two main experiments which were implemented based on ML classifiers and DL models. It is worthy to mention that in both experiments the same proposed datasets were used.

The rest of the paper is organized as follows. Section 2 explains the research methodology that has been followed to investigate the existing relevant and irrelevant YouTube videos for children through SA based on user comments by using two techniques: ML and DL techniques. Section 3 offers the experimental results of various ML classifiers and DL models. The results discussion explains in section 4. Finally, a conclusion will be covered in section 5.

## 2. METHODS AND MODEL ARCHITECTURE

This section explains the methods that were used to carried out this research. It represents the dataset pre-processing, splitting, and features extraction approaches that will be used in the conducted experiments as shown in Figure 2.

### 2.1. Data collection

Building the dataset consists of several steps which have been executed, they are the dataset collection, data cleaning, and data annotation activities, respectively. Initially, the utilization of comments of various YouTube videos have been extracted using YouTube's API v3. Using the unique identifier (ID) that appears at the end of each YouTube video's URL, as summarized in Table 1. The proposed criteria helps in selecting videos, detecting sentiment based on comments into positive and negative, extracting more knowledge regarding the suitability of videos for children, and to categorize videos into three types: full movies, animation trailers, and inappropriate videos for children.

- The selection of YouTube videos depends on the number of viewers, the number of comments, numbers of likes, and title of video (relevant/irrelevant).
- The most popular video appearing on YouTube trends page for children.

The collected dataset covered four years of uploading YouTube videos (from 2017 to 2021). This approach resulted in the acquisition of more than 14,000 seed comments. Focusing to obtain a set of the most

popular videos, the videos were divided into three categories (animation trailers, full movies, and inappropriate videos for children). The datasets contain all the metadata related to each comment like the user ID, date, time, and the number of likes. Therefore, the data was cleaned manually by removing the duplicated comments and by removing any non-English comments. Hence, 10,272 comments have been deleted from the nine videos that were selected of three different categories. After that, these comments were manually labelled into three classes; positive, negative, and natural, all with the supervision of three expert annotators. However, it is to be mentioned that there has been an exception of this in the neutral class because the number of comments in this class overwhelm the other classes. Finally, the binary dataset which contains 4,456 comments, has been categorized to have 3,012 positive and 1,445 negatives as shown in Figure 3.

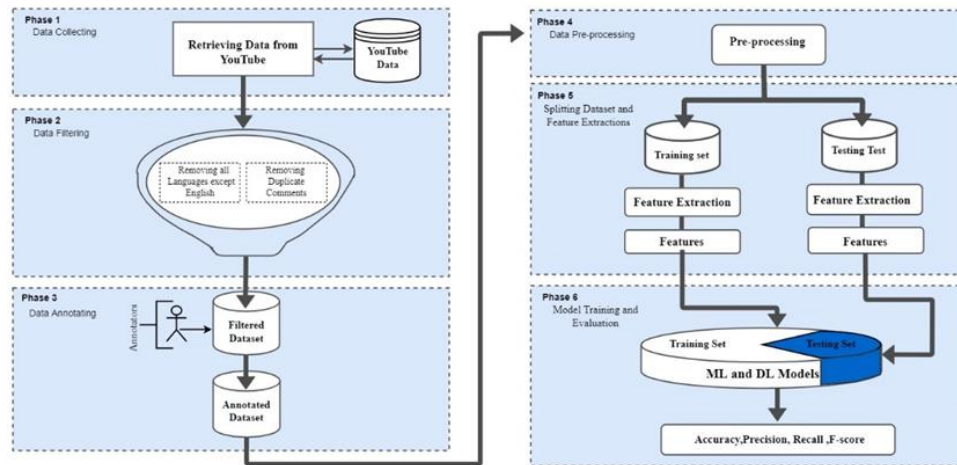


Figure 2. Proposed model

Table 1. YouTube videos used in dataset analysis

#	Date	Id	Categories	Views	Comments (duplicate)	Comments before filtering	Comments (binary)	Positive	Negative
1	Jan 11, 2021	lhxUGk9Mb1A	Inappropriate	17,368,21	43	567	129	35	95
2	Jan16, 2021	6U8HWwTJeCM	Inappropriate	33,807,230	1574	2,213	570	183	387
3	Feb3, 2021	UWw6t1K3Jd4	Inappropriate	5,383,693	2773	711	294	197	97
4	Jul 5, 2020	QO0p-711DJ0	Full Movie	1,472,489	24	407	201	148	53
5	Aug6, 2020	eMoSqI5O9kQ	Tailer animation	2,605,894	3099	777	301	154	147
6	Oct7, 2019	taE3PwurhYM	Tailer animation	4,091,762	38	3,963	2367	1973	394
7	Oct17, 2019	TIZUNqs9hng	Full Movie	1,414,056	3611	291	167	82	85
8	Sep12, 2018	6UiNnW_2STI	Tailer animation	2,456,542	2847	1152	326	147	179
9	Mar 1, 2017	cTa78nRiO24	Full Movie	1,051,358	8	191	101	93	8

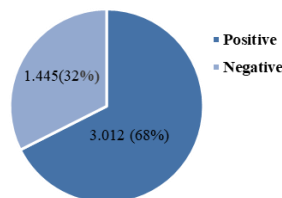


Figure 3. Sentiment analysis of the dataset

## 2.2. Data Pre-processing

In this section, the pre-processing activity is performed on the datasets and this stage includes cleaning the data before training on a model. The comments' pre-processing involves getting rid of all the

undesired and unnecessary words such as the emoji symbol and stop words, it also included stemming. To analyse sentiment, many comments are composed of symbols that need to be replaced by special tags. The processes are described as follows:

- Removing spaces and quotes at the end of the comments.
- Removing special characters such as: / # \$ % ^ & \* | [ ] ; : { } - + ( ) < > ? ! %.
- Removing emoticons and emoji symbols.
- Removing stop words such as: on, in, at, about, am, is, are, was, etc.
- Removing the repetitions of words and letters such as: “This is awssommmmmeee” would become “This is awesome”.
- Converting the comments into lowercase.
- Stemming has been named stem packages and is available in NLTK. Moreover, there are various types of stemming. This is used to return word form to its root namely removing suffix such as -ING, ED, ION-IONS to reduce the words size in dataset and achieve better performance in SA. For Example, the stemming of the words ("Interesting") which means ("Interest").

### 2.3. Feature extraction

The feature extraction step is necessary since it is the input to train ML/DL classifiers. The features are extracted from the comments using TF-IDF and word embedding representations. Unigram, bigram, and trigram features are used. For unigram (1-gram) features, the word appears independent and does not take into consideration other words in the document therefore using n-gram models with  $n=[1,1]$ . While bigram (2-gram) the appearance of the two words joint as one feature by using n-gram models with  $n=[1,2]$ . For trigram (3-gram) features which include both bigram and unigram models with  $n=[1,3]$ . In each of n-grams, specific max\_features 2,500; 5,000 and 10,000 were used.

#### 2.3.1. TF-IDF

TF-IDF is defined as a statistical measure utilized to know the importance of word that appears in a document [26]. TF-IDF called (term frequency–inverse document frequency) is calculated as:

$$TF - IDF = \frac{TF(t,d)}{\log\left(\frac{N}{IDF}\right)} \quad (1)$$

TF is the rate of word frequency in a text/total word frequency in text. Moreover, IDF is  $\log$  (total numbers of text/number of texts that appears the word used in TF).

#### 2.3.2. Word embedding

The word embedding is a technique for determining the syntactic and semantic context of a word by using information from a text corpus. For NLP tasks, this approach has been commonly used. Additionally, it's also known as word representations due to all words in a text corpus which are represented as vectors in different dimensions [27].

### 2.4. Model implementation

In this phase, the dataset was divided into testing and training by k-fold where the number of folds was  $k=5$ , used four for training and one of them for testing as shown in Figure 4. Then, ML and DL techniques were applied to classify videos and comments to compare both techniques in terms of performance. While algorithms that used on ML are seven classifiers: RF, LR, DT, MNB, KNN, Adaboost, and SVM. For the DL models, three models were selected including ANN, CNN and LSTM.



Figure 4. K-fold method [28]

### 2.5. Model evaluation

The aim of this phase is to evaluate the efficiency and performance of all classification models based on computing a confusion matrix which used four binary classification metrics: true positive (TP),

false positive (FP), false negative (FN), true negative (TN) all considering the confusion matrix as shown in Table 2, these were indicated as:

- True positive (TP), refers the positive comments correctly classified as positive.
- False positive (FP), refers the positive comments incorrectly classified as positive.
- False negative (FN), refers the negative comments incorrectly classified as negative.
- True negative (TN), refers the negative comments correctly classified as negative.

Table 2. The Confusion matrix

Actual Class	Prediction class	
	Positive	Negative
	TP FP	FN TN

Through the confusion matrix, the evaluation results are calculated in terms of accuracy, accuracy, recall, and f-score as illustrated in the following part:

Accuracy is a measure of how accurate the classifier is. It's defined by the following formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Precision is a measure of positive comments correctly classified over the total number of positive comments correctly classified and incorrectly classified and it calculated as:

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

While recall is a measure of model performance to detect the correct classifier in the dataset and it computed as:

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

F-measure (F) is a measure to obtain the value between precision and recall, and it is calculated as:

$$F_{measure} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

### 3. EXPERIMENTAL RESULTS

This section presents the experimental results that were conducted using the most common ML classifiers and DL models. There are six total experiments, all of which include the utilization of two stemmers and different feature extraction size in both approaches ML and DL with N-gram methods.

#### 3.1. An experiment on ML classifiers

This section describes and evaluates the results of the ML classifiers on the proposed dataset. Therefore, there are seven ML classifiers which have been selected based on previous studies including RF, LR, MNB, DT, KNN, AdaBoost, and SVM. It is important to highlight that the n-gram features were extracted using TF-IDF. The unigram, bigram, and trigram features were extracted and collected based on a max feature parameter. This was set to 2,500; 5,000 and 10,000. Accuracy, precision, recall, and f-score were all used as evaluation metrics of the model's performances. Two stemming techniques were utilized in the conducted experiments as follows: snowball stemmer and porter stemmer. The following results refer to average and median values that were calculated to investigate the performance of all the classifiers for comment detecting as illustrated in tables, indicate that accuracy, precision, recall, and f-score regarding unigram, bigram, and trigram in terms of average and median values.

##### 3.1.1. Experiment one: ML classifiers with a maximum of 2,500 features

In the first experiment, ML classifiers were used based on different evaluation metrics where the maximum number of features was 2,500 and the snowball stemmer was utilized as shown in Table 3. The results are presented for unigrams, bigrams, and trigrams. It is obvious from the experiments conducted that the LR classifier obtained the highest scores compared to the RF, MNB, DT, KNN, AdaBoost, and SVM. The lowest scores were obtained using the SVM and KNN classifiers in terms of accuracy, precision, recall, and f-score. In terms of unigrams, the LR classifier reached an accuracy of 0.90 compared to accuracies of 0.89, 0.85, 0.85, 0.74, 0.89, and 0.74 obtained using RF, MNB, DT, KNN, AdaBoost, and SVM, respectively.

Table 3. All videos 2,500 features snowball stemmer

Classifier	Feature Extraction	Average				Median			
		Acc	Precision	Recall	Fscore	Acc	Precision	Recall	Fscore
RF	Unigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
LR		<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>
MNB		0.85	0.87	0.85	0.84	0.85	0.87	0.85	0.84
DT		0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
KNN		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.68</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.69</b>
AdaBoost		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
SVM	Bigrams	<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.68</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.69</b>
RF		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>
LR		<b>0.89</b>	<b>0.90</b>	<b>0.89</b>	<b>0.89</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>
MNB		0.86	0.87	0.86	0.86	0.86	0.87	0.86	0.86
DT		0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
KNN		<b>0.74</b>	<b>0.80</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.80</b>	<b>0.75</b>	<b>0.69</b>
AdaBoost	Trigrams	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
SVM		<b>0.74</b>	<b>0.80</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.80</b>	<b>0.75</b>	<b>0.69</b>
RF		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
LR		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>
MNB		0.87	0.88	0.87	0.86	0.86	0.87	0.86	0.85
DT		0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
KNN	Trigrams	<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.80</b>	<b>0.75</b>	<b>0.69</b>
AdaBoost		<b>0.89</b>	<b>0.88</b>	<b>0.89</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>
SVM		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.80</b>	<b>0.75</b>	<b>0.69</b>

In terms of porter stemmer, Table 4 illustrates the evaluation scores computed for various classifiers as the number of features was 2,500. It is clear from the experimental results that the LR classifier obtained the highest accuracy, precision, recall, and f-measure compared to RF, MNB, DT, KNN, AdaBoost, and SVM. In terms of unigram, the LR and RF classifiers reached an accuracy of 0.90 while MNB, DT, KNN, AdaBoost, and SVM attained accuracies of 0.85, 0.85, 0.74, 0.89, and 0.74 in order.

Table 4. All videos 2,500 features porter stemmer

Classifier	Feature Extraction	Average				Median			
		Acc	Precision	Recall	Fscore	Acc	Precision	Recall	Fscore
RF	Unigrams	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
LR		<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>
MNB		0.85	0.87	0.85	0.84	0.85	0.87	0.85	0.84
DT		0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
KNN		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.69</b>
AdaBoost		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.88</b>
SVM	Bigrams	<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.69</b>
RF		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.90</b>	<b>0.89</b>	<b>0.90</b>	<b>0.89</b>
LR		<b>0.89</b>	<b>0.90</b>	<b>0.89</b>	<b>0.89</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>
MNB		0.86	0.87	0.86	0.86	0.86	0.87	0.86	0.86
DT		0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
KNN		<b>0.75</b>	<b>0.80</b>	<b>0.75</b>	<b>0.69</b>	<b>0.75</b>	<b>0.80</b>	<b>0.75</b>	<b>0.69</b>
AdaBoost	Trigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>
SVM		<b>0.75</b>	<b>0.80</b>	<b>0.75</b>	<b>0.69</b>	<b>0.75</b>	<b>0.80</b>	<b>0.75</b>	<b>0.69</b>
RF		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
LR		<b>0.89</b>	<b>0.90</b>	<b>0.89</b>	<b>0.89</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>
MNB		0.87	0.88	0.87	0.86	0.87	0.88	0.87	0.86
DT		0.85	0.85	0.85	0.85	0.84	0.84	0.84	0.84
KNN	Trigrams	<b>0.75</b>	<b>0.80</b>	<b>0.75</b>	<b>0.69</b>	<b>0.75</b>	<b>0.80</b>	<b>0.75</b>	<b>0.69</b>
AdaBoost		<b>0.89</b>	<b>0.88</b>	<b>0.89</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>
SVM		<b>0.75</b>	<b>0.80</b>	<b>0.75</b>	<b>0.69</b>	<b>0.75</b>	<b>0.80</b>	<b>0.75</b>	<b>0.69</b>

### 3.1.2. Experiment two: ML classifiers with 5,000 maximum features

In the second experiment, the performances of various ML classifiers depend on different evaluation metrics where the maximum number of features was 5,000 and the snowball stemmer was utilized as shown in Table 5. The results are presented for the unigrams, bigrams, and trigrams. It is obvious from the experiments

conducted that the LR classifier reached the highest scores compared to RF, MNB, DT, KNN, AdaBoost, and SVM. On other hand, the lowest scores were obtained using the SVM and KNN classifiers in terms of accuracy, precision, recall, and f-score. In terms of unigram, the RF, LR and AdaBoost classifiers all reached an accuracy of 0.90 compared to accuracies of 0.85, 0.84, 0.74, and 0.74 obtained using MNB, DT, and KNN respectively.

Table 5. All Videos 5,000 features snowball stemmer

Classifier	Feature extraction	Average				Median			
		Acc	precision	Recall	Fscore	Acc	precision	Recall	Fscore
RF	Unigrams	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
LR		<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>
MNB		0.85	0.87	0.85	0.84	0.84	0.87	0.84	0.83
DT		0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
KNN		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.68</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.69</b>
AdaBoost		<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>
SVM		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.68</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.69</b>
RF	Bigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>
LR		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.88</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
MNB		0.85	0.87	0.85	0.84	0.85	0.87	0.85	0.84
DT		0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
KNN		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.68</b>
AdaBoost		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
SVM		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.68</b>
RF	Trigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
LR		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.88</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
MNB		0.85	0.87	0.85	0.84	0.85	0.87	0.85	0.84
DT		0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
KNN		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.68</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.68</b>
AdaBoost		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
SVM		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.68</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.68</b>

In terms of porter stemmer, Table 6 illustrates the evaluation scores computed for various classifiers as the maximum number of features was 5,000. It is clear from the experimental results that the LR classifier obtained the highest scores in accuracy, precision, recall, and f-measure, compared to RF, MNB, DT, KNN, AdaBoost, and SVM. In terms of unigram, the RF, LR, and AdaBoost classifiers all reached an accuracy of 0.90 while MNB, DT, KNN, and SVM each attained accuracies of 0.85, 0.85, 0.74, and 0.74 respectively.

Table 6. All videos 5,000 features porter stemmer

Classifier	Feature extraction	Average				Median			
		Acc	Precision	Recall	Fscore	Acc	Precision	Recall	Fscore
RF	Unigrams	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
LR		<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>
MNB		0.85	0.87	0.85	0.84	0.85	0.87	0.85	0.83
DT		0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
KNN		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.68</b>	<b>0.75</b>	<b>0.80</b>	<b>0.75</b>	<b>0.69</b>
AdaBoost		<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>
SVM		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.68</b>	<b>0.75</b>	<b>0.80</b>	<b>0.75</b>	<b>0.69</b>
RF	Bigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
LR		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.88</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
MNB		0.85	0.87	0.85	0.84	0.85	0.87	0.85	0.84
DT		0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
KNN		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.68</b>
AdaBoost		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
SVM		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.68</b>
RF	Trigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
LR		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.88</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
MNB		0.85	0.87	0.85	0.84	0.86	0.87	0.86	0.85
DT		0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
KNN		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.68</b>
AdaBoost		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
SVM		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.68</b>

### 3.1.3. Experiment three: ML classifiers with 10,000 maximum features

In the third experiment, the performance of various ML classifiers was based on different evaluation metrics where the maximum number of features were 10,000 and snowball stemmer was utilized as shown in Table 7. The results are presented for the unigrams, bigrams, and trigrams. It is obvious from the experiments



conducted that the RF, LR and AdaBoost classifiers all obtained the highest scores compared to the MNB, DT, KNN, and SVM classifiers. Specifically, the lowest scores were obtained using SVM and KNN classifiers in terms of accuracy, precision, recall, and f-score. In terms of unigrams, the RF, LR, and AdaBoost classifiers reached an accuracy of 0.90 compared to accuracies of 0.85, 0.85, 0.74, and 0.74 obtained by MNB, DT, KNN, and SVM respectively.

Table 7. All videos 10,000 features snowball stemmer

Classifier	Feature Extraction	Average				Median			
		Acc	Precision	Recall	Fscore	Acc	Precision	Recall	Fscore
RF	Unigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
LR		<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.91</b>	<b>0.90</b>	<b>0.91</b>	<b>0.90</b>
MNB		0.85	0.87	0.85	0.84	0.86	0.87	0.86	0.84
DT		0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
KNN		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.69</b>
AdaBoost		<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>
SVM	Bigrams	<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.69</b>
RF		<b>0.89</b>	<b>0.90</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
LR		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.88</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
MNB		0.83	0.86	0.83	0.81	0.84	0.87	0.84	0.82
DT		0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
KNN		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.68</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.69</b>
AdaBoost	Trigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
SVM		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.68</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.69</b>
RF		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.88</b>
LR		<b>0.88</b>	<b>0.89</b>	<b>0.88</b>	<b>0.88</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
MNB		0.83	0.86	0.83	0.82	0.85	0.87	0.85	0.83
DT		0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
KNN		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.69</b>
AdaBoost		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
SVM		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.69</b>

In terms of porter stemmer, Table 8 reports the evaluation scores which were computed for different classifiers as the maximum number of features was 10,000. It is clear from the experimental results that the RF, LR and AdaBoost classifiers all reached the highest scores in terms of accuracy, precision, recall, and f-measure compared to MNB, KNN, DT, and SVM. In terms of unigrams, the RF, LR and AdaBoost classifiers reached an accuracy of 0.89, 0.90 and 0.90 compared to the accuracies of 0.85, 0.85, 0.74, and 0.74 obtained by MNB, DT, KNN, and SVM respectively.

Table 8. All videos 10,000 features porter stemmer

Classifier	Feature extraction	Average				Median			
		Acc	Precision	Recall	Fscore	Acc	Precision	Recall	Fscore
RF	Unigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
LR		<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>
MNB		0.85	0.87	0.85	0.84	0.86	0.87	0.86	0.85
DT		0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
KNN		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.69</b>
AdaBoost		<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>
SVM	Bigrams	<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.69</b>
RF		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
LR		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.88</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
MNB		0.83	0.86	0.83	0.81	0.84	0.87	0.84	0.83
DT		0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
KNN		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.68</b>	<b>0.75</b>	<b>0.80</b>	<b>0.75</b>	<b>0.69</b>
AdaBoost	Trigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
SVM		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.68</b>	<b>0.75</b>	<b>0.80</b>	<b>0.75</b>	<b>0.69</b>
RF		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>
LR		<b>0.88</b>	<b>0.89</b>	<b>0.88</b>	<b>0.88</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
MNB		0.84	0.87	0.84	0.82	0.85	0.87	0.85	0.83
DT		0.85	0.85	0.85	0.85	0.86	0.86	0.86	0.86
KNN		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.69</b>
AdaBoost		<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>
SVM		<b>0.74</b>	<b>0.81</b>	<b>0.74</b>	<b>0.69</b>	<b>0.75</b>	<b>0.81</b>	<b>0.75</b>	<b>0.69</b>

### 3.2. An experiment on DL models

This section presents the experiment results based on DL models. This is in order to evaluate the performance of the experiment. There are three classifiers which have been selected based on previous studies including CNN, LSTM, and ANN. To measure the performance of these classifiers, it will depend on TF-IDF and word embedding to extract the words. The following results refer to average and median values that were calculated to investigate the performance of all classifiers for comment detecting as illustrated in Tables, indicate that accuracy, precision and recall, and f-score regarding word embedding in terms of average and median values. The comments in the collected datasets were collected from nine different videos. The features were extracted using TF-IDF for ANN models using word embeddings for CNN and LSTM models. Two stemming techniques were utilized in the conducted experiments as follows: snowball stemmer and porter stemmer.

#### 3.2.1. Experiment four: DL models with 2,500 maximum features and word embedding

In the fourth experiment, the performance of various DL classifiers based on different evaluation metrics are shown. The maximum number of features was 2,500 and snowball stemmer was utilized as portrayed in Table 9. It is clear from the experimental results the ANN and LSTM classifiers obtained the highest scores in terms of accuracy, precision, recall, and f-score. Specifically, the ANN and LSTM classifiers both reached an accuracy of 0.89 while CNN attained an accuracy of 0.70.

In terms of porter stemmer, Table 10 illustrates the evaluation scores computed for various classifiers as the maximum number of features was 2,500. It is clear from the experimental results that the ANN and LSTM classifiers both obtained the highest scores in terms of accuracy, precision and recall, and f-score. Specifically, the ANN and LSTM classifiers reached an accuracy of 0.89 while CNN attained an accuracy of 0.73.

Table 9. All Videos 2,500 features snowball stemmer

Classifier	Feature extraction	Average				Median			
		Acc	Precision	Recall	Fscore	Acc	Precision	Recall	Fscore
ANN	Unigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
	Bigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
	Trigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
CNN	Word Embedding	<b>0.70</b>	<b>0.68</b>	<b>0.70</b>	<b>0.66</b>	<b>0.70</b>	<b>0.69</b>	<b>0.70</b>	<b>0.66</b>
LSTM		<b>0.89</b>	<b>0.88</b>	<b>0.89</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>

Table 10. All videos 2,500 features porters stemmer

Classifier	Feature extraction	Average				Median			
		Acc	Precision	Recall	Fscore	Acc	Precision	Recall	Fscore
ANN	Unigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
	Bigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
	Trigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
CNN	Word Embedding	<b>0.73</b>	<b>0.71</b>	<b>0.73</b>	<b>0.70</b>	<b>0.73</b>	<b>0.72</b>	<b>0.73</b>	<b>0.72</b>
LSTM		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.88</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>

#### 3.2.2. Experiment Five: DL models with 5,000 maximum features and word embedding

In the fifth experiment, the performance of various DL classifiers when the maximum number of features was 5,000 and snowball stemmer was utilized are shown in Table 11. It is obvious from the experiments conducted that the ANN and LSTM classifiers obtained the highest scores compared to CNN in terms of accuracy, precision, recall, and f-score. In terms ANN and LSTM classifiers, they both reached accuracies of 0.90, while CNN attained an accuracy of 0.70.

Table 11. All videos 5,000 features snowball stemmer

Classifier	Feature extraction	Average				Median			
		Acc	Precision	Recall	Fscore	Acc	Precision	Recall	Fscore
ANN	Unigrams	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>	<b>0.90</b>	<b>0.89</b>
	Bigrams	0.89	0.89	0.89	0.89	0.89	0.88	0.89	0.88
	Trigrams	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89
CNN	Word Embedding	<b>0.70</b>	<b>0.69</b>	<b>0.70</b>	<b>0.66</b>	<b>0.70</b>	<b>0.69</b>	<b>0.70</b>	<b>0.65</b>
LSTM		<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>

In terms of porter stemmer, Table 12 shows the evaluation scores computed for various classifiers when the maximum number of features was 5,000. It is clear from the experimental results that the ANN and LSTM classifiers obtained the highest scores in terms of accuracy, precision, recall, and f-score. In terms ANN and LSTM classifiers reached an accuracy of 0.89 while CNN attained an accuracy of 0.73.

Table 12. All videos 5,000 features porters stemmer

Classifier	Feature extraction	Average				Median			
		Acc	Precision	Recall	Fscore	Acc	Precision	Recall	Fscore
ANN	Unigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
	Bigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
	Trigrams	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
CNN	Word Embedding	<b>0.73</b>	<b>0.72</b>	<b>0.73</b>	<b>0.71</b>	<b>0.74</b>	<b>0.74</b>	<b>0.74</b>	<b>0.72</b>
LSTM		<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.90</b>	<b>0.89</b>	<b>0.90</b>	<b>0.89</b>

### 3.2.3. Experiment six: DL Models with 10,000 maximum features and word embedding

In the sixth experiment, the performance of various DL classifiers based on different evaluation metrics where the maximum number of features was 10,000 and snowball stemmer was utilized as shown in Table 13. It is obvious from the experiments conducted that the LR and LSTM classifiers obtained the highest scores compared to CNN in terms of accuracy, precision, recall, and f-measure. Specifically, ANN and LSTM classifiers reached accuracies of 0.90 and 0.88, while CNN attained an accuracy of 0.72 respectively.

Table 13. All videos 10,000 features snowball stemmer

Classifier	Feature extraction	Average				Median			
		Acc	Precision	Recall	Fscore	Acc	Precision	Recall	Fscore
ANN	Unigrams	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
	Bigrams	0.89	0.89	0.89	0.89	0.88	0.88	0.88	0.88
	Trigrams	0.89	0.89	0.89	0.88	0.89	0.89	0.89	0.89
CNN		<b>0.72</b>	<b>0.71</b>	<b>0.72</b>	<b>0.68</b>	<b>0.71</b>	<b>0.69</b>	<b>0.71</b>	<b>0.66</b>
LSTM	Word Embedding	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>

In terms of porter stemmer, Table 14 reports the evaluation scores computed for various classifiers when the maximum number of features was 10,000. It is clear from the experimental results that the ANN and LSTM classifiers obtained the highest scores compared to CNN in terms of accuracy, precision and recall, and f-score. Notably, ANN and LSTM classifiers, reached accuracies of 0.90 and 0.88 while CNN attained an accuracy of 0.77, respectively.

Table 14. All videos 10,000 features porters stemmer

Classifier	Feature extraction	Average				Median			
		Acc	Precision	Recall	Fscore	Acc	Precision	Recall	Fscore
ANN	Unigrams	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>
	Bigrams	0.89	0.89	0.89	0.89	0.88	0.88	0.88	0.88
	Trigrams	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89
CNN		<b>0.77</b>	<b>0.76</b>	<b>0.77</b>	<b>0.74</b>	<b>0.74</b>	<b>0.73</b>	<b>0.74</b>	<b>0.71</b>
LSTM	Word Embedding	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>	<b>0.89</b>	<b>0.88</b>	<b>0.88</b>

Generally, the ANN classifier using the snowball and Porters stemmer through unigrams achieved the highest accuracy on both cases of the experiment. This can be further seen through viewing the results of the experiments through all the different amounts of Features the classifier was tested on and comparing the accuracies of the other classifiers. This is all as seen in Tables 9-14.

## 4. RESULTS DISCUSSION

In this section, the results of retrieving YouTube videos for children based on SA were reached through two main experiments, and the results were discussed and analyzed. These experiments were

conducted based on ML and DL models. Additionally, a comparison of the results has been made. The results associated with experiments of ML showed that the LR, RF, and AdaBoost classifiers reached the highest score in terms of accuracy of 90%. On the other hand, the SVM and KNN classifiers obtained the lowest-performing scores for accuracy with those scores being 74%.

In accordance with the previous studies' results, Yasin *et al.* [19] have demonstrated that the LR classifier achieves 87% accuracy which is slightly close to the present LR classifier accuracy score of 90%, while it has been illustrated that the RF and SVM classifier both achieved 82%, Yafooz and Alsaeedi [2] has concluded that the SVM classifier achieved an accuracy of 84% and has concluded that the KNN classifier achieved 61% in terms of accuracy. which contrast in the experiment done by [14]. Moreover, the results attained from the conducted experiments which utilized DL models showed that the ANN and LSTM classifiers both reached the highest accuracy score of 90%. While the CNN classifier obtained the lowest accuracy score of 77%. In accordance with the previous studies results, it has been illustrated that the CNN classifier achieved an accuracy of 85.7% [29] and demonstrated that the CNN classifier achieved an accuracy of 88% which differs from what was found in this experiment [30]. In the experimental results of machine learning, the RF, LR and AdaBoost all achieved equal accuracies with the ANN and LSTM classifiers of deep learning. Thus, the RL, RF, AdaBoost, ANN, and LSTM classifiers have outperformed other classifiers in terms of accuracy. Notably, the dataset that was used in the related studies differ from the dataset in this research. However, in accordance with the results of previous studies, the RF, LR, AdaBoost, ANN, and LSTM are the best performance models.

## 5. CONCLUSION

Sentiment analysis techniques are a process for detecting how the user feels. Recently, this was applied in several fields. This research project applied sentiment analysis to construct a model that classifies YouTube video comments into positive and negative categories. This is to detect the suitability of videos for children. Moreover, the dataset was collected based on comments retrieved from YouTube using an API, manual annotation was applied for the comments. The annotation process has been carried out on the introduced dataset. Additionally, the experiment models were applied using two approaches, ML and DL, based on sentiment analysis techniques to evaluate the performance both of ML and DL classifiers. ML classifiers such as RF, LR, MNB, DT, KNN, AdaBoost, and SVM were applied. While DL classifiers such as ANN, CNN and LSTM have been applied. Moreover, three types of n-gram features were applied on ML by using TF-IDF. However, in DL, the word embedding was applied on CNN and LSTM. Whereas the ANN used three types of n-gram features which were applied to evaluate the performance of both of ML and DL classifiers. The experimental results indicated the LR, RF, AdaBoost, ANN, and LSTM outperformed other classifiers in terms of accuracy.

The results of the experiments, specifically in ML, show that the LR, RF and AdaBoost classifiers all had the best accuracies in classifying Youtube video comments through unigrams when being tested with 5,000 features alongside both the stemmers (porters and snowball). Additionally, the DL experiments show that the highest accuracies were in general achieved by ANN though unigrams, followed by LSTM through word embeddings when tested by 5,000 features alongside the snowball stemmer. For future work, several ideas could help improve the framework. Using sentiment analysis with video-to-text approaches could identify and filter irrelevant videos. Combining Naïve Bayes with SVM could improve accuracy in sentiment classification and use the pretrained models will improve the model accuracy. Additionally, testing the model on an Arabic dataset could explore sentiment analysis in a new domain.

## FUNDING INFORMATION

Authors state no funding involved.

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Amal Alrehaili	✓	✓	✓	✓	✓		✓	✓	✓		✓			
Abdullah Alsaeedi	✓	✓		✓		✓	✓	✓	✓	✓		✓	✓	✓
Wael M.S. Yafooz	✓			✓		✓			✓	✓		✓	✓	✓

C : Conceptualization	I : Investigation	Vi : Visualization
M : Methodology	R : Resources	Su : Supervision
So : Software	D : Data Curation	P : Project administration
Va : Validation	O : Writing - Original Draft	Fu : Funding acquisition
Fo : Formal analysis	E : Writing - Review & Editing	

## CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## DATA AVAILABILITY

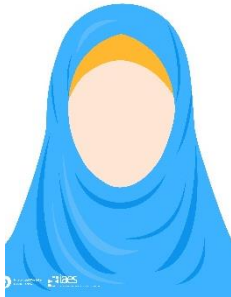
- Data availability is not applicable to this paper as no new data were created or analyzed in this study.




## REFERENCES

- [1] L.Yue, W. Chen, X. Li, W. Zuo and M. Yin, "A survey of sentiment analysis in social media," *Knowledge and Information Systems*, vol. 60, pp. 617-663, 2019, doi:10.1007/s10115-018-1236-4.
- [2] W. M. Yafooz and A. Alsaedi, "Sentimental analysis on health-related information with improving model performance using machine learning," *Journal of Computer Science*, vol.17, no.2, pp.112-122,2021, doi:10.3844/jcssp.2021.112.122.
- [3] O. Oueslati, E. Cambria, M. B. HajHmida, and H. Ounelli, "A review of sentiment analysis research in Arabic language," *Future Generation Computer Systems*, pp.408-430, 2020, doi: 10.48550/arXiv.2005.12240.
- [4] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F., Chang and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, pp. 3-14, 2017, doi: 10.1016/j.imavis.2017.08.003.
- [5] F. I. Tanesab, I. Sembiring and H. D. Purnomo, "Sentiment analysis model based on youtube comment using support vector machine," *International Journal of Computer Science and Software Engineering (IJCSSE)*, vol.6, no. 8, pp.180-185, 2017, doi: 10.1109/ICOMITEE.2019.8920923.
- [6] M. H. Abd El-Jawad, R. Hodhod and Y. M. K. Omar, "Sentiment analysis of social media networks using machine learning," *2018 14th International Computer Engineering Conference (ICENCO)*, Cairo, Egypt, 2018, pp. 174-176, doi: 10.1109/ICENCO.2018.8636124.
- [7] M. Birjali, M. Kasri, A. B. Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Systems*, vol.226, pp.107-134, 2021, doi:10.1016/j.knosys.2021.107134.
- [8] M. Z. Asghar, S. Ahmad, A. Marwat, and F. M. Kundi, "Sentiment analysis on youtube: A brief survey," *arXiv preprint, arXiv:1511.0914*, 2015, doi:10.48550/arXiv.1511.09142.
- [9] A. Alrehaili, A. Alsaedi and W. Yafooz, "Sentiment analysis on YouTube videos for kids: review," *9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pp.1-5, 2021, doi: 10.1109/ICRITO51393.2021.9596364.
- [10] R. Novendri, A. S. Callista, D. N. Pratama, and C. E. Puspita, "Sentiment analysis of YouTube movie trailer comments using Naïve Bayes," *Bulletin of Computer Science and Electrical Engineering*, vol. 1, no.1, pp.26-32, 2020, doi: 10.25008/bcsee.v1i1.5.
- [11] G. S. Chauhan and Y. K. Meena, "YouTube video ranking by aspect-based sentiment analysis on user feedback," *Soft Computing and Signal Processing*, pp. 63-71, 2019, doi: 10.1007/978-981-13-3600-3\_6.
- [12] R. Pradhan, "Extracting sentiments from YouTube comments," *2021 Sixth International Conference on Image Information Processing (ICIIP)*, Shimla, India, 2021, pp. 1-4, doi: 10.1109/ICIIP53038.2021.9702561.
- [13] S. I. Alqahtani, W. M. S. Yafooz, A. Alsaedi, L. Syed, and R. Alluhaibi, "Children's Safety on YouTube: A Systematic Review," *Applied Sciences*, vol. 13, no. 6, p. 4044, 2023, doi: 10.3390/app13064044.
- [14] A. Khan et al., "Sentiment classification of user reviews using supervised learning techniques with comparative opinion mining perspective," *Science and Information Conference*, pp. 23-29, 2019, doi: 10.1007/978-3-030-17798-0\_3.
- [15] R. Tahir, F. Ahmed, H. Saeed, S. Ali, F. Zaffar, and C. Wilson, "Bringing the kid back into YouTube kids: detecting inappropriate content on video streaming platforms," *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Vancouver, BC, Canada, 2019, pp. 464-469, doi: 10.1145/3341161.3342913.
- [16] S. Reddy, N. Srikanth, and G. Sharvani, "Development of kid-friendly youtube access model using deep learning," *Data Science and Security*, pp. 243-250, 2020, doi: 10.1007/978-981-15-5309-7\_26.
- [17] S. Alghowinem, "A safer youtube kids: An extra layer of content filtering using automated multimodal analysis," *Proceedings of SAI Intelligent Systems Conference*, pp. 294-308, 2018, doi: 10.1007/978-3-030-01054-6\_21.
- [18] R. Amanda and E. S. Negara, "Analysis and implementation machine learning for YouTube data classification by comparing the performance of classification algorithms," *Jurnal Online Informatika*, vol. 5, no. 1, pp. 61-72, 2020, doi: 10.15575/join.v5i1.505.
- [19] S. Yasin, K. Ullah, S. Nawaz, M. Rizwan and Z. Aslam, "Dual language sentiment analysis model for YouTube videos ranking based on machine learning techniques," *Pakistan Journal of Engineering and Technology*, vol. 3, no. 2, pp. 213- 218, 2020.
- [20] R. Xia, C. Zong and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Information sciences*, vol. 181, no.6, pp.1138-1152, 2011, doi: 10.1016/j.ins.2010.11.023.
- [21] N. F. Da Silva, E.R. Hruschka and E. R. Hruschka Jr, "Tweet sentiment analysis with classifier ensembles," *Decision Support Systems*, pp.170-179, 2014, doi: 10.1016/j.dss.2014.07.003.
- [22] M. M. Fouad, T. F. Gharib, and A.S. Mashat, "Efficient twitter sentiment analysis system with feature selection and classifier ensemble," *International Conference on Advanced Machine Learning Technologies and Applications*, pp. 516-527, 2018, doi :10.1007/978-3-319-74690-6\_51.
- [23] M. J. Alam, T. Chowdhury, S. Hossain, S. Chowdhury, T. Das, "Child tracking and hidden activities observation system through mobile app," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 22, no. 3, pp 1659-1666, 2021, doi: 10.11591/ijeecs.v22.i3.pp1659-1666.
- [24] A. M. Ramadhani and H. S. Goo, "Twitter sentiment analysis using deep learning methods," *2017 7th International Annual Engineering Seminar (InAES)*, Yogyakarta, Indonesia, 2017, pp. 1-4, doi: 10.1109/INAES.2017.8068556.
- [25] N. M. Ali, M. M. Abd El Hamid and A. Youssif, "Sentiment analysis for movies reviews dataset using deep learning models," *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, vol. 9, 2019, doi: 10.5121/ijdkp.2019.9302.




- [26] K. Srividya, and A. M. Sowjanya, "Aspect based sentiment analysis using POS tagging and TFIDF," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 8, no. 6, 2019, doi: 10.35940/ijeat.F7935.088619.
- [27] A. Alayba, "Twitter sentiment analysis on health services in Arabic," Coventry University, 2019. [Online]. Available: <https://pureportal.coventry.ac.uk/en/studentTheses/twitter-sentiment-analysis-on-health-services-in-arabic> (accessed 5/12/2024).
- [28] M. S. Basarslan, and F. Kayaalp, "Sentiment analysis with machine learning methods on social media," *Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 9, no. 3, 2020, doi: 10.14201/ADCAIJ202093515.
- [29] R. Kaushal, S. Saha, P. Bajaj and P. Kumaraguru, "KidsTube: Detection, characterization and analysis of child unsafe content and promoters on YouTube," in *14th Annual Conference on Privacy, Security and Trust (PST)*, 2016, doi: 10.1109/PST.2016.7906950.
- [30] S. Postalcioglu and S. Aktas, "Comparison of neural network models for nostalgic sentiment analysis of YouTube comments," *Hittite Journal of Science & Engineering*, pp.215-221, 2020, doi: 10.17350/HJSE19030000191.

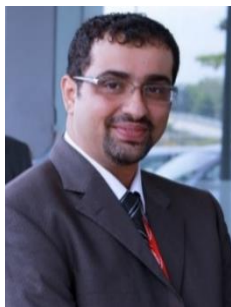
## BIOGRAPHIES OF AUTHORS






**Amal Alrehaili**    received the B.Sc. degree in Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia, in 2017, M.Sc. degree of computer Science from Taibah University in 2021. Her research interests include Natural Language Processing, Image processing, Deep Learning and Machine Learning. She can be contacted at email: [amal.alrehily@gmail.com](mailto:amal.alrehily@gmail.com).



**Abdullah Alsaeeidi**    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 Associate Professor at the Computer Science Department, Taibah University, and Madinah, Saudi Arabia. His research interests include software engineering, software model inference, grammar inference, and machine learning. He can be contacted at email: [aasaeeidi@taibahu.edu.sa](mailto:aasaeeidi@taibahu.edu.sa).



**Wael M.S. Yafooz**    is a professor of Artificial Intelligence in the computer Science Department, Taibah University, Saudi Arabia. He received his bachelor's 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 is an IEEE Senior Member and has obtained the Fellow of the Higher Education Academy (FHEA) recognition. 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. 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) or [wyafooz@taibahu.edu.sa](mailto:wyafooz@taibahu.edu.sa).