Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 211-221. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2860

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

A BIG DATA BASED EMOTION DETECTION FRAMEWORK FOR SOCIAL MEDIA USING APACHE SPARK.

Wafa Saadi*[1], Fatima Zohra Laallam[2] and Messaoud Mezati[3]

[1]Computer Science department, Mohamed Khider University, Biskra, Algeria
[2, 3]Kasdi Merbah University, Ouargla, Algeria

[1]https://orcid.org/0009-0009-7794-1224, [2]https://orcid.org/0000-0003-2875-3576, [3]https://orcid.org/0009-0001-1996-5625

Email: *wafa.saadi@univ-biskra.dz, laallam.fatima_zohra@univ-ouargla.dz, mezati.messaoud@univ-ouargla.dz

ARTICLE INFO

Article History
Received: October 27, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Social Big Data,
Spark Apache,
Emotion Detection,
Machine learning,
YouTube Comments.

ABSTRACT

YouTube is considered as one of the most widely used video-sharing platforms in the world. Users can express their reactions to videos through comments, which often convey emotions that can be automatically identified using computational techniques. Emotion detection on YouTube data presents a challenging task due to the heterogeneity, unstructured nature, and large scale of user generated contents. In this study, we develop an emotion detection framework implemented on Apache Spark, an open-source platform for distributed Big Data processing. The proposed system integrates Machine Learning algorithms with Natural Language Processing (NLP) techniques and leverages Spark's MLlib library to classify emotions expressed in YouTube comments. To efficiently deal with the complexity and noise inherent in largescale multimedia data, several preprocessing and feature extraction steps are introduced. The K-Means clustering algorithm is used after data preparation for the corpus automatic annotation, the resulting labeled Dataset is labeled according to the Ekman emotional model with six basic emotions. The selected classifiers are trained using the resulting labeled Dataset. Experimental results demonstrate that the proposed approach improves both scalability and accuracy, making it suitable for leveraging the emotion detection in social Big Data environments.

I. INTRODUCTION

The exponential growth of data volumes in both industry and research presents significant opportunities alongside substantial computational challenges. The large volume of data mainly named Big Data [1] is a comprehensive term that refers to a wide range of methodologies, technologies, and infrastructures including hardware, software, and real-time data collection and analysis. In contrast to traditional data analytics, which typically processes only a limited subset of data, Big Data analytics [2] involves the processing of the whole data. As data sizes have increasingly exceeded the processing capacity of single machines, there has arisen a critical need for systems capable of scaling computations across multiple nodes. Consequently, this demand has driven the emergence of numerous cluster-based programming models designed to address a wide range of computational workloads. Within the first category, the Apache Spark framework is classified, as it represents a fundamental tool for distributed data processing and large-scale analytics in Big Data environments. Social networking data constitute a significant source of Big Data due to their considerable volume. Social Big Data [3] or as generally called Social media platforms have become one of the most influential channels where users express their thoughts, emotions, opinions, and experiences related to various topics and businesses. Traditionally, such expressions were shared in textual form. Social media has emerged as a prominent platform for online interaction, fundamentally transforming the way of communication, social engagement, and thinking. At the same time, it has facilitated an exponential growth of user-generated content. In recent years, the massive volume of textual data produced on social media platforms has become one of the most representative sources of big data. Therefore, the extraction and analysis of such user-generated information reflect significant importance for understanding and promoting social development.

Understanding user's emotions is crucial for assessing how a product, service, or event reflects public perception. Emotions constitute a fundamental aspect of human personality and play a crucial role in everyday life. Although numerous definitions of emotion exist, they are commonly described as *"complex patterns of reactions encompassing experiential, behavioral, and physiological components."*[4] Emotions are often conflated with feelings or moods; however, they represent distinct psychological phenomena characterized by specific triggers, intensity, and duration. Emotion detection and analysis, also known as emotion recognition, represents a specialized area within affective computing [5] that aims to automatically identify and categorize human emotions based on textual, vocal, physiological, or multimodal inputs. The emotion detection operation is basically funded on an emotion model [6], [7] where we can find multiple model categories [8] in the literature. Emotion detection on social media [9], [10] enables the identification and analysis of emotional patterns within large populations toward specific subjects. In recent years, applications of emotion detection [11], [12] have become increasingly evident in contexts such as political campaigns, movie evaluations, brand perception, videos reviews and various other domains. YouTube represents one of the most widely used video-sharing platforms in recent era, providing users with the opportunity to engage and express their opinions on video content. Viewers can interact with videos primarily through comments, which frequently convey underlying emotional states.

These emotional cues embedded within user comments can be automatically identified and analyzed using computational techniques for emotion detection. In the current study, we will represent a pipeline for emotion detection of YouTube comments for English languages. The framework based on the analysis of the YouTube comments using the MLlib library [13] of the Spark Apache framework [14], a fundamental Big Data analysis framework. Apache Spark has emerged as one of the leading frameworks in big data computing with a robust stack containing several helping tools for better Big Data handling. Based on the offered machine learning algorithms we conducted a system. In the first phase, we collected and prepared the corpus, which is composed of YouTube comments. The preparation process results of a cleaned and embedded Data. The next step is fundamental for the Dataset preparation, where we apply a clustering algorithm for the data automatic annotation. The labelling is based on the Ekman basic emotional model. The last phase is based on the labeled dataset where we selected and trained a machine learning classifiers using the MLlib library. The remainder of this paper is organized as follows. Section 2 presents the literature review and related studies. In Sections 3, we present our methodology, which includes the proposed framework. Section 4 discusses the experimental results. Finally, Section 5 presents the conclusion and perspectives.

II. THEORETICAL REFERENCE

II.1 BIG DATA

II.1.1 Big Data Definition

From a computer science perspective, data is information that can be stored and used by computer program. As a first definition, Big Data is a collection of a large and complex data that traditional data processing tools are incapable to treat. More specific technical definition, Big data is a set of data with high volume, high velocity and/or high variety that requires new forms of processing tools for improved decision making, Knowledge discovery and understanding.

II.1.2 Big Data Characteristics

To maintain a clear conceptual distinction between large data collections and big data, it is crucial to refer to a set of defining characteristics that distinguish truly massive datasets. In scientific and technological contexts, the classification of data as big data does not depend only on its size, but rather on multiple dimensions typically called Vs representing the main characteristics that enable the classification of data as Big Data or otherwise. Initially comes the volume of data, where the term Big Data essentially refers to the big size of data, emphasizing its size as a defining characteristic. In this context, volume represents the total amount or quantity of data generated by one or multiple applications. The data size directly influences the computational and storage requirements, as well as the processing strategies necessary to efficiently manage and analyze such data. The velocity takes the second position; it refers to the speed of data generation. It is a measure of how fast the data is generated and processed. The velocity plays a fundamental role in the processing challenges of Big Data.

In another perspective, Big Data is composed of a variety of data, where data is generated from multiple sources and it consists of various forms and formats. This introduces variety in data and consequently introduces 'complexity'. The variety is due to the availability of a large number of heterogeneous platforms in the industry. This means that the type to which Big Data belongs to is also an important characteristic that needs to be known for appropriate processing of data. This characteristic helps in effective use of data according to their formats, thus maintaining the importance of Big Data. Finally, the veracity is in addition considered as an important characteristic to consider the quality of data captured; it refers to the reliability and integrity of data, encompassing variations in coverage, accuracy, and relevance. Ensuring that datasets are sufficiently accurate is essential, as critical business and analytical decisions often rely on them. While many researchers believed in three and four Vs (i.e. Volume, Velocity, Variety and Veracity) for Big Data characteristics, another concept that needs to be addressed is the discussions that were taking place between researchers on the five Vs [15] and seven Vs [16] of big data.

II.1.3 Big Data Analysis

The primary objectives of Big Data analysis are resumed in two goals; the first goal is to design effective methodologies capable of accurately predicting future observations. The second goal is to extract meaningful insights into the relationships between features and response variables, thereby advancing scientific understanding. Although numerous additional Vs dimensions have been proposed in the literature, researchers consistently acknowledge Volume, Velocity, and Variety as the three fundamental characteristics of Big Data. These core attributes were initially identified in early foundational studies and continue to be emphasized as the primary dimensions

defining Big Data. Big Data are characterized by both high dimensionality and large sample sizes, which together introduce several distinctive challenges. Where the volume of data often leads to noise accumulation, spurious correlations, and incidental homogeneity, complicating accurate inference. When combined with variety of data, it further results in high computational costs and algorithmic instability, affecting model efficiency and scalability. Moreover, Big Data are characterized by the variety and velocity of data as presented in the figure 1, collected at different time points and through various technologies. These characteristics introduce experimental variations and statistical biases, thereby necessitating the development of adaptive and robust analytical methods capable of handling such complexity.

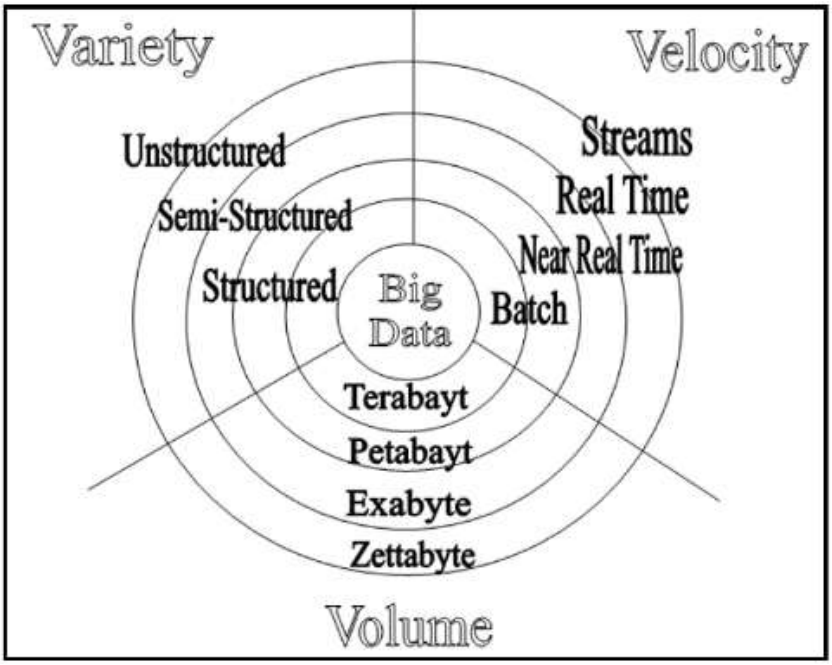

Figure 1: The three Big Data Vs.
Source: [17].

In conclusion, the different Vs characterizing Big Data need tools and techniques for data storage, applications, managements, meaning and analysis in different application domain like education and learning analytics, healthcare and biomedical research, finance and banking, social media analysis and various others interested in derive knowledge from amounts of data. In particular, data continues to grow exponentially; new technologies are evolving accordingly. Frameworks such as Hadoop, Apache Spark, and Apache Storm have gained significant popularity due to their ability to handle large-scale data processing with high efficiency. These systems are continuously advancing, leading to the development of specialized libraries such as Spark's MLlib that facilitate the integration of Machine Learning techniques within Big Data based environment. In the next section, we will introduce one of the most used frameworks designed for the process of data in the Big Data context.

II.2 APACHE SPARK

The exponential growth of data volumes in both industrial and research domains presents significant opportunities, yet also introduces significant computational challenges. As data growth have surpassed the processing capabilities of individual machines, it has become necessary to develop systems capable of distributing computations across multiple nodes. Consequently, a rise of cluster-based programming paradigms has emerged, each designed to address a variety of computational workloads. Among the most popular Big Data frameworks in recent years are: Apache Spark for distributed processing of data, Apache Kafka for real-time streaming, Hadoop HDFS is for distributed storage, Databricks is a unified Big Data and AI platform and Snowflake and Google BigQuery are top cloud analytics solutions. Initially, these paradigms were relatively specialized, with new models being introduced to meet the specific requirements of emerging application domains.

Apache Spark [18] is an open-source distributed computing framework that is designed for large scale data processing, it is considered as one of the most prominent and widely adopted cluster computing systems. Supported by a highly active open source community, Spark remains among the most dynamic projects within the Apache ecosystem figure 2. It delivers high performance for both batch and streaming workloads. This efficiency is achieved through its advanced Directed Acyclic Graph (DAG) scheduler, query optimizer, and physical execution engine. By storing data in the memory of worker nodes, Spark significantly outperforms Hadoop in scenarios that require multiple iterative operations. Apache Spark offers a more efficient and user friendly analytics environment compared to Hadoop MapReduce, enabling programs to execute up to 100 times faster in memory and up to 10 times faster on disk. Moreover, Spark provides extensive support for multiple programming languages, including Java, Python, and R, and can operate seamlessly across various environments such as Hadoop, Mesos, standalone clusters, and cloud platforms.

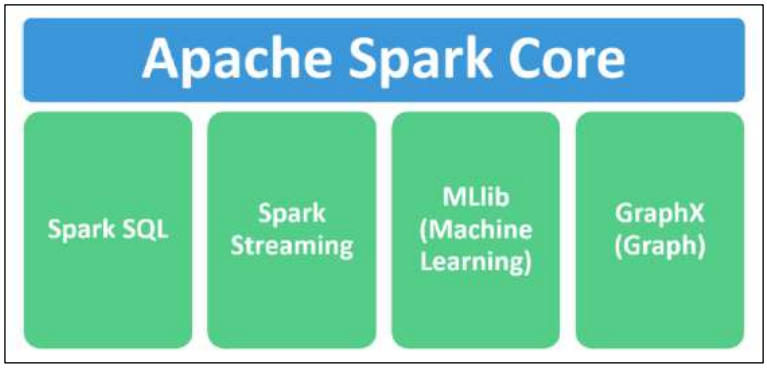

Figure 2: Apache Spark Ecosystem.
Source: Authors, (2026).

II.2.1 MLlib

MLlib [13] is a scalable, user-friendly, and comprehensive machine learning library developed as part of the Apache Spark ecosystem, offering both high performance and ease of use in distributed computing environments. It comprises a comprehensive set of machine learning algorithms, it is designed to process and analyze large-scale data in parallel. However, executing MLlib algorithms [19], [20]within Spark can present considerable technical complexity and operational effort challenges that become progressively more complex because of the intensive growth of dataset sizes.

II.2.2 Core Features of MLlib

MLlib provides a robust, flexible and comprehensive library for large-scale machine learning. Its core features include a comprehensive collection of algorithms for classification, regression, clustering, and collaborative filtering, enabling efficient analysis of massive datasets in distributed environments. Designed with scalability and usability in mind, MLlib simplifies the construction and optimization of machine learning workflows, offering a streamlined experience comparable to that of scikit-learn. Beyond algorithmic implementation, it encompasses advanced functionalities for feature extraction, selection, and scaling, as well as built in tools for model evaluation using standard performance indicators such as accuracy and precision. The library supports two main programming interfaces: the RDD-based API, which provides fine-grained control over distributed computations, and the DataFrame-based API, which enhances ease of use and performance optimization for high-level analytical tasks.

The next section introduces one of the typical fields of application of MLlib in the social big data analytics domain witch is the emotion detection in social media, more particular emotion detection from text data derived from YouTube comments.

II.3 EMOTION DETECTION FROM YOUTUBE COMMENTS

Emotions constitute an essential aspect of human life, fundamentally influencing the way individuals perceive and interpret the world. Over the past three decades, numerous approaches have been developed to advance emotion analysis ranging from manual techniques, such as psychologist-designed questionnaires, to computer-based methods. Nowadays, computer assisted emotion recognition has found widespread applications. Humans express their emotions through various modalities, including written language, vocal tone, facial expressions, physiological responses, body gestures, postures, and other physiological signals. For example, emotion detection from physiological signals [21-24] is employed in the development of smart homes and intelligent work environments. In addition, facial recognition techniques [25-27] are increasingly used in consumer services, education, and security-related domains, among others. To systematically categorize and interpret the emotional expressions, several theoretical emotion models have been proposed [28-30]. Selecting an appropriate emotion model is essential for accurately recognizing and interpreting emotions across different modalities.

Such a model should define a set of permissible emotional categories that are relevant and suitable for the specific application context or scenario. With the rapid development and evolution of social media platforms, an increasing number of users now express their emotions and opinions online in response to various events or topics, ranging from feelings of affection and agreement to dislike or sadness. Consequently, research focused on developing effective computational techniques to analyze such user-generated content and enable emotion-aware systems has garnered substantial attention from both the academic and industrial communities in recent years. In this context, analyzing the YouTube contents intended to support emotion detection objectives, has progressively evolved in recent years. Comments on YouTube videos related to a given topic often contain spontaneous emotional expressions, offering valuable insights into public sentiment.

II.4 RELATED WORKS

Emotion detection in the context of YouTube has attracted increasing scholarly attention in recent years due to the platform's massive user base and the richness of affective information embedded in user-generated content. For instance, The research presented in [31] explores the relationship between color and emotion by analyzing the rich spectrum of colors present in YouTube videos to uncover the emotional nuances they communicate. This study explores how colors influence emotional responses in movies and other visual media. By integrating classical color theory, psychological research, and modern media analysis, it seeks to uncover new ways of predicting and interpreting emotional effects in digital content. In a different perspective, recent research on YouTube video classification has shifted from purely topical categorization toward emotion based analysis, for example, the research performed in [32] have introduced models that classify videos into six basic emotions, happiness, anger, disgust, fear, sadness, and surprise, using supervised and

unsupervised learning methods, highlighting the importance of emotional information in multimedia content understanding. An other example of research in this area elaborated in [33] has examined how different modalities, text, audio, and visual features like color, interact to influence viewers emotional reactions. By analyzing textual components, including video titles, descriptions, and transcripts, alongside audiovisual cues to identify the elements that most strongly evoke emotions across various genres.

The research indicates that textual information, particularly the combination of titles, descriptions, and transcriptions, plays a central role in influencing audience emotions, highlighting the significance of linguistic context in multimodal emotion analysis. In the context of emotion analysis on YouTube comments, an extensive research has been conducted in the literature, employing a wide range of techniques and approaches to detect emotions embedded within the textual content of user comments. For example, in the study performed in [34], the authors conducted experiments on emotion classification on Indonesian YouTube comments. A corpus containing 8,115 YouTube comments is collected and manually labelled using the Ekman basic emotional model, with one added neutral label. From the experiments, they conclude that using word embedding on emotion classification task on YouTube comment can increase the accuracy of the classification and that the best method to use word embedding for the classification is by using Convolutional Neural Network algorithm. An additional work realized in [35] where the authors presented a comprehensive set of techniques to identify sentiment and extracted emotions from Bangla texts by collecting a dataset of comments from various YouTube videos.

They built deep learning based models to classify the Bangla sentences with a three class (positive, negative, neutral) and a five class (strongly positive, positive, neutral, negative, strongly negative) sentiment labels. They also built models to extract the emotion of the Bangla sentence as any one of the six basic emotions (anger, disgust, fear, joy, sadness and surprise). Finely, they evaluated the performance of their model using a new dataset of Bangla, English and Romanized Bangla comments from different types of YouTube videos. Some research has employed machine-learning technique, precisely, the supervised technique to detect emotions in YouTube comments, the work presented in [36] analyzed users expressed emotions, the study conducts a sentiment analysis based on textual comments. The authors evaluated the performance of a Long Short-Term Memory (LSTM) model on two distinct datasets, the IMDB movie reviews dataset and a YouTube video comments dataset available on GitHub, in order to examine the generalization capability and effectiveness of a single model across different but semantically related domains. The proposed LSTM model achieves good testing accuracies on the datasets, demonstrating competitive performance compared to existing sentiment analysis approaches.

The stady performed in [37] investigated the role of psychological and affective features in improving automated cyberbullying detection. It integrated personality traits, derived from the personality models, and emotional categories, based on Ekman's model extracted from textual YouTube comments. The authors evaluated whether incorporating these higher-level behavioral cues could enhance the classification effectiveness of ensemble machine learning models (Random Forest and AdaBoost), demonstrating that the fusion of personality and emotion based features gives rise to superior performance compared to traditional linguistic feature sets. Some other research has employed unsupervised approaches to identify emotional patterns in multilingual datasets, including English and Arabic comments related to social and political conflicts, highlighting the cross-linguistic and contextual complexity of emotional expression online. For example, the study performed in [38] where the main objectives were the automatic annotation of a multilingual corpus with emotional label based on the Ekman emotional model, with the investigation of the psychological impact of crises on individuals, by analyzing the emotional content embedded in YouTube comments associated with war-related videos.

The study focuses on the clustering algorithm K-means, which is an unsupervised model, used for emotion detection from YouTube comments related to the Gaza war, using data in both English and Arabic. It explores the role of Emoji's in enhancing emotional interpretation. Generally, the findings confirm the effectiveness of the clustering techniques in detecting emotional patterns from unstructured and diverse YouTube data. The proposed framework presented a scalable, language-aware solution for unsupervised emotion detection, making it particularly suitable for multilingual and sensitive contexts. Another initiative using the unsupervised techniques is presented in [39] where the authors presented there system architecture divided into two phases : at first they collected a corpus of YouTube comments from various videos, Emotion labeling is not required in this phase. In the second phase, an unsupervised machine learning approach was adopted to group or classify comments according to their emotional tone without relying on predefined emotion labels. The unsupervised machine learning algorithm was based on the work presented in [40], with modifications to improve accuracy, and utilizing YouTube comments as a rich resource for publicly available text.

In another orthogonal context, using the Big Data frameworks like Apache Spark has demonstrated significant improvement. The research performed in [41] addresses the challenge of analyzing massive volumes of audience comments on YouTube videos, where the main goal is to use machine learning algorithms in a distributed environment mainly Apache Spark to automatically analyze the large dataset of viewer comments, determine audience sentiment, and help content creators improve their strategies. The authors introduced an Improved Novel Ensemble Method (INEM), ensemble-based machine learning model that combines multiple classifiers to improve sentiment prediction. Using Spark significantly improved scalability and speed for processing and analyzing big datasets of YouTube comments. Using Twitter data the researchers in [42] address the challenge of automatically detecting emotional states expressed by users in Arabic language tweets, particularly within the context of the COVID-19 pandemic. The authors affirm that beyond simple positive/negative sentiment, fine-grained emotion classification like joy, fear, anger, sadness, disgust, surprise plays a crucial role in understanding human behavior in crises.

The aim of the research is to develop a real-time online tool that handle streaming Arabic tweets and predicts one of six basic emotions, using the Spark Apache tools to collect, pre-process, feature-extract, and apply the trained model in real time to each tweet. In the same context, the system described in [43] highlights the power of the Apache Spark platform in order to perform sentiment analysis on large volumes of data from Twitter. The motivation was that, given the volume and velocity of Twitter data streams, traditional single machine or batch approaches are insufficient. The authors built a pipeline that can scale, process the data effectively, classify tweets by sentiment, and thereby support applications such as public opinion monitoring, marketing analytics, or social behavior analysis. Using the Spark platform, machine-learning algorithms mainly: Naïve Bayes, Logistic Regression, Support Vector Machine, Decision Tree that were applied via the MLlib library, enabled efficient handling of larger tweet volumes and faster processing compared to non-distributed methods. The Table 1 summarizes the selected researches related to our study.

Table 1: Summary of prior research.

Ref.	Research Focus	Data Source	Methodology / Model	Emotion Model	Key Findings
[31]	Relationship between color and emotion in media	YouTube videos	Color spectrum analysis integrating color theory, psychology, and media analysis	Plutchik emotional model (8 emotions)	Colors strongly influence emotional response; visual features contribute to emotional interpretation.
[32]	Emotion-based video classification	YouTube videos	Supervised & unsupervised learning	Ekman emotional model (Six basic emotions)	Models effectively classify videos into six emotional categories, emphasizing the role of emotional context in multimedia understanding.
[33]	Multimodal emotion analysis (text, audio, visual)	YouTube videos	Feature fusion: textual (titles, descriptions, transcripts) + audiovisual	Russell and Mehrabian model	Textual information, especially combined metadata, plays a dominant role in predicting audience emotions from video.
[34]	Emotion classification on Indonesian YouTube comments	YouTube comments	Word embeddings + CNN	Ekman emotional model + Neutral	Word embeddings improve accuracy; CNN performs best for emotion classification.
[35]	Emotion and sentiment detection in Bangla YouTube comments	Bangla, English, Romanized Bangla YouTube comments	Deep-learning models	Sentiment (3-class, 5-class)&Ekman emotions	Created multilingual datasets; models perform well for both sentiment and emotion tasks.
[36]	Cross-domain emotion analysis (YouTube & IMDB)	YouTube comments & IMDB reviews	LSTM (supervised)	Sentiment analysis	LSTM generalizes well across domains; competitive performance on multiple datasets.
[37]	Cyberbullying detection enhanced by emotion & personality traits	YouTube comments	Ensemble ML (Random Forest, AdaBoost)	Ekman + Personality traits	Fusion of personality and emotion features improves detection accuracy beyond linguistic baselines.
[38]	Multilingual unsupervised emotion detection (English & Arabic)	War-related YouTube comments	K-Means clustering (unsupervised)	Ekman emotional Model	Effective at uncovering emotional patterns; emojis enhance interpretation; suited to multilingual contexts.
[39]	Unsupervised emotion detection for automatic datasets annotation	YouTube comments	Unsupervised algorithm based on [40]	Ekman emotional Model	Automatically clusters comments by emotional tone; works without predefined labeled dataset.
[41]	Distributed emotion analysis using Big Data frameworks	Large-scale YouTube comments	Apache Spark + Improved Novel Ensemble Method (INEM)	Sentiment analysis	Spark ensures scalability & speed; ensemble model boosts predictive accuracy on massive datasets.
[42]	Real-time Arabic emotion detection during COVID-19	Twitter data	Apache Spark pipeline	Ekman emotional Model	Real-time detection of six emotions; Spark supports live streaming and online analysis.
[43]	Scalable sentiment analysis on social-media data streams	Twitter data	Spark MLlib (Naïve Bayes, LR, SVM, DT)	Sentiment analysis	Distributed Spark framework enables fast, large-scale sentiment classification.

Source: Authors, (2026).

III. MATERIALS AND METHODS

III.1 DATASET CREATION

The dataset is created by the extraction of comments from two videos using YouTube API version 3.0. We manually select these videos from the CNN and BBC news channels in English language based on their popularity (number of views and number of comments) dated respectively 2025 and 2021. We limit the number of comments for each video up to **4490**, which results a total number of **8980** comments written in English. Table 2 provides a description of our dataset features.

Table 2: Dataset description.

Dataset features	Number of comments
Total number of comments	8980
Comments with Emojis	932
Comments without Emojis	8048

Source: Authors, (2026).

The dataset annotation is based on the Ekman emotional model where six basic emotion, which defines six fundamental emotion categories: happiness, sadness, fear, disgust, anger, and surprise. The annotation process is mainly based on the pipeline described by the work defined in [38] where we used the K-means model which is an unsupervised machine learning technique. Based on the same strategy, we performed a preprocessing pipeline to prepare the data for the labeling process.

III.2 DATASET PREPARATION

The Dataset preparation process begins after the data collection, which results a corpus of unlabeled YouTube comments described in the section above. The resulting Dataset is annotated by the K-means clustering model based on the Ekman's basic emotional model. The figure 3 illustrates the different steps of the dataset annotation process. Data preprocessing enhances text quality and consistency through cleaning and normalization. It ensures cleaner and more consistent textual input for emotion detection. It removes noise (URLs, mentions, symbols), normalizes text (repetitions, hashtags, whitespace, case), and eliminates stop words to improve model accuracy. Numerical values are retained, as they may hold meaningful contextual or emotional significance. After cleaning, text tokenization is performed using BERT tokenizer, which splits text into sub word units and converts them into token IDs for model input. This enables BERT's bidirectional encoding to capture full contextual meaning which is crucial for emotion detection. The high dimensional output from BERT increased computational complexity and noise during clustering. To avoid this issue, Principal Component Analysis (PCA) was applied to reduce dimensionality while preserving key variance and data structure, resulting in a more efficient and effective clustering process.

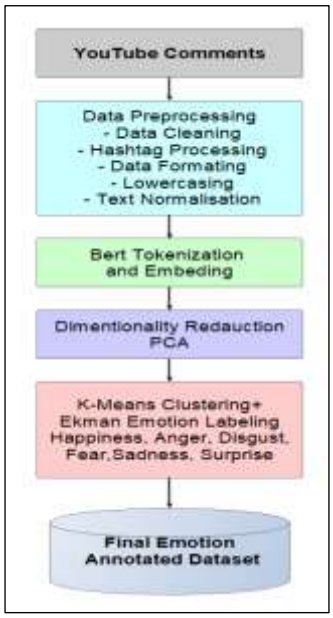

Figure 3: Data annotation Pipeline.
Source: Authors, (2026).

III.3 CLUSTERING TECHNIQUE

Unsupervised learning techniques provide an effective way for textual data annotation with minimal human intervention, motivating the use of clustering for automatic text annotation. The annotation of the YouTube comments Dataset is assured by using a clustering technique basically K-means, which efficiently handles large datasets and balanced clusters, to assign emotional labels to YouTube comments. This approach produces a labeled dataset that forms the basis for training supervised models for accurate emotion detection. In the next section we will present the selected collection of machine learning techniques where the main objective is using the resulting labeled dataset after the clustering step in the training for preparing the model to the emotion prediction and recognition.

III.4 MACHINE LEARNING TECHNIQUES

In this work, we utilized three classification algorithms in order to implement the emotion detection system. At this level, the selected classifiers are chosen from the machine learning models defined in the MLlib [13] Spark's library. We selected a collection composed of classic models where the objective is to compare their accuracy and decide the best model in the context of emotion

recognition from amount volume of text specifically in the context of YouTube comments analysis. The selected model are basically the Naïve Bayes, Logistic Regression and Random Forest classifier.

III.4.1 Naive Bayes

The Naïve Bayes classifier is a probabilistic multiclass classification algorithm grounded in Bayes' theorem. In this model, each data instance is represented as a feature vector, under the assumption that all features are conditionally independent given the class label. The algorithm is computationally efficient, requiring only a single pass through the training data to estimate parameters. During training, the conditional probability distribution of each feature with respect to each class is computed. Subsequently, Bayes' theorem is applied to determine the most probable class label for unseen instances.

III.4.2 Logistic Regression

The Logistic Regression is a statistical classification model in which the dependent variable can assume one of several discrete values. It employs the logistic function to model the relationship between the input features and the probability of belonging to a specific class. While the model is primarily designed for binary classification, it can be extended to handle multiclass problems using different strategies.

III.4.3 The Random Forest Classifier

The Random Forest algorithm is an ensemble learning technique based on the bagging principle. It addresses the main limitation of a single decision tree by constructing multiple decision trees on randomly selected subsets of the dataset. The final prediction is then obtained by taking the majority vote of the individual trees. The Random Forest Classifier in Apache Spark's MLlib is an ensemble-learning model based on the Random Forest algorithm, where his primary role is tasks classification. It is based on the construction of a multitude of decision trees during training where it returns as result the class with the highest vote among the individual trees. In the next section, we will present the different obtained results with a discussion of the most relevant among them.

III.5 EVALUATION METRICS

The performance of the used machine learning algorithms is measured using a set of evaluation metrics, in the first range the accuracy is used; it is the most widely used. For many years, accuracy served as the basic criterion for comparing machine learning models effectiveness. However, it is not sufficient alone to make an accurate decision of the model performance. Accordingly, we used the Precision, the Recall and the F-score for more precise performance measurement of the used models. Accuracy measures the overall correctness of predictions, while Precision and Recall evaluate the model's ability to correctly identify relevant classes. The F1-score provides a harmonic mean between Precision and Recall, providing a more robust performance indicator in imbalanced datasets. It ranges from zero to one where a higher F1-score indicates better model performance.

IV. RESULTS AND DISCUSSIONS

In this section, we present and analyze the experimental results obtained, along with an evaluation of the effectiveness of the tree used algorithms (Logistic Regression, Naïve Bayes, and Random Forest) from the MLlib Spark's library applied to the resulting labeled dataset. In addition, we discusses the performance of the proposed models based on various evaluation metrics. The goal is to evaluate the effectiveness and overview capability of each algorithm on the selected dataset by analyzing and comparing the obtained results. To evaluate model performance, we used standard metrics notably: Accuracy, Precision, Recall, and F1-score. Table 3 presents the performance of the three used classification models based on the four standard evaluation metrics presented above. As illustrated, the Logistic Regression and Naïve Bayes models achieved the highest accuracy values, demonstrating superior performance to Random Forest Classifier by a considerable margin.

Table 3: Performance results for the three algorithms.

Models	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.9289805269186713	0.9293267627553385	0.9289805269186712	0.9282860057563744
Naïve Bayes	0.8178694158075601	0.8370928837933002	0.8178694158075601	0.8217422717081797
Random Forest Classifier	0.5664375715922108	0.49976124982922965	0.5664375715922108	0.41087645030132036

Source: Authors, (2026).

For instance, the Logistic Regression obtained an accuracy of 92.3%, with a precision of 91.8% and an F1-score of 92.0%, indicating consistent performance across classes. Overall, Logistic Regression achieved the best performance across all metrics, with an accuracy of approximately 0.93 and an F1-score of 0.93, indicating consistent and reliable classification results. Naïve Bayes followed, obtaining an accuracy of 0.82 and an F1-score of 0.82, which suggests satisfactory performance but slightly lower precision and recall compared to Logistic Regression. In contrast, the Random Forest Classifier recorded the lowest performance, with an accuracy of 0.57 and an F1-score of 0.41. This poor performance may be due to overfitting, suboptimal parameter tuning, or an insufficient number of estimators, which limited its generalization ability on the dataset. In summary, the comparative results indicate that Logistic Regression provides the most effective balance between accuracy, precision, and recall for this emotion detection task, outperforming both Naïve Bayes and Random Forest. The Figure 4 illustrates the visualization via a bar chart comparing the performance metrics of the three models. We can clearly notice that Logistic Regression outperforms the other models across all metrics (Accuracy, Precision, Recall, and F1-Score), while the Random Forest Classifier shows the weakest performance overall.

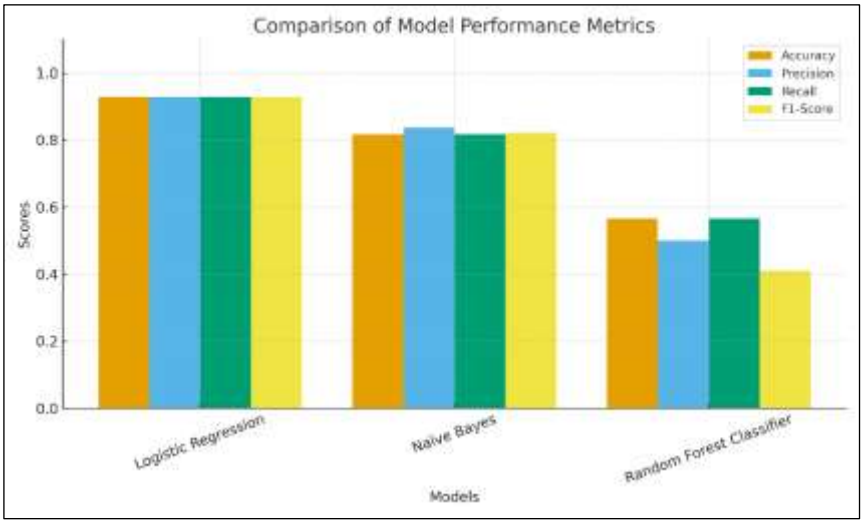

Figure 4: Performance comparison of the three models.
Source: Authors, (2026).

V. CONCLUSIONS

In this study, we presented our framework designed for the emotion detection from YouTube comments, leveraging the machine learning capabilities of the Spark's Apache MLlib library. We proposed a scalable emotion detection framework for analyzing YouTube comments using Apache Spark large-scale abilities. The integration of Machine Learning algorithms with Natural Language Processing techniques enabled efficient processing and classification of emotions across large dataset. By employing K-Means clustering for automatic corpus annotation and supervised classifiers trained on the labeled dataset, the system demonstrated improved accuracy and scalability compared to traditional approaches. The experimental evaluation demonstrated that the framework could efficiently process large volumes of data while maintaining high performance, where the Logistic Regression surpassed 92% in both metrics, accuracy and F1-Score, making it a promising solution for real-world social media analytics. Moreover, the study underlines the importance of scalable architectures in Big Data driven emotion analysis. Future research will focus on the integration of deep learning architectures and multimodal data (text, audio, and video) to improve the precision and expressiveness of emotion detection systems in complex online environments. In the same context, the work can be adopted for steaming data using the Spark Steaming library for Emotion detection tasks in large steaming social data.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Wafa Saadi, Fatima Zohra Laallam and Messaoud Mezati.
Methodology: Wafa Saadi.
Investigation: Wafa Saadi.
Discussion of results: Wafa Saadi, Fatima Zohra Laallam and Messaoud Mezati.
Writing – Original Draft: Wafa Saadi**.**
Writing – Review and Editing: Wafa Saadi.
Resources: Wafa Saadi.
Supervision: Fatima Zohra Laallam and Messaoud Mezati.
Approval of the final text: Wafa Saadi, Fatima Zohra Laallam and Messaoud Mezati.

VIII. REFERENCES

[1] A. Badshah, A. Daud, R. Alharbey, A. Banjar, A. Bukhari, and B. Alshemaimri, 'Big data applications: overview, challenges and future', Artif. Intell. Rev., vol. 57, no. 11, p. 290, Sep. 2024, doi: 10.1007/s10462-024-10938-5.

[2] R. Kamal and P. Saxena, Big Data Analytics. Mcgraw Hill Education, 2019.

[3] M. Ianni, E. Masciari, and G. Sperlí, 'A survey of Big Data dimensions vs Social Networks analysis', J. Intell. Inf. Syst., vol. 57, no. 1, pp. 73–100, Aug. 2021, doi: 10.1007/s10844-020-00629-2.

[4] W. G. Parrott, Emotions in Social Psychology: Essential Readings. in Key readings in social psychology. Psychology Press, 2001. [Online]. Available: https://books.google.dz/books?id=FNBnKLvFdM4C

[5] R. W. Picard, Affective Computing. The MIT Press, 1997. doi: 10.7551/mitpress/1140.001.0001.

[6] P. Ekman, 'Basic Emotions', in Handbook of Cognition and Emotion, 1st ed., T. Dalgleish and M. J. Power, Eds., Wiley, 1999, pp. 45–60. doi: 10.1002/0470013494.ch3.

[7] R. Plutchik, 'A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION', in Theories of Emotion, Elsevier, 1980, pp. 3–33. doi: 10.1016/B978-0-12-558701-3.50007-7.

[8] R. Plutchik and H. Kellerman, Theories of emotion. in Emotion, theory, research, and experience, no. v. 1. New York: Academic Press, 1980.

[9] H. Zhang, ‘A Study of Human Emotion Analysis Based on Social Media’, in Proceedings of the 2023 2nd International Conference on Social Sciences and Humanities and Arts (SSHA 2023), vol. 752, M. F. B. Sedon, I. A. Khan, M. C. BiRkök, and K. Chan, Eds., in Advances in Social Science, Education and Humanities Research, vol. 752. , Paris: Atlantis Press SARL, 2023, pp. 174–180. doi: 10.2991/978-2-38476-062-6_23.

[10] K. A. S, M. S, and R. R. L, ‘Emotional Analysis on Social Media Platform’, in 2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), Chennai, India: IEEE, Mar. 2025, pp. 1–6. doi: 10.1109/ICDSAAI65575.2025.11011895.

[11] J. M. Garcia-Garcia, V. M. R. Penichet, and M. D. Lozano, ‘Emotion detection: a technology review’, in Proceedings of the XVIII International Conference on Human Computer Interaction, Cancun Mexico: ACM, Sep. 2017, pp. 1–8. doi: 10.1145/3123818.3123852.

[12] A. Saxena, A. Khanna, and D. Gupta, ‘Emotion Recognition and Detection Methods: A Comprehensive Survey’, J. Artif. Intell. Syst., vol. 2, no. 1, pp. 53–79, 2020, doi: 10.33969/AIS.2020.21005.

[13] ‘MLlib: RDD-based API - Spark 4.0.1 Documentation’. Accessed: Oct. 20, 2025. [Online]. Available: https://spark.apache.org/docs/latest/mllib-guide.html

[14] ‘Apache SparkTM - Unified Engine for large-scale data analytics’. Accessed: Oct. 20, 2025. [Online]. Available: https://spark.apache.org/

[15] T. L. Nguyen, ‘A Framework for Five Big V’s of Big Data and Organizational Culture in Firms’, 2018 IEEE Int. Conf. Big Data Big Data, pp. 5411–5413, 2018.

[16] M. A. Khan, M. F. Uddin, and N. Gupta, ‘Seven V’s of Big Data understanding Big Data to extract value’, in Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education, 2014, pp. 1–5. doi: 10.1109/ASEEZone1.2014.6820689.

[17] S. Sagiroglu and D. Sinanc, ‘Big data: A review’, in 2013 International Conference on Collaboration Technologies and Systems (CTS), 2013, pp. 42–47. doi: 10.1109/CTS.2013.6567202.

[18] M. Zaharia et al., ‘Apache Spark: a unified engine for big data processing’, Commun. ACM, vol. 59, no. 11, pp. 56–65, Oct. 2016, doi: 10.1145/2934664.

[19] X. Meng et al., ‘MLlib: Machine Learning in Apache Spark’, J. Mach. Learn. Res., vol. 17, no. 34, pp. 1–7, 2016.

[20] L. Theodorakopoulos, A. Karras, and G. A. Krimpas, ‘Optimizing Apache Spark MLlib: Predictive Performance of Large-Scale Models for Big Data Analytics’, Algorithms, vol. 18, no. 2, p. 74, Feb. 2025, doi: 10.3390/a18020074.

[21] X. Li et al., ‘EEG Based Emotion Recognition: A Tutorial and Review’, ACM Comput. Surv., vol. 55, no. 4, pp. 1–57, Apr. 2023, doi: 10.1145/3524499.

[22] L. Wang, J. Hao, and T. H. Zhou, ‘ECG Multi-Emotion Recognition Based on Heart Rate Variability Signal Features Mining’, Sensors, vol. 23, no. 20, p. 8636, Oct. 2023, doi: 10.3390/s23208636.

[23] J. Perdiz, G. Pires, and U. J. Nunes, ‘Emotional state detection based on EMG and EOG biosignals: A short survey’, in 2017 IEEE 5th Portuguese Meeting on Bioengineering (ENBENG), Coimbra, Portugal: IEEE, 2017, pp. 1–4. doi: 10.1109/ENBENG.2017.7889451.

[24] Aloy Anuja Mary G, ‘Emotion Detection Through Electrocardiogram Signal Classification in an IOT Environment with Deep Neural Networks’, J. Electr. Syst., vol. 20, no. 3, pp. 1620–1630, May 2024, doi: 10.52783/jes.3657.

[25] Prof. K. Sarvakar, ‘A Survey Of Face Emotion Recognition Using Deep Learning Methods’, Jun. 03, 2025, In Review. doi: 10.21203/rs.3.rs-6794812/v1.

[26] H. A. Shehu, W. N. Browne, and H. Eisenbarth, ‘Emotion categorization from facial expressions: A review of datasets, methods, and research directions’, Neurocomputing, vol. 624, p. 129367, Apr. 2025, doi: 10.1016/j.neucom.2025.129367.

[27] H. Liu, ‘Emotion Detection through Body Gesture and Face’, 2024, arXiv. doi: 10.48550/ARXIV.2407.09913.

[28] K. J. Lakshmi Bai, A. Harshita, T. Gaddam, S. Mirtipati, and D. Peyyala, ‘A Comprehensive survey on Emotion Recognition’, in 2024 IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG), Indore, India: IEEE, Dec. 2024, pp. 1–6. doi: 10.1109/ICTBIG64922.2024.10911870.

[29] A. R. Murthy and K. M. Anil Kumar, ‘A Review of Different Approaches for Detecting Emotion from Text’, IOP Conf. Ser. Mater. Sci. Eng., vol. 1110, no. 1, p. 012009, Mar. 2021, doi: 10.1088/1757-899X/1110/1/012009.

[30] R. A. García-Hernández et al., ‘A Systematic Literature Review of Modalities, Trends, and Limitations in Emotion Recognition, Affective Computing, and Sentiment Analysis’, Appl. Sci., vol. 14, no. 16, p. 7165, Aug. 2024, doi: 10.3390/app14167165.

[31] M. C. Cakmak, M. Shaik, and N. Agarwal, ‘Emotion Assessment of YouTube Videos using Color Theory’, in Proceedings of the 2024 9th International Conference on Multimedia and Image Processing, Osaka Japan: ACM, Apr. 2024, pp. 6–14. doi: 10.1145/3665026.3665028.

[32] Y.-L. Chen, C.-L. Chang, and C.-S. Yeh, ‘Emotion classification of YouTube videos’, Decis. Support Syst., vol. 101, pp. 40–50, Sep. 2017, doi: 10.1016/j.dss.2017.05.014.

[33] N. Yousefi, M. C. Cakmak, and N. Agarwal, ‘Examining Multimodel Emotion Assessment and Resonance with Audience on YouTube’, in Proceedings of the 2024 9th International Conference on Multimedia and Image Processing, Osaka Japan: ACM, Apr. 2024, pp. 85–93. doi: 10.1145/3665026.3665039.

[34] J. Savigny and A. Purwarianti, ‘Emotion classification on youtube comments using word embedding’, in 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA), Denpasar: IEEE, Aug. 2017, pp. 1–5. doi: 10.1109/ICAICTA.2017.8090986.

[35] N. Irtiza Tripto and M. Eunus Ali, ‘Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments’, in 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), Sylhet: IEEE, Sep. 2018, pp. 1–6. doi: 10.1109/ICBSLP.2018.8554875.

[36] P. Sharma, S. Gupta, P. Gangwar, P. Kumar, and N. Saxena, ‘Decoding emotions: LSTM-based sentiment analysis for movie reviews and YouTube comments’, presented at the 4TH INTERNATIONAL CONFERENCE ON INNOVATION IN IOT, ROBOTICS AND AUTOMATION (IIRA 4.0), Moradabad, India, 2025, p. 020044. doi: 10.1063/5.0246918.

[37] V. Balakrishnan and S. K. Ng, 'Personality and emotion based cyberbullying detection on YouTube using ensemble classifiers', Behav. Inf. Technol., vol. 42, no. 13, pp. 2296–2307, Oct. 2023, doi: 10.1080/0144929X.2022.2116599.

[38] W. Saadi, F. Z. Laallam, M. Mezati, C. R. HALIMI, and E. H. BRIKI, 'Emotion Detection in Conflict Discourse: An Unsupervised Approach to Multilingual YouTube Comments', J. Inf. Syst. Eng. Manag., vol. 10, no. 58, pp. 7–15, 2025.

[39] D. Yasmina, M. Hajar, and A. M. Hassan, 'Using YouTube Comments for Text-based Emotion Recognition', Procedia Comput. Sci., vol. 83, pp. 292–299, 2016, doi: 10.1016/j.procs.2016.04.128.

[40] A. Agrawal and A. An, 'Unsupervised Emotion Detection from Text Using Semantic and Syntactic Relations', in 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Macau, China: IEEE, Dec. 2012, pp. 346–353. doi: 10.1109/WI-IAT.2012.170.

[41] K. Subha and N. Bharathi, 'Spark-based Big Data Sentiment Analysis of Social Media Comments', in 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), Theni, India: IEEE, Nov. 2023, pp. 164–169. doi: 10.1109/ICSCNA58489.2023.10370518.

[42] N. Abdelhady, I. E. Elsemman, and T. H. A. Soliman, 'A real-time predicting online tool for detection of people's emotions from Arabic tweets based on big data platforms', J. Big Data, vol. 11, no. 1, p. 171, Nov. 2024, doi: 10.1186/s40537-024-01035-z.

[43] A. Baltas, A. Kanavos, and A. K. Tsakalidis, 'An Apache Spark Implementation for Sentiment Analysis on Twitter Data', in Algorithmic Aspects of Cloud Computing, T. Sellis and K. Oikonomou, Eds., Cham: Springer International Publishing, 2017, pp. 15–25.