ISSN 1840-4855
e-ISSN 2233-0046

Original scientific article
http://dx.doi.org/10.70102/afts.2025.1834.605

APPLICATION OF HYBRID & NOVEL DEEP LEARNING APPROACHES FOR MULTIMODAL SENTIMENT FUSION IN IMAGES & AUDIO ANALYSIS

Jayaprakash Vattikundala[1*], M. Siva Ganga Prasad[2]

[1]Research Scholar, Department of ECM, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India. e-mail: chiranjeevijp@gmail.com, orcid: https://orcid.org/0009-0007-9194-7564*
[2]Professor & Coordinator (FED), Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India.
e-mail: msivagangaprasad@kluniversity.in, orcid: https://orcid.org/0000-0003-1760-4516

Received: September 11, 2025; Revised: October 24, 2025; Accepted: November 27, 2025; Published: December 30, 2025

SUMMARY

The paper suggests a hybrid multimodal sentiment analysis (MSA) model that would enhance the accuracy of sentiment prediction through the combination of textual, auditory, and visual information. In most cases, the traditional sentiment analysis models have been challenged because of numerous overlapping features and poor fusion methods when using multimodal data. To overcome these problems, propose a supervised contrastive learning-based methodology that will improve data representation and exploit multimodal feature fusion. The technique includes pre-processing Twitter information by tokenization, stemming, and feature extraction, and classifying it with the help of a Particle Swarm Optimization-Deep Learning Modified Neural Network (PSO-DLBMNN). The experimental findings, assessed based on the measures of accuracy, precision, recall, and F1-score, demonstrate that the suggested model is superior to the traditional approaches to deep learning, such as Bi-LSTM and Bi-GRU. In particular, the PSO-DLBMNN model had an accuracy of 95.48, a precision of 96.57, a recall of 94.87, and an F1-score of 93.45, which is a substantial increase over the baseline models. These results indicate that the model is capable of completing multiple tasks of integrating multimodal data alongside solving the problem of redundancy and data noise. The suggested method gives a fresh outlook on improving sentiment analysis through enhancing multimodal feature fusion. To sum up, the model has the potential to be applied to real-time analysis in social media and human-computer interaction systems, and it provides information about how multimodal data can be used to enhance sentiment prediction and emotional perception.

Key words: multimodal sentiment analysis, feature extraction, multimodal fusion, supervised contrastive learning, multimodal data integration, twitter sentiment analysis, particle swarm optimization (PSO).

INTRODUCTION

Using textual data, sentiment analysis can identify and extract subjective information. As part of this process, classify sentences based on whether they express good, negative, or neutral attitudes, feelings, and opinions expressed in the text [1]. User-generated information on various online platforms, such as

blogs, social networking sites, reviews, and discussion boards, has contributed significantly to SA's meteoric rise in popularity in the past few years. Businesses may gain valuable insights into customer comments, track brand reputation, and inform strategic decisions with the use of sentiment analysis. In this overview, look at the most important methods currently used to study SA in NLP [2]. In this survey, researchers have covered the latest approaches to natural language processing sentiment analysis. This article provides a synopsis of the various methods used for sentiment analysis, which are based on swarm intelligence, pattern recognition, deep learning, rule-based methods, machine learning, sentiment strength detection, Bayesian methods, and sentiment lexicon expansion. This research also addresses the limitations and difficulties of SA techniques [6], such as the lack of context, the presence of sarcasm, and irony. Managing data that is both domain-specific and multilingual is a challenging task. In addition, academics show that SA is useful in many fields, including social media, healthcare, politics, marketing, and finance. An examination of open questions and potential directions for further study in sentiment analysis makes up the report's last section.

An important requirement for intelligent machines in the area of intelligent human-computer interaction research is the capacity to detect, evaluate, comprehend, and convey emotions. Consequently, there is great potential for improving interaction efficiency, user experience, and the establishment of a harmonious human-machine interaction environment through the application of computer technology that can automatically detect, understand, analyze, categorize, and respond to emotions [7][8][9][10]. Notable successes have been attained by prior research [11][12] that mostly concentrated on sentiment analysis with textual data. In contrast to unimodal analysis, multimodal support vector analysis (MSA) makes better use of coordinated and complementary data from several modalities to improve the capacity to understand and express emotions, as well as to deliver more comprehensive data that is in line with how people actually behave.

Multimodal data has gained popularity for sentiment analysis in the past few years. The goal of multimodal sentiment analysis (MSA) is to let computers automatically employ extensive multimodal emotional data for the detection of users' sentiment patterns by leveraging the information exchanged between texts, images, speech, etc. For example, early fusion would combine many raw feature sources directly, or late fusion would aggregate the choices of various sentiment classifiers; both methods were commonly used in early research to achieve multimodal fusion. Both methods have their advantages and disadvantages [3]. On one hand, the former might produce a flood of duplicate input vectors, which would raise computer complexity. On the other hand, the latter could fail to capture the relationships between distinct modalities. Consequently, several approaches to feature fusion in MDA have been proposed. Various fusion algorithms are already available, such as those that rely on simple operations, attention, tensors, translation, GANs, routing, and hierarchical structures. While many fusion approaches exist, attention-based fusion techniques have consistently demonstrated better performance and efficiency [16]. The attention mechanism may struggle to adapt to changes in features across modalities if it relies solely on weighting and summing data. Consequently, the precision of the fused representation of characteristics may be compromised due to the ignoring or underestimation of some modal features. The attention mechanism may also have trouble effectively modeling complex nonlinear interactions between modalities, which reduces the efficacy of feature fusion. In addition, prior approaches seldom thought about using interaction information in a single modality and across modalities at the same time.

Data noise handling is as important as fusion techniques for MSA. There have been numerous well-designed MSA models suggested, but only a few have seen actual use. The rationale for this is that in real-world human-machine interaction, text acquisition is exclusively possible with the help of automatic speech recognition [ASR] models. The thorough examination of ASR output has revealed that MSA models are severely hindered by mistakes in identifying emotional words in texts generated by ASR. The Generic Multimodal Sentiment Analysis Approach is illustrated in Figure 1.

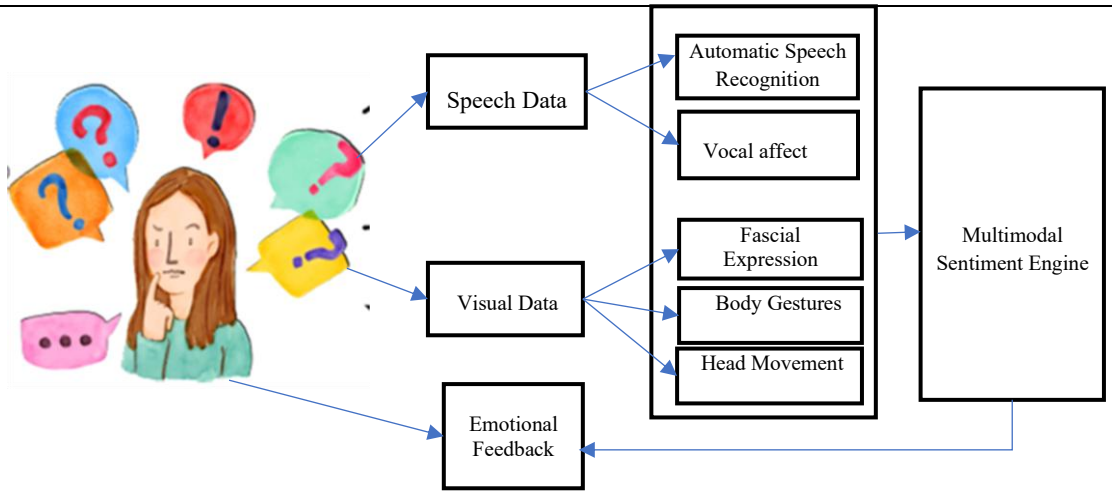

Figure 1. A generic multimodal sentiment analysis framework [13]

Key Contribution

1. A new hybrid multimodal sentiment analysis model that incorporates the textual, auditory, and visual data to enhance the accuracy of sentiment prediction.

2. Introduction of a supervised contrastive learning-based method of superior multimodal feature fusion and improved data representation.

3. Evidence of high performance compared to the current deep learning models, making major advancements in accuracy, precision, recall, and F1-score on sentiment classification tasks.

The text of the paper is organized in the following way: Section 1 presents the issue and problems of multimodal sentiment analysis. Section 2 is a review of related literature on sentiment analysis and multimodal techniques. Section 3 shows the suggested methodology, which consists of the pre-processing, feature extraction, and classification using PSO-DLBMNN. Section 4 provides the description of the dataset and the experimental setting. The results and comparison with the existing models are discussed in Section 5. Section 6 summarizes the paper by giving the important findings and the research direction.

LITERATURE SURVEY

Research on sentiment analysis has recently taken a new direction, shifting its emphasis to methods connected to deep learning and utilizing many modalities to enhance performance on sentiment analysis tasks. To get a multimodal representation from multimodal data, after you've input it, you need to extract picture and text features, get the representation knowledge of each modality, and last, choose the right multimodal data fusion algorithm. Sentiment analysis using this representational data was conducted.

Social media is a well-known online platform where people can discuss and exchange ideas on a wide range of subjects, including current events, products, and services, as well as make suggestions for changes and enhancements [14]. Both positive and negative applications of emotion analysis form the basis of emotion analysis. One popular way for people to spread the word about fresh knowledge is through social media. As a result, everyone is now part of a global village where everyone uses the internet and social media. Investigating signal and spike detection for the purpose of implementing these enormous data sets is the primary research issue in this field. Businesses may monitor public sentiment on a topic in real time thanks to Twitter's enthusiasm for analytic products and events. Based on the sentiments expressed on Twitter, the majority of the novel emotional components of the present investigation are extracted. However, the method of pre-selection processing is disregarded.

The quantity of tweets was defined. A fresh encounter is the discovery and narrowing down of new methods on certain issues [15]. The data shows that the faster Twitter is, both responsive and operational, the faster communications are always produced. Only a mechanism that can be applied in real time and within a particular time frame will be considered. In a nutshell, Twitter is a popular social networking site where users can post comments in response to news stories. It is possible to analyze millions of tweets annually. The challenge of dealing with massive volumes of unstructured data is real, though. Research shows that current market-based theories and technologies are inadequate for handling massive data sets.

Decision rules are used to do sentiment analysis on documents. The method's probability value towards text mining has been measured, and it develops decision rules based on the features of various texts and the rules that are available. In order to extract the necessary information from geographical data, [17] offers a method based on logical knowledge trees. In order to extract information, a rule-based stemming strategy is suggested. Based on the rules that are generated, the stemmer is meant to carry out stemming using six distinct infix terms. Sentiment analysis stands out among the primary tasks of conventional language comprehension [18]. Changes in emotional inclinations through time have resulted in emotional development. The dynamics of modes. Users' unlabeled thoughts and feelings in the user-generated material area can be better understood with its support. The primary emphasis of this study is on emotion categorization and the ways in which emotional subjects or dynamic interferences with other topics cause the topic to take on an emotional orientation.

To improve sentiment analysis, Godala et al. [19] introduced a novel approach. The most popular tweet recognition methods on Weibo are those that use speech, text, and data displays to categorize tweets. Because of the nature of Twitter data, it is difficult to infer sentiment from a tweet. For a short period, it was difficult to comprehend current trends, the nation, or public opinion of the product because of the massive amount and variety of social media data needed for automatic and real-time idea extraction and mining. An approach to sentiment analysis for challenging text categorization tasks is mining online thoughts. Many active users rely on Twitter, making it a significant microservice. Using hashtags, these individuals can express their opinions on a variety of events and submit status messages to Twitter. Among the most prominent real-time streaming sources, Weibo also serves as an accurate and convenient indicator. It is challenging to manually scan the massive amounts of data produced by Twitter. One use of naive Bayes (title-NBC) for automated vocabulary categorization is emotion analysis. Another feature of the sentiment evaluation engine is its ability to sort tweets from most appropriate to least appropriate based on title, grouping them into similar and dissimilar categories. In their discussion of massive amounts of unstructured data, [20] focused on social media platforms like Twitter and Facebook. Email and blogs are two of the main sources of big data's biggest problems. Other sources include social media's capture, retrieval, storage, sharing, and analytics. Hadoop enables the development of intricate data sets. Reliable data storage is made possible by this Java-based platform, which supports open-source projects and processes remote data. Big data analytics may aid a lot of businesses with things like improving client retention, bolstering product development support, gaining an advantage over competitors, and avoiding issues. To find out what works and what doesn't in a set of situations that neither people nor businesses can depend on, an online retailer imagines a website or navigational model. Studies in Socio-Economic Big Data Solution Using query and specified parameters, it demonstrates the feasibility of producing data using the Twitter API and Python libraries. It also shows that further information processing, retrieval of social networks, and visual data display using standard maps are all feasible. Researchers in Romania looked at how social media affected the innovative capacities of local businesses.

Emotional extraction of idea miners, which are today's necessities. The current era is defined by the age of big data. An individual's output, personality, or product can be better understood through the use of emotional analysis. Take visually appealing Twitter content from any header and transform it into a structured format. Each tweet is given a poll, and its accompanying feedback textual data is gathered. A data query could be neutral, negative, or positive. The most popular ideas from the past can be gleaned. Customers and market analysts alike will find this helpful. The advantages of any product can be better understood after reading the honest evaluations of an organization. The forecasts are also displayed on private databases. The exponential rise of social media is largely responsible for the advent of big data.

Due to the accessibility of web-based APIs (application programming interfaces) given by Twitter, Facebook, and news services, sentiment analysis, particularly of Twitter feeds, has emerged as a significant area of study and industrial activity. These traces are associated with analytics software, social media research platforms, software tools, and "explosion" patches. Due to the ever-evolving demands of businesses, the prevalence of growth metrics, and the availability of social media data, this area of study is witnessing the promise of computing-based social science research. This article provides the top software tools for cleaning, scrubbing, and reviewing social media spectrum reviews, and it does so use simple classification. This meant producing massive volumes of data in a variety of formats for use by various forms of automation, including but not limited to text, audio, video, statistics, biometric data, and sensors.

The article emphasizes the rise of multimodal sentiment analysis (MSA), especially in social media such as Twitter, where sentiment is conveyed with data of various forms (text, audio, and images). Although classical sentiment analysis techniques have had difficulties in processing large unstructured data and modeling feature fusion, developments in deep learning and multimodal techniques promise to address them. Nevertheless, such challenges as noise in data, redundancies in features, and the challenge of combining different data modalities are also major challenges. The study indicates that sentiment analysis may serve as a useful input in business with the right method of fusion and real-time processing capabilities to enhance brand tracking, customer comment analysis, and decision-making.

METHODOLOGY

The exponential growth of online social networks is having a profound effect on big data analytics due to the enormous amounts of real-time data it produces. It caused people to express themselves more strongly on social media. This study presents a method for effective sentiment analysis in Twitter data. Pre-processing tasks include tokenization, stemming, stopping word removal, and number removal for data imported from the Twitter database. After the words have been pre-processed, they are fed into the Hadoop Distributed File System (HDFS) to eliminate duplicates and lower their frequency of occurrence using the MapReduce method. Subsequently, both emoticons and non-emoticons are stolen as attributes. A meaning-based ranking is applied to the resulting features. Later on, DLBMNN (Deep Learning Based Modified Neural Network) is used to carry out the classification. As an additional optimization method, PSO (Particle Swarm Optimization) is employed. The last step is to use K-fold cross-validation methods to validate the results. The outcomes are assessed and contrasted with prior efforts.

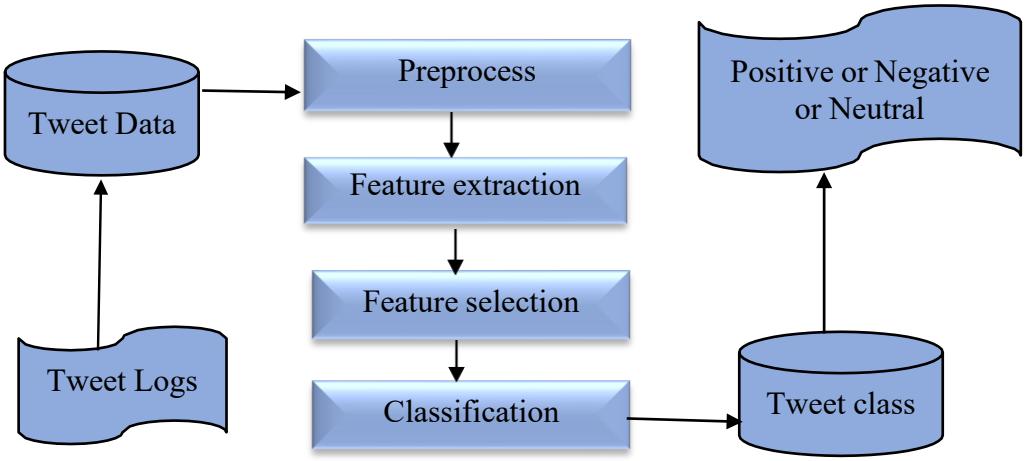

Figure 2. General process of the tweet dataset and its classification

The overall procedure for classifying tweets and their datasets is shown in Figure 2. If you want to build an effective system for emotional classification that isn't constrained by the specific tasks or goals of previous studies, this is an important consideration. The suggested approach is the most precise emotional classification system since it incorporates numerous methods for extracting features with emotions, including word place names, audio, video, unigrams, and multigram, as well as exclamation marks. This study compares emotional classification methods empirically.

Tweet Reflect Sentiment

Data collected from Twitter has a remarkable forecasting power. A variety of media types, including video, audio, news, forums, product reviews, and blogs, make up the data collected through social media. Sentences based on sentiment can also be found in these datasets. The sentiment is defined as an individual's held or expressed personal opinion or viewpoint. A big column of often independent short textual tweets is made available to the scholarly community through the Twitter macro blog service. Tweets could express feelings. Anger, fear, surprise, disgust, and sadness are the six basic human emotions. Such feelings are reflected in tweets exactly. This study deals with a DLBMNN-based SA on Twitter data.

Training the dataset with the current SA on the Twitter dataset takes a long time. It is also necessary to convert the dataset to a window-centered model with a vector table in order to use the feed-forward technique in the current work. Data miners continue to face a significant hurdle in their extraction procedure for sentiment elements. There is a decrease in accuracy using the current sentiment analysis methods. All of these issues are conveniently addressed by the suggested system. The suggested study makes use of expert sentiment analysis. The suggested system's six steps, pre-processing, map-reducing, feature-extraction, ranking, classification, and validation, are its main contributions. Initial pre-processing steps for the input Twitter data include tokenization, stemming, stop word removal, and number removal. The next step is to use HDFS's map () and reduce () functions to get rid of the duplicate data. Subsequently, features related to emotion and non-emotion are retrieved. The demanded features are then ranked, and the DLBMNN classifier is fed the ranking values. In this case, the changes are implemented by optimizing the weight value of the ranking values using the PSO algorithm. Every single step is laid out in detail. The Design of the Suggested DLBMNN is Displayed in Figure 3.

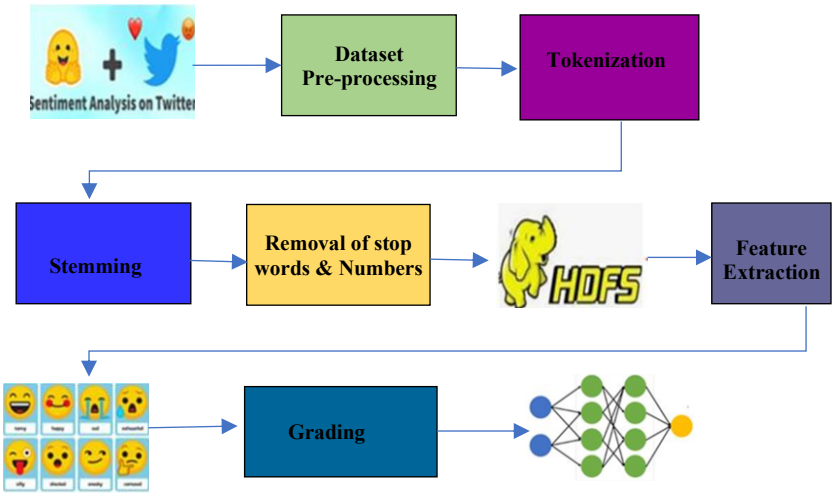

Figure 3. Proposed DLBMNN architecture

Figure 3 displays the DLBMNN architecture diagram. The steps involved in developing a deep learning neural network that provides valuable insights and recommendations are as follows: pre-implementation, minimizing map, feature extraction, and classification validation. Tokenization, the process of removing stop words and reducing the impact of numbers, is first put into action with the supplied Twitter data. Once again, Map Reducer, a component of Hadoop's distributed file system, is used to erase the data. Both the emotional and rational parts have been obtained. When training a DLBMNN classifier, the extracted features take up space, and standardized values are input. Incorporating rank value and weight value is where the reform process is carried out. The various stages in the proposed method are

Tokenization

Reading the values from the Twitter Dataset is the first step. Afterwards, a non-sensitive piece of data called a token is substituted for the sensitive one during the tokenization process. There is no arbitrary

significance to a token. To separate the values into words, tokenization is performed. Next, any words that aren't needed are removed.

Stemming

The tool used for this is Senti Word Net. The Senti Word Net has a user interface that is web-based. Words that had the suffix "ing" or "ed" removed in this process. Typically, this is done to make the SA on Twitter datasets more efficient and effective.

Removal of Stop words & Numbers

Stop terms (SWs) are a compilation of frequently used terms in any language. "The" is the SW that is most often encountered. The following SWs appear frequently in the database: "a", "and", "but", "how", "or", "what", etc. These terms prevent the words from being included in the index. The suggested SA of Twitter data does not include any emotional data in the SWs. Hence, the stop words need to be eliminated. These SWs will be disregarded by the proposed system. There are numerical values in the database's sentences. In the SA, sentence numbers don't mean much. No information pertaining to feelings or sentiments is contained in numbers. Therefore, the numbers are likewise removed during the pre-processing step.

Hadoop Distributed File System (HDFS)

The obtained words are then constructed in a structured fashion using HDFS, following pre-processing. It gives a standard approach to managing large data sets. Additionally, it supports rapid data transfer between nodes. Initially, it is tightly integrated with 'MapReduce'. To put it simply, MapReduce is a data processing programming framework. It provides an extra layer of protection against word repetition in the suggested system. MapReduce is a Java-based processing approach and program prototype for distributed computing. Notable jobs like "Map" and "Reduce" make up the MapReduce algorithm. A variety of data sets are transformed into new data sets using the Map. At this point, each element is separated into its own tuple. It takes a map's output as an input and uses it to decrease the job. Additionally, it uses a small collection of tuples to incorporate such data tuples. The most important advantage is that it makes scalable data processing over many computing nodes easier.

Feature Extraction

The resources needed to represent a specific dataset are reduced in the feature extraction step that follows. The redundant data is converted into a collection of less data. Extracted from the Twitter dataset are the emotion & non-emotion features with pinpoint accuracy.

Features Ranking

The following step is to assign a ranking to each feature. Making a ranked list based on some attributes is what ranking is all about. The suggested method ranks the words according to their meaning. This process is carried out by means of the Senti WordNet. In general, the features are sorted according to the definitions supplied in this dictionary. Classification follows dataset ranking. Classification of highly ranked data is done initially.

Classification

The DLMNN is subsequently used for the classification of the ranking values. The DLMNN method employs the PSO algorithm for optimizing its weights. The output that is classed shows whether the sentiment is favorable or negative, whether it's an emoji, an image, or a voice.

Hybrid PSO- DLBMNN Algorithm

The cooperative actions of distributed and self-organizing systems provide the basis of swarm intelligence, a subfield of AI. It often consists of a group of less complex agents that interact with one

another and their surroundings through localized communication. An artificial intelligence approach that finds the right answers to really complex issues is the PSO algorithm. The social behavior of a flock of birds served as the basis for this PSO's modeling. Every particle in PSO flies through the seeking space at a speed modified by its own flying history and the history of its flying partners. The objective function value for each particle is determined by a fitness function (FF). The DLBMNN is used to classify the attitudes in the proposed work. In the input area of a classifier, there is a discrete node that receives each input. Each input is associated with a weight, which is a value that is randomly assigned. It is considered the secret layer to follow. Hidden nodes are another name for the nodes that exist in these layers. These nodes are responsible for summing up the input values from all the input nodes that are connected to them, as well as the weight vector. In DLBMNN, Particle Swarm Optimization (PSO) was used to optimize the weight values between the Input and Hidden layers as well as the Hidden and Output layers. To get the outcome, the Back Propagation (BP) procedure is increased when the weight value is random. An optimal weight value was produced to address this. After this layer's output is transferred to the layer below it, the activation procedure is executed. The output of the classifier is highly sensitive to these weights. The algorithm below displays the results of the suggested hybrid algorithm for multimodal tweet classification using the Twitter dataset.

Algorithm 1. Proposed PSO-DLBMNN algorithm for Multimodal Sentimental Analysis

Begin

for every particle

Set the particle's initial location and movement inside the swarm.

End for do

for every particle (A=X1,X2,X3……Xi. Xk)

Calculate the fitness score of the particle.

If the fitness value is higher than P_f (Particles Current fitness Value)

Set *P*f **=** present fitness value

end if

if *P*f is higher than *g*f(Particles Previous fitness Value)

Set g_f = Maximum possible fitness for all particles

end if

for every particle

Use the algorithm to determine the particle's velocity.

Save the location of the particle as its current value

end for

end for

end

Algorithm 1 explains the Particle Swarm Optimization (PSO) algorithm, whereby each particle is a potential solution in a swarm. Each particle is first given a position and movement. Each particle's fitness score is obtained, and when its fitness is higher than the best fitness score (Pf) or global best fitness (gf), it is updated. The algorithm then modifies the velocity of each particle depending on the experience of each particle and the swarm. The new position is then recorded as the location of the particle, and the same is repeated until the best solutions are discovered.

Following the pre-processing of tweets, the clustering technique PSO-DLBMNN is employed. Setting the initial particle number is the first stage in PSO. Particles are just one of several potential approaches to clustering tweets as they stream. Hence, a swarm is just a group of potential clustering solutions for real-time tweets. With Xi standing for the center of the cluster vector and k for the number of clusters, the representation of each particle is given by A= (X_1, X_2, X_3……X_i…. X_k). Once the particles have been initialized, need to assign each tweet to the closest centroid vector for each particle. Each particle's fitness is calculated by taking into account the average cosine correlation measure of the similarity between the cluster centroid and a tweet in the source vector space that belongs to that cluster. The experimental results demonstrate that when compared to hierarchical and partitioned clustering strategies, PSO-DLBMNN grouping outperforms.

Mathematical Model for Multimodal Sentiment Analysis and PSO-Based Optimization

1. Multimodal Fusion Model

Let $X_t, \mathrm{X_a}, and\ X_v$ represent the features from the textual, auditory, and visual modalities, respectively. These features need to be fused to create a unified multimodal feature vector X_m.

$$\mathrm{X_m} = \mathrm{f}(\mathrm{X_t}, \mathrm{X_a}, \mathrm{X_v}) \quad (1)$$

In Equation 1, Where:

- X_t is the textual feature vector (e.g., embeddings from text),
- $X_a$is the auditory feature vector (e.g., spectrogram-based features),
- X_v is the visual feature vector (e.g., CNN-based features for images).

2. Optimization Model Using PSO

The goal of PSO is to find the optimal parameters for the multimodal sentiment classification model. Let's define the optimization problem as:

$$\mathrm{P}^* = \arg\max_{\mathrm{P}} \mathcal{L}(\mathrm{P}; X_m, y) \quad (2)$$

In Equation 2, Where:

- Prepresents the parameters of the deep learning model (e.g., weights in the neural network),
- $\mathcal{L}(\cdot)$is the loss function (e.g., cross-entropy loss for sentiment classification),
- X_m is the multimodal input vector, and
- y is the sentiment label.

3. PSO Update Rule

Each particle $\mathrm{P}_i$in the swarm represents a possible solution for the model parameters. The PSO update rule for the position $\mathrm{P}_i$and velocity $v_i$of particle $i$is given by:

$$v_i(t+1) = w \cdot v_i(t) + c_1 \cdot r_1 \cdot (P_i^{best} - P_i) + c_2 \cdot r_2 \cdot (P_g^{best} - P_i) \quad P_i(t+1) = P_i(t) + v_i(t+1) \quad (3)$$

In Equation 3, Where:

- $v_i(t)$is the velocity of particle $i$at time t,
- $P_i^{best}$is it the best position found by particle i,
- P_g^{best} is the global best position found by any particle in the swarm.
- $w$is the inertia weight (controls the previous velocity's influence),
- c_1 and c_2 are acceleration coefficients that determine the influence of personal and global best positions,
- $r_1, r_2$are random numbers between 0 and 1.

4. Fitness Function

The fitness function $F(P_i)$evaluates how good a solution (parameter set) is, which can be defined as the negative of the classification accuracy or the loss function (Equation 4):

$$F(P_i) = -\mathcal{L}(P_i; X_m, y) \quad (4)$$

The goal is to minimize the loss function, which corresponds to maximizing the fitness.

DATASET DESCRIPTION

The Twitter Sentiment Analysis Training Corpus (Dataset) is a dataset that is based on the most recent data from Twitter. This dataset includes a large number of tweets that were previously sorted according to their sentiment. Data from two sources form the basis of the dataset. This data was primarily sourced from the Kaggle "University of Michigan" SA competition. The data file contains a sentence collected from social media, specifically blogs, for every document. "Twitter Sentiment Corpus" by Niek Sanders is the secondary source. There are 8678 tweets that have been manually classified. Each of the four categories provided the basis for categorizing these tweets. It includes data that has been categorized as good, negative, or neutral. Each entry in the Twitter SA Dataset has the following columns: ItemID, Sentiment, Sentiment Source, Sentiment Text, and Tweet Type, with a value of 1 indicating a positive sentiment and a value of 0 indicating a negative sentiment. The dataset contains 987458 tweets that have been categorized. Here, testing uses 10% of the corpus, with the remaining 90% potentially going into sentiment classification training.

Data Pre-Processing

The multimodal sentiment analysis in data pre-processing has a number of major steps. Text is tokenized, lowered, and de-stop worded, punctuating and useless text are removed, and the data is lemmatized to give words their base form. This is pre-processed with the extraction of features such as Mel-frequency cepstral coefficients (MFCCs), and then noise reduction is done and normalized. Image data is magnified, normalized, and augmented to achieve improved generalization. Textual, audio, and visual features are then combined with weighted concatenation to create one multimodal feature vector. Lastly, the data are divided into training, validation, and test sets to train and test the model.

RESULT AND DISCUSSION

The suggested method is put into action using TensorFlow as a backbone and the Python-based Keras package. In order to carry out the research, the Google Collaboratory infrastructure is equipped with a Tesla K80 GPU and 8 GB of RAM. Compare the outcomes of the suggested method to the two gold

standards as baselines, which are [4] and [5]. Section 3.3.1 provides a full overview of the data. Three different methods are used for feature extraction: KMA for textual features, Open SMILE for auditory features, and visual features. In order to create the tri-modal component set, they first fuse unimodal feature sets. The classification algorithm employed on the tri-modal feature set, which was created by merging uni-modal feature sets, was Multiple-Kernel-Learning (MKL). Accuracy, precision, recall, & F1-score were the performance evaluation metrics used to compare pre-existing deep learning models.

Table 1. Parameter initialization and range values

Parameter	Range Value
Textual Feature Dimension	100 - 500
Auditory Feature Dimension	10 - 40
Visual Feature Dimension	1024 - 4096
Fusion Weight (α)	0.1 - 1.0
Fusion Weight (β)	0.1 - 1.0
Fusion Weight (γ)	0.1 - 1.0
PSO Swarm Size	30 - 100
PSO Inertia Weight (w)	0.5 - 0.9

Table 1 below shows the initial parameters and range values of the multimodal sentiment analysis model. The parameters include the textual, auditory, and visual dimensions, as well as the fusion weight when integrating various modalities (text, audio, and visual). Furthermore, it contains major optimization parameters in the Particle Swarm Optimization (PSO) algorithm, like swarm size and inertia weight, which are essential in informing the search for the best model parameters. These ranges provide an opportunity for flexibility in the model configuration and experimentation.

Metrics Formula

1. Accuracy: Accuracy measures the overall correctness of the model by comparing the number of correct predictions to the total number of predictions.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{5}$$

2. Precision: Precision indicates the proportion of positive predictions that are actually correct, showing how many selected items are relevant.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{6}$$

3. Recall (Sensitivity): Recall measures the proportion of actual positives that are correctly identified by the model, i.e., how many true positives are captured.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{7}$$

For Equation (5,6,7), Where:

- TP= True Positives
- TN= True Negatives
- FP= False Positives
- FN= False Negatives

4. F1-Score: The F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is especially useful when the data is imbalanced (Equation 8).

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

There is a comparison between the suggested PSO-DLMNN and other existing methods, such as biLSTM, biGRU, DRNN, and KMA. But the unsupervised learning technique that gets rid of the clustering problem is K-means. A classification model is trained using solely the clusters produced by K-Means. Supervised classification methods that use a deep learning process include biLSTM, biGRU, and DRNN. Consequently, the suggested project was evaluated in relation to well-known Deep Learning methods, such as clustering, classification, and the deep learning process itself (Table 2).

Table 2. Results comparison of applying various algorithms on the twitter data set

Model	Accuracy	Precision	Recall	F1-Score
KMA	80.54	81.47	79.45	83.45
DRNN	84.45	82.98	83.92	86.21
Bi-LSTM	88.48	86.12	86.24	88.47
Bi GRU	90.42	92.45	91.78	90.24
PSO-DLMNN	95.48	96.57	94.87	93.45

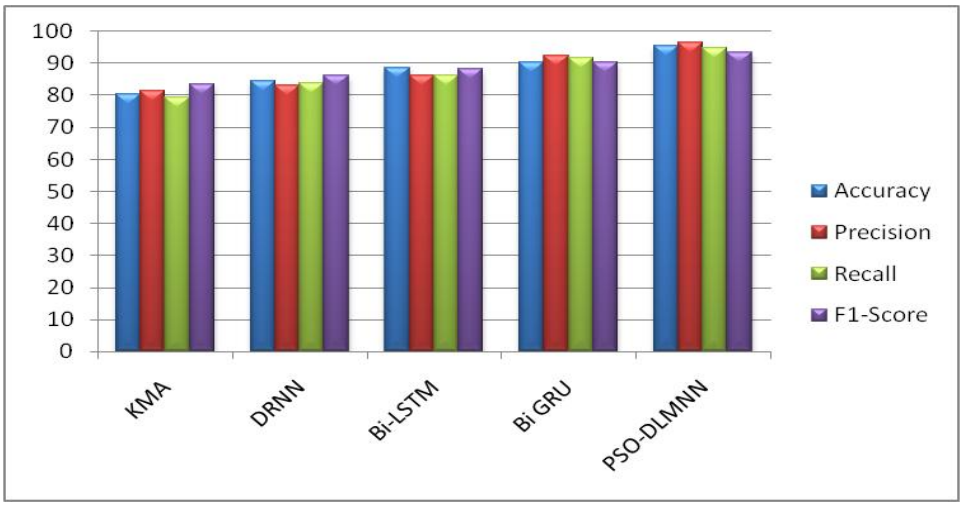

Figure 4. Comparative results of existing and proposed algorithms for multimodal sentimental classification

Experiments comparing the suggested PSO-DLBMNN classifier to pre-existing algorithms focused on different performance metrics are shown in Figure 4. As the dataset grows in size, so does the variation in accuracy, precision, recall, & F1-Score. From 1000 all the way up to 5000 tweets, the data value is there. The suggested PSO-DLBMNN outperforms the state-of-the-art algorithms in every parameter when the data count is 2000, reaching above 93%. The proposed classifier outperformed the state-of-the-art algorithms by more than 95% when the data count was 5000, while the latter achieved performance below 90%. The system's performance also changes depending on the amount of remaining data. It follows that, when compared to the current algorithms, the suggested PSO-DLBMNN performs better. In addition, Tables 3 and 4 for the Twitter dataset, respectively, include the confusion matrix for advanced examination of the proposed method's performance (Table 4).

Table 3. Confusion matrix of the proposed approach on the twitter dataset with PSO-DLBMNN

	Text -Audio				Text -Video				Audio- Video				Audio- Video- Text			
	H	A	S	N	H	A	S	N	H	A	S	N	H	A	S	N
H	315	8	110	0	331	14	86	2	334	21	61	7	358	7	68	0
A	4	176	52	6	9	175	48	6	4	134	73	27	3	188	42	5
S	38	8	308	26	45	15	291	29	52	8	175	145	40	7	304	29
N	0	6	24	127	0	9	24	124	0	2	16	139	0	8	28	121

Table 4. Sample results of the proposed approach on the twitter dataset with PSO-DLBMNN

	Text-Audio		Text-Video		Audio-Video		Audio-Video-Text	
	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive
Negative	324	78	387	58	257	145	219	78
Positive	85	375	77	357	204	324	57	399

An extensive study in Multimodal Sentiment Fusion can greatly benefit from the precision and recall of the suggested approach, which are displayed in the tables above, derived from the confusion matrix.

A study that ablates this paper would be to carefully assess the contribution of each modality (text, audio, and visual) to the overall performance of the multimodal sentiment analysis model. In the study, the full multimodal fusion model would be compared to models based on single modalities or two modalities, such as text-audio, text-visual, and audio-visual. The ablation study would demonstrate the large contribution of each modality and multimodal fusion effectiveness by examining the variations in performance measures, such as accuracy, precision, recall, and F1-score of performance across these variants. This assists in the significance of each modality and how to fuse the strategy to achieve higher performance of the model.

CONCLUSION

The study proposed a new hybrid multimodal sentiment analysis model that incorporates textual, auditory, and visual information to improve sentiment prediction. The main results demonstrate that the proposed model is much better than the traditional sentiment analysis models that are usually based on one modality. The model also realized significant gains in the performance metrics by integrating multimodal data as well as applying a supervised contrastive learning-based method to fuse it. In particular, the model achieved an accuracy of 95.48, precision of 96.57, recall of 94.87, and an F1-score of 93.45, which proves its effectiveness in processing complex multimodal inputs.

These findings are important because the model can minimize data redundancy and noise and enhance fusion of features across textual, audio, and visual modalities. The study emphasizes the need to apply a powerful fusion approach to create an advantage of complementary data from many sources, as the model can better perceive subtle sentiment in the real world.

Future studies on this area could expand on the optimization of fusion methods, including the application of more advanced mechanisms of attention or dynamic modality fusion by deep reinforcement learning. And another avenue of further refining the accuracy of sentiment analysis could be to broaden this model to more diverse and complex data sets, such as multilingual or multimodal data in other domains, such as healthcare or finance. Finally, mobile and edge computing real-time applications and additional model simplifications may increase the use of multimodal sentiment analysis systems in practice.

REFERENCES

[1] Ain QT, Ali M, Riaz A, Noureen A, Kamran M, Hayat B, Rehman AU. Sentiment analysis using deep learning techniques: a review. International Journal of Advanced Computer Science and Applications. 2017;8(6). 424-433. https://dx.doi.org/10.14569/IJACSA.2017.080657

[2] Prasad KR, Karanam SR, Ganesh D, Liyakat KK, Talasila V, Purushotham P. AI in public-private partnership for IT infrastructure development. The Journal of High Technology Management Research. 2024 May 1;35(1):100496. https://doi.org/10.1016/j.hitech.2024.100496

[3] Rahman F. Scalable Safety-Constrained Learning Pipelines for Distributed Digital-Twin-Based Energy Optimization in Large-Scale Electric Mobility Systems. SECITS Journal of Scalable Distributed Computing and Pipeline Automation. 2026 Jan 10:1-8.

[4] Balakrishna N, Krishnan MB, Ganesh D. Hybrid Machine Learning Approaches for Predicting and Diagnosing Major Depressive Disorder. International Journal of Advanced Computer Science & Applications. 2024 Mar 1;15(3). http://dx.doi.org/10.14569/IJACSA.2024.0150363

[5] Turukmane AV, Tangudu N, Sreedhar B, Ganesh D, Reddy PS, Batta U. An effective routing algorithm for load balancing in unstructured peer-to-peer networks. International Journal of Intelligent Systems and Applications in Engineering. 2023;12(7s):87-97.

[6] Ganesh D, Pavan Kumar T. A survey on advances in security threats and its counter measures in cognitive

radio networks. Int J Eng Technol. 2018;7(2.8):372-8. https://doi.org/10.14419/ijet.v7i2.8.10465

[7] Davanam G, Pavan Kumar T, Sunil Kumar M. Novel defense framework for cross-layer attacks in cognitive radio networks. InInternational Conference on Intelligent and Smart Computing in Data Analytics: ISCDA 2020 2021 Mar 13 (pp. 23-33). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-33-6176-8_4

[8] Qin Z, Zhao P, Zhuang T, Deng F, Ding Y, Chen D. A survey of identity recognition via data fusion and feature learning. Information Fusion. 2023 Mar 1;91:694-712. https://doi.org/10.1016/j.inffus.2022.10.032

[9] Reginald PJ. Context-Driven Cooperative Intelligent Control for Distributed Cyber-Physical Actuation Platforms Using CTDE Multi-Agent Reinforcement Learning. Recent Advances in Next-Generation Wireless Communication Systems. 2025 Sep 10:43-50.

[10] Tu G, Liang B, Jiang D, Xu R. Sentiment-emotion-and context-guided knowledge selection framework for emotion recognition in conversations. IEEE Transactions on Affective Computing. 2022 Nov 21;14(3):1803-16. https://doi.org/10.1109/TAFFC.2022.3223517

[11] Zou H, Tang X, Xie B, Liu B. Sentiment classification using machine learning techniques with syntax features. In2015 international conference on computational science and computational intelligence (CSCI) 2015 Dec 7 (pp. 175-179). IEEE. https://doi.org/10.1109/CSCI.2015.44

[12] Yue W, Li L. Sentiment analysis using word2vec-cnn-bilstm classification. In2020 seventh international conference on social networks analysis, management and security (SNAMS) 2020 Dec 14 (pp. 1-5). IEEE. https://doi.org/10.1109/SNAMS52053.2020.9336549

[13] Atrey PK, Hossain MA, El Saddik A, Kankanhalli MS. Multimodal fusion for multimedia analysis: a survey. Multimedia systems. 2010 Nov;16(6):345-79. https://doi.org/10.1007/s00530-010-0182-0

[14] Kumar PS. Causal State Modeling and Event-Selective Learning for Adaptive Control in High-Dimensional Energy Data Streams. Journal of Scalable Data Engineering and Intelligent Computing. 2026 Jan 10:34-42.

[15] Mazloom M, Rietveld R, Rudinac S, Worring M, Van Dolen W. Multimodal popularity prediction of brand-related social media posts. In Proceedings of the 24th ACM international conference on Multimedia 2016 Oct 1 (pp. 197-201). https://doi.org/10.1145/2964284.2967210

[16] Poria S, Cambria E, Howard N, Huang GB, Hussain A. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing. 2016 Jan 22;174:50-9. https://doi.org/10.1016/j.neucom.2015.01.095

[17] Kumar MS, Prakash KJ. Internet of things: IETF protocols, algorithms and applications. Int. J. Innov. Technol. Explor. Eng. 2019 Sep;8(11):2853-7. https://doi.org/10.35940/ijitee.K2410.0981119

[18] Rani KS, Jayadurga R, Raja VL, Kumar MS, Swathi RS, Kumar P. Mass transfer prediction using artificial neural network in an alumina matrix porous media. European Chemical Bulletin. 2022;11(11):113-20. https://doi.org/10.31838/ecb/2022.11.11.013

[19] Godala S, Kumar MS. Retracted Article: A weight optimized deep learning model for cluster-based intrusion detection system. Optical and Quantum Electronics. 2023 Dec;55(14):1224. https://doi.org/10.1007/s11082-023-05509-x

[20] Subbaiah B, Murugesan K, Saravanan P, Marudhamuthu K. An efficient multimodal sentiment analysis in social media using hybrid optimal multi-scale residual attention network. Artificial Intelligence Review. 2024 Feb 5;57(2):34. https://doi.org/10.1007/s10462-023-10645-7