Feature Fusing with Vortex-Based Classification of Sentiment Analysis Using Multimodal Data

V. Sunil Kumar, S. Renukadevi, B. M. Yashaswini, Vindhya P. Malagi, and Piyush Kumar Pareek

Abstract For the purpose of identifying the felt emotions and intentions behind multimodal data, multimodal sentiment investigation seeks to semantic info acquired across different modalities. The primary focus of this field of study is the design of a novel fusion system that can efficiently and effectively aggregate data from several sources. However, the ability to use the independence and connection across modalities is lacking in prior work, preventing optimal performance. Consequently, the work suggests a unique approach to visual and textual sense modality and then fuses these features using different kernel learning techniques (MKL). After that, we feed the combined dataset into an Extreme Learning Machine (ELM), where we use the Vortex Search Algorithm to choose the most appropriate bias and weight (VSA). When everything is said and done, trials are run on two publicly accessible datasets, and the results are better than the current gold standard for multimodal sentiment analysis. For instance, whereas the current ML only manages 94% accuracy, the suggested ELM-VSA achieves 98.50%.

Keywords Multimodal sentiment analysis · Multiple kernel learning · Extreme learning machine · Vortex search algorithm · Feature fusion

V. Sunil Kumar (✉) · P. K. Pareek
Nitte Meenakshi Institute of Technology, Bengaluru, Karnataka, India
e-mail: sunil.kumar@nmit.ac.in

P. K. Pareek
e-mail: piyush.kumar@nmit.ac.in

S. Renukadevi · B. M. Yashaswini · V. P. Malagi
Dayananda Sagar College of Engineering, Bengaluru, Karnataka, India
e-mail: renukadevi-aiml@dayanandasagar.edu

B. M. Yashaswini
e-mail: yashaswini-aiml@dayanandasagar.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

N. R. Shetty et al. (eds.), *Advances in Computing and Information*, Lecture Notes in Electrical Engineering 1104, https://doi.org/10.1007/978-981-99-7622-5_32

1 Introduction

These days, most people get their news and information from social media platforms like disputes, and economic discussions are all commonplace on the web [1]. Customers dislike traditional stores because of the time commitment involved, and instead like purchasing online. The majority of their goods purchases are made on e-commerce sites like as Amazon, Flipkart, Snapdeal, etc. After making a purchase, consumers often share their thoughts on it on platforms like Facebook, Twitter, and WhatsApp [2, 3]. Customers may use the information provided by their reviews to make educated purchasing decisions by weighing the items' benefits and drawbacks against those of competing products and companies.

There is a lot of data being shared across all internet users, which makes it hard to find the relevant data. Since this is the case, Natural Language Processing (NLP) is increasingly interested in cutting-edge studies using Machine Learning (ML) [4]. Naive Bayes, (SVM), Decision Tree, etc., are only few examples of the classifiers that may be generated using ML methods. It is not sufficient to grasp textual modality, feelings or views, and emotions derived from opinions [5, 6]. Data on facial expressions may be found in the visual modality, whereas both video and textual data provide strong indicators for learning about emotions. Interest in solving the challenge of extracting actionable insights from social media data has led to the development of a number of Multimodality Sentiment Analysis algorithms that include text, audio, and video elements. Emotions and analytical approaches may be generated by the grouping of visual modalities [7].

In general, the presence of many modalities within the same data segment is typically a positive thing, since it provides more information for semantic and affective disambiguation [8]. The multimodal fusion component is vital for MSA since it is this aspect that attempts to integrate info from all input sense modalities to comprehend the emotion underlying the viewed data. With the goal of improving prediction accuracy, multimodal fusion integrates data from several sources that are often complimentary to one another [9]. Early fusion and model fusion are the three main categories into which fusion methods may be categorised [10]. In early fusion, characteristics from different modalities are combined as soon as they are retrieved. Combining many characteristics into a single one before feeding it into a classification or regression model is the simplest kind of early fusion. Early fusion has the potential benefit of recording the interplay between many modalities [11]. However, because of the varying sample rates, it requires feature alignment, which may be problematic due to high dimensionality [12]. The unimodal predictions are fused using voting, weighting, or an extra learnt model during late fusion, which makes it possible to utilise a different model for each modality. The increased adaptability afforded by late fusion is applicable to all unimodal models. But, it does not take into account the granular interplay between the many modes of expression. The goal of model fusion is to combine the benefits of early fusion and late fusion into a single, cohesive process.

Grouping characteristics into categories and assigning a unique kernel function to each category are central to MKL, a feature selection approach [13]. MKL helped even more since it efficiently combined data from several modalities, which is something we had been struggling with up until now. The subsequent is a synopsis of the paper's findings and their implications:

- The article successfully detects emotion regardless of the speaker by combining video, audio, and text modalities. The first originality is the usage of MKL to integrate the three approaches. We employ several kernels to adapt to distinct modalities and, as a result, obtain more accuracy than previous methods, which use a single kernel classifier to fuse all three senses.
- Finally, ELM does the sentiment analysis, while VSA is utilised to determine the weight and bias of ELM optimally.

The outstanding parts of the paper are structured as shadows: Methods currently in use for analysing emotional content are analysed in Sect. 2. Section 3 delivers a quick overview of the suggested perfect. In Sect. 4, we see the results of comparing the proposed model to the current methods for validation. Section 5 wraps up the study and suggests where the research should go next.

2 Related Works

With the goal of extracting datasets, Ye et al. [14] have designed a cross-modal data. This research takes a two-pronged approach to extracting meaning from text and pictures by using masked language modelling and masked auto-encoders. Our SMP model outperforms the state of the art on a variety of multimodal sentiment classification studies, both at the sentence level and with a focus on specific targets. As an added measure, we undertake ablation experiments and case studies to prove that our SMP model is reliable.

A new framework, HyCon, for hybrid contrastive illustration has been proposed by Mai et al. [15]. The model is able to completely investigate cross-modal interactions, learn inter-sample and inter-class associations, and lessen the modality gap since the research conducts concurrently. Additional tools, such as the refinement term and the modality margin, are provided to facilitate enhanced learning of unimodal pairings. In addition, it creates a pair selection system to find the most illuminating negative and positive pairings and give them appropriate weights. By generating numerous training pairings automatically, HyCon mitigates the unfavourable impact of small datasets and improves generalisation. Multimodal sentiment analysis and emotion detection are two areas where the technique has been shown to excel above baselines in extensive experimental testing.

Two-phase multi-task is the name of a multimodal framework proposed by Yang et al. [16] (TPMSA). It uses a unique multi-task learning technique to explore the classification capacity of each representation and a two-stage training strategy to maximise the use of the pre-trained model. The research performed tests on MOSI and

CMU-MOSEI. The findings demonstrate the superiority of our suggested approach over the state-of-the-art SOTA technique on both datasets, spanning the majority of criteria.

In this work, Yan et al. [17] offer a multi-tensor fusion which can successfully be employed for multimodal emotional intensity prediction by capturing intra-modal dynamics and inter-modal interactions. The suggested technique significantly outstrips state-of-the-art alternatives on the CMU-MOSI and CMU-MOSI datasets. While the suggested method only attempts to keep coarse-grained modal fusion, additional work on fine-grained modal fusion is planned to further decrease the sum of parameters needed by the model and enhance the precision with which inferences and predictions of emotional intensity may be made.

From the multimodal datasets, Salur et al. [18] presented a deep feature extraction using deep learning techniques (BiLSTM, CNN). The resulting feature sets were then categorised using a soft voting-based ensemble learning model, which was undertaken after feature selection was performed on the picture features. The suggested model's efficacy was evaluated using two separate benchmark datasets of text-image pairings. Experiment results showed that the suggested perfect outperformed other competing models on the same datasets.

Chen et al. [19] weighted cross-modal attention mechanism which takes into account both the temporal correlation information and the spatial dependency information of each modality, while also dynamically adjusting the weight of each modality across various time steps. Multimodal tasks are co-trained with their equivalent unimodal subtasks so that the mutually beneficial interactions between the modalities may be studied in greater depth. The model greatly outperforms previous attempts on the CMU-MOSI dataset, setting a new state-of-the-art record. Our model outperforms the competition by a wide margin, as evidenced by the fact that it obtains the highest F1-score on the CMU-MOSEI dataset for the binary classification, the seven-class task, and the regression task and is only outperformed in accuracy by the multimodal split attention fusion (MSAF) perfect with aligned data for the classification.

3 Proposed Methodology

Figure 1 shows working flow of projected model.

3.1 Problem Definition

Unimodal raw sequences X mR(l md m) from the same video fragment are used as input for models in MSA tasks, with l m denoting the arrangement length and d m the dimension of the representation vector for modality m. In this study, we focus on the combination of text, visual, and auditory data, denoted by the notation mt, v,

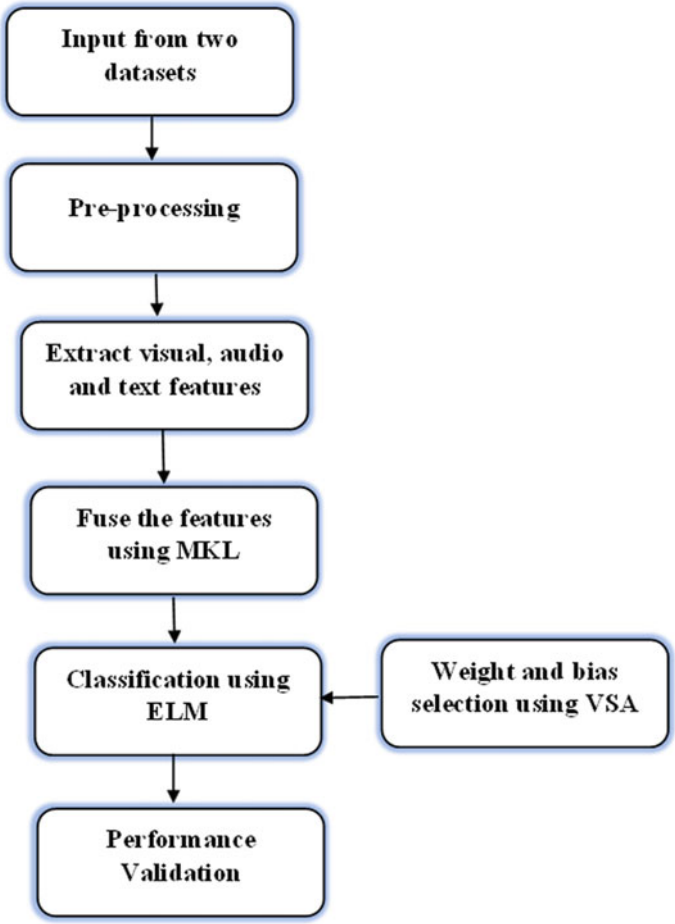

Fig. 1 Flow of projected model

a. To accurately forecast a truth value *y* that represents the sentimentality strength, the developed model must first extract various input vectors to build an uniform illustration.

3.2 Datasets and Metrics

We test our MSA methods using the CMU-MOSI [20] and CMU-MOSEI [20] datasets, both of which are freely accessible to the academic community. There are 89 unique narrators included in the 2199 CMU-MOSI utterance video segments found over 93 videos. In order to indicate the polarity of conveyed emotion, each

Table 1 Dataset split

Split	CMU-MOSEI (dataset 2)	CMU-MOSI (dataset 1)
Test	4659	686
Train	16,326	1284
Validation	1871	229
All	22,856	2199

segment is manually annotated with a sentiment value in the range of −3 to + 3. The CMU-MOSEI dataset improves upon the original CMU-MOSI by doubling its size. It includes 23,454 YouTube videos that are reviews of various films. Identical to CMU-MOSI in its labelling approach. Table 1 details the criteria by which the two datasets were separated.

3.3 Preprocessing

Following many earlier studies, we convert into standard tensors, as mentioned for our model, and enable fair rivalry with additional baselines.

For this procedure, we employ a textual modality known as "word embedding sources" [21]. This means that in all tests, the approach uses the embedding sources to encode the input raw text.

Modality of sight. In particular, we use Facet, an logical tool based on the Face Action Coding Systems (FACS) [22] to extract facial characteristics, in our research on CMU-MOSI and CMU-MOSEI. As a consequence, the vector lengths of the MOSI and MOSEI datasets are 47 and 35, respectively.

The Acoustical Modalities. The COVAREP [23] professional acoustic analysis framework was used to extract the acoustic characteristics.

Harmonisation of Methods. We conducted studies using word-aligned input signals. The researchers in this work employed P2FA [24] to synchronise the visual and auditory inputs to the same textual resolution. By averaging their representation vectors into a new one, the tool automatically sorts several frames into various groups and matches each group with a token. Across all trials, the study's models' text embeddings came from BERT-base-uncased.

3.4 Extracting Visual Data

Sentiment analysis of visual material at scale may aid in obtaining an accurate subject sentiment. The main problems with video sentiment detection are that it is computationally expensive and that training datasets are often poorly labelled, reducing the likelihood that the trained model will generalise well to unseen data. Due to

the massive size of the video data, we only look at every tenth frame throughout the training process. It uses the constrained local model (CLM) to locate the face's contour in each image. A lower resolution is used to scale down the cropped frame further. By doing so, we can significantly cut down on the quantity of video data used for training. A video's picture sequence serves as the input. We do this by combining the photos taken at times t and $t + 1$ to show how they are related through time. To acquire the 2D characteristics of Layer 1 from the modified input, we use kernels of varied diameters, shown as Kernels 1, 2, and 3.

Train the model on a variety of face sizes and shapes to increase its transferability to new problems. In addition, we train the model using product review videos from one domain and then test it using videos from a separate domain to ensure it is speaker-independent. All video frames were reduced in resolution by a factor of two during preprocessing. To obtain temporal convolution features, we fused together every successive frame of a movie into a single frame. By padding with zeros, we were able to get all the images down to exactly 250 by 500 pixels.

3.4.1 Extracting Features from Audio Data

Each video clip was annotated, and then audio characteristics were retrieved automatically for the research. We also utilised a sliding window of 100 ms to extract features from audio at a frame rate of 30 Hz. Specifically, it employed the free and open-source programme openSMILE to calculate the characteristics. The pitch and volume of a person's voice may be automatically extracted using this set of tools. The voices were normalised using Z-standardisation. To determine whether samples included or lacked human speech, a threshold was applied to the voice intensity after normalisation. Several Low-Level Descriptors (LLD) and statistical functionals of them make up openSMILE's extracted features. Kurtosis, quartiles, interquartile ranges, slope of linear regression, etc., are all examples of functionals. When we added up all the possible uses for each LLD, we got a total of 6,373.

3.4.2 Extracting Features from Textual Data

The length of a sentence determines whether the higher-order qualities are brief and specific, or extensive and encompassing the whole phrase. Each word in the text was used to create a 306-dimensional vector that was then concatenated with two other components to serve as input for the CNN feature extractor.

To embed a word: This research made use of a freely accessible word2vec dictionary that was trained with the use of Google News' 100 million-word corpus and the continuous bag-of-words architecture. Each word is signified as a vector of 300 dimensions in this dictionary. We resorted to random vectors for any entries that could not be located in this particular dictionary.

- We used a six-dimensional binary vector to represent the six basic components of speech (noun, verb, adjective, adverb, preposition, and conjunction). Our chosen part-of-speech tagger was Stanford Tagger.

The input vector is written as s (1:*n*) = s 1 s 2… s *n* if the instance has n words. The notation s I R k denotes a k-dimensional feature vector for the word s I (here, *k* = 306). All of the texts utilised in these analyses were no more than 65 words long.

3.5 Feature Selection and Fusion

Through the use of feature selection, the total number of features employed in the study was drastically reduced. The principal component analysis (PCA) method and the cyclic correlation method were the two methods used (PCA). The key idea behind CFS is that powerful feature subsets should consist of characteristics that are strongly connected with the target class yet uncorrelated with one another. In contrast, PCA employs an orthogonal transformation to the data to produce a set of linearly independent variables. The significance of the variables in the data may be used to rank them. Using principal component analysis (PCA), we may eliminate superfluous elements, allowing for dimensionality reduction and analogy. In this case, we choose the top K features from each method, with K being a number determined via experimentation. It was found that $K = 300$ when mixing audio, video, and text.

However, feature selection is handled separately across all three modes of investigation (unimodal, bimodal, and multimodal). To achieve feature-level fusion, the feature vectors obtained for the three modalities are added together. Clearly, the aggregated feature vectors from several modalities exhibit a wide range of variation. So, we trained an MKL algorithm classifier using the produced vectors and the correct sentiment polarity labels from the training set, using the SPF-GMKL implementation [25], which is prepared to handle heterogeneous data.

Cross-validation was used to settle on suitable parameters for the classifier. Eight kernels were employed in the study, including five RBF with gammas between 0.01 and 0.05 and three polynomial kernels with powers of 2, 3, and 4. Though Simple-MKL was also used, its output was much less outstanding.

3.6 Extreme Learning Machine (ELM)

The results obtained using traditional Neural Network (NN) approaches leave much to be desired in accuracy. For this reason, we present the Extreme Learning Machine (ELM), a Single-Layer Feed-forward Network (SLFN) with a learning speed one thousand times greater than that of any other existing feed forward network. Typically, in ELM, we have an input layer, a hidden layer, and an output layer. There is a random element to the pairing of weights ($w\ i$) and biases ($b\ j$) between the input and hidden

layers. The computational effort and complexity are reduced by using this method. In this study, the Gaussian activation function is applied to awaken the dormant neurons. Analytical estimates are made for the weights in the output. In the current study, we focus on solar power production as the end goal of the ELM and take into account temperature and irradiance as inputs. With 'N' random training sets ($x\ i = [x\ i_1$,$x\ i_2$,…x in]), the mathematical formula of SLFN may be written. "TRn and $t\ i = [t\ i_1, t\ i_2,\ldots, t$ im]" If we write TRn and $t\ i = [t\ i_1$,$t\ i_{2,\ldots,}\ t$ im], we get the following: TRm), 'L' is the number of hidden nodes, and $h(x)$ is the activation function (1).

$$\sum_{i=1}^{L} \beta_i h_i(x_j) = \sum_{i=1}^{L} \beta_i h_i(w_i x_j + b_i) = Y_j \tag{1}$$

where $w_i = [w_{i1}, w_{i2}, \ldots w_{in}]^T$ is the input weight matrix; $\beta_i = [\beta_{i1}, \beta_{i2}, \ldots \beta_{in}]^T$ is the output weight matrix to ELM which is measured as Y_j. Equation (1) is described fleetingly in Eq. (2).

$$H\beta = T \tag{2}$$

where H is the output matrix layer and is portrayed in Eq. (3).

$$H(w, b, x) = \begin{bmatrix} h_1(w_1 x_1 + b_1) & \cdots & h_L(w_L x_1 + b_L) \\ h_1(w_1 x_2 + b_1) & \cdots & h_L(w_L x_2 + b_L) \\ \vdots & \ddots & \vdots \\ h_1(w_1 x_N + b_1) & \cdots & h_L(w_L x_N + b_L) \end{bmatrix} \tag{3}$$

The output weight matrix can be signified as shadows in Eq. (4).

$$\beta = H^{+}T \tag{4}$$

Matrix H_+ is shown here as its Moore–Penrose generalised inverse. Prediction precision is determined entirely by $w\ i$ and $b\ i$, with b being the most important element. Unlike the VSA, which chooses $w\ i$ and $b\ i$ at random, the ELM may benefit from having these parameters adequately decided.

3.6.1 Selection of Weight and Bias Using Vortex Search Algorithm (VSA)

Vertical movement of fluids in a stirred tank triggers the VSA algorithm, a meta-heuristic optimization process. This approach, like other single-solution algorithms, makes advantage of simplified generation procedures. Any iteration of VSA populations may be converted into a state-of-the-art unified answer with the aid of values. Furthermore, the speed with which the update and seek operations are carried out in

the search region during each iteration run is essential for showing a single answer. This steadiness within the proposed VSA is accomplished by a vortex-like search pattern. The stacked circles represent some of the techniques used in vortex sampling.

Producing the initial solution

The preliminary procedure personalises 'center/parameters of ELM' μ_0 and 'radius' r_0. In this phase, the early 'center' (μ_0) can be intended using Eq. (5).

$$\mu_0 = \frac{upperlimit + lowerlimit}{2} \tag{5}$$

where parameters, which can be distinct in vector of $d \times 1$ dimensional space of network. In adding, σ_0 is the early *radius* r_0 produced with Eq. (6).

$$\sigma_0 = \frac{\max(upperlimit) - \min(lowerlimit)}{2} \tag{6}$$

Generating the candidate solutions

How populations change throughout time. The purpose of the approach of producing candidate solutions is to find Ct(s) in any given number of iterations. Assuming a Gaussian distribution, the VSA is constructed as follows: C0(s) = s 1, s 2,…,s m, m = 1, 2, 3,…, n, where s represents the solution and m represents the total sum of candidate solutions. The number n represents the total number of possible answers. Equation reveals the multivariate Gaussian distribution (Eq. (7))

$$p(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} exp\left\{-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu)\right\} \tag{7}$$

In equation (where d signifies the dimension, x represents the d1 vector of the random variable, m represents the sample mean (or centre), and c represents the covariance matrix), d represents the dimension (Eq. 7). When the diagonal components (i.e. variances) of a set of values are equal and the off-diagonal elements (i.e. covariance) are zero, as shown by Eq. (8), the values have a spherical distribution, where d is the dimension. The value is determined by assuming independent random variables with no correlation.

$$\Sigma = \sigma^2.[I]_{d \times d} \tag{8}$$

I'm using the *dd* identity matrix and the beginning radius (r 0) to represent the variance of the distribution in Eq. (8).

Extra of the current solution

The existing method of selection is changed. Though there are other ways to solve this puzzle, only one will do for our purposes; thus, we will memorise it directly from C0 (s). This solution is substituted for the current circle's epicentre. Candidate

solutions must be verified to be inside the search spaces before being considered for selection.

$$s_k^i = \begin{cases} rand.\left(upperlimit^i - lowerlimit^i\right) + lowerlimit^i, s_k^i < lowerlimit^i \\ rand.\left(upperlimit^i - lowerlimit^i\right) + lowerlimit^i, s_k^i > upperlimit^i \end{cases} \tag{9}$$

Random here refers to the equal distribution of k and I. In order to discover a better alternative, VSA shifts to using s as the new centre and Eq. (7) to reduce the size of the vortex. Accordingly, the novel set of solutions may be generated, C1 (s). If the selected response is more advantageous than the optimal one, it might be considered the optimal response going forward and committed to memory.

The radius decrement procedure

In the VSA, the inverse with each iteration. The incomplete gamma function shown in Eq. (10) has several uses in probability theory, especially in connection with the Chi-square distribution.

$$\gamma(x, a) = \int_0^x e^{-t} t^{a-1} dta0 \tag{10}$$

where x is an independent random variable and a shape parameter greater than zero. The balancing gamma function, (x, a), is typically presented in the same way as the incomplete gamma function (Eq. What is $a(a)$ in Eq. (11)) (Table 2).

$$\Gamma(x, a) = \int_0^\infty e^{-t} t^{a-1} dta0 \tag{11}$$

4 Results and Discussion

4.1 Evaluation Criteria

Each observation's predicted probability, which varies from 0 to 1 and denotes the likelihood of belonging to the positive class, is calculated in the classification problem. The confusion matrix shown in Table 3 is created by using a threshold of 0.5 to categorise each observation as either positive or negative.

$$\text{Accuracy} = (TP + TN)/(TP + FP + TN + FN) \tag{12}$$

Table 2 Pseudocode of VSA procedure

Initialising step
Algorithm parameters: Input the population size, the lower and upper bounds
Fitness of best solution
Centre of the circle (μ_0), Eq. (5)
Radius of the circle (σ_0), Eq. (6)
Repeat
Create candidate solutions within the circle by Eq. (7)
If exceeded, then shift values into the boundaries by Eq. (9)
Select best solution to replace the current centre
Decrease the standard deviation (radius) for the next iteration by Eq. (11)
End
Output
Best solution found so far S_{best}

$$\text{Recall(or Sensitive or True Positive rate)} = TP/(TP + FN) \tag{13}$$

$$\text{Precision} = TP/(TP + FP) \tag{14}$$

$$F1 - \text{Measure} = 2\frac{}{\text{Precision} \times \text{Recall}}\text{Precision} + \text{Recall} \tag{15}$$

$$\text{False Positive rate} = FP/(FP + TN) \tag{16}$$

However, experts in the field concur that these metrics are inadequate for use in sentiment analysis. Also, the misclassification costs in sentiment analysis will vary among the four groups of classification outcomes shown in the confusion matrix. When we make a wrong prediction and it turns out to be untrue, the resulting expenses are usually higher than those associated with a false positive, as the latter just necessitates the expense of an inquiry.

Therefore, we use the area under the ROC curve (AUC) value as an overarching performance metric in addition to the aforementioned metrics. There is a graphical representation of the ROC curve (Receiver Operating Characteristic) that shows how well the receiver performs. It compares the TPR to the FPR at various cutoffs. The

Table 3 Definition of the confusion matrix

		Predicted class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

AUC is preferred over accuracy as a measure of performance since it does not rely on a threshold. When an AUC value is close to 1, the model's performance is optimal.

4.2 Analysis of Proposed Model Using Two Datasets

The generic ML techniques are considered for comparison, which is implemented on two datasets and results are averaged in Tables 4 and 5. Because, the existing techniques use different combinations of datasets for MSA.

In the analysis of accuracy, the proposed ELM has 97%, where existing techniques achieved nearly 93–95% of accuracy. The reason for better performance is that the proposed model's weight and bias are optimally selected by using VSA. The existing techniques did not use any optimization models for selection of weights. In the analysis of AUC, the ELM-VSA achieved 97%, where ELM has 95%, DT and RF have 92%, SVM has 95%, and NB has 95%. The recall performance of ELM-VSA is poor, but it achieved better performance than existing models. For instance, DT and RF have 22%, SVM and ELM have 32%, NB has 39%, where the ELM-VSA has 42.19%. From this analysis, it is clearly proves that the proposed ELM-VSA achieved better presentation than existing procedures.

In the analysis of AUC, the proposed ELM-VSA has 97%, ELM has 96%, NB has 95%, RF and SVM have 94%, and DT has 93%. When comparing with all techniques, DT and RF achieved less accuracy (i.e. 94%), SVM has 95%, NB has 96%, ELM has

Table 4 Comparative analysis of proposed model in first dataset

Classification	AUC	F1-measure	Precision	Recall	Accuracy
DT	0.929	0.282	28.02	28.40	93.80
RF	0.929	0.257	29.79	22.62	94.10
SVM	0.935	0.311	29.56	32.82	93.90
NB	0.958	0.373	34.94	39.97	94.03
ELM	0.957	0.340	36.25	32.06	95.24
ELM-VSA	0.976	0.398	40.82	42.19	97.27

Table 5 Comparative analysis of proposed model in second dataset

Classification	AUC	F1-measure	Precision	Recall	Accuracy
DT	0.936	0.418	41.44	41.89	94.44
RF	0.941	0.358	40.10	33.01	94.50
SVM	0.949	0.437	44.59	41.98	95.89
NB	0.956	0.468	42.60	52.02	96.78
ELM	0.968	0.518	55.74	50.62	97.26
ELM-VSA	0.971	0.540	58.34	55.08	98.50

97%, and ELM-VSA has 98% of accuracy. When all the techniques are tested with precision and recall, the ELM-VSA has 55% of recall and 58.34% of precision, RF has low performance, i.e. 40% of precision and 33% of recall, and ELM has 55% of precision and 50% of recall. Figures 2, 3, 4, 5, and 6 show the graphical analysis of proposed model with existing practices in terms of various metrics.

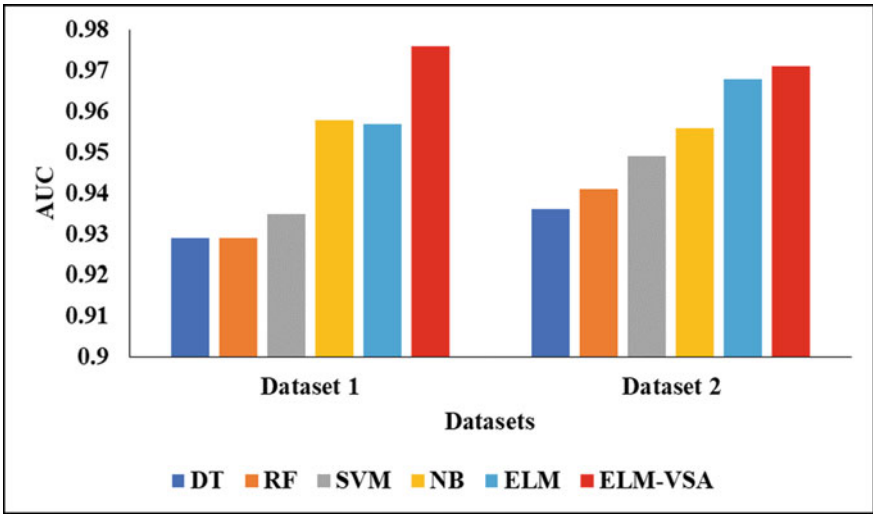

Fig. 2 AUC comparison on two datasets

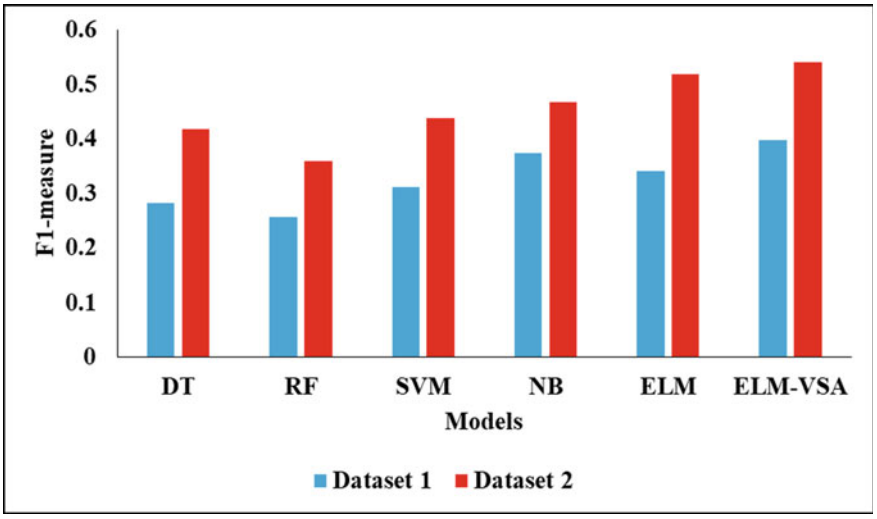

Fig. 3 F1-measure comparison on two datasets

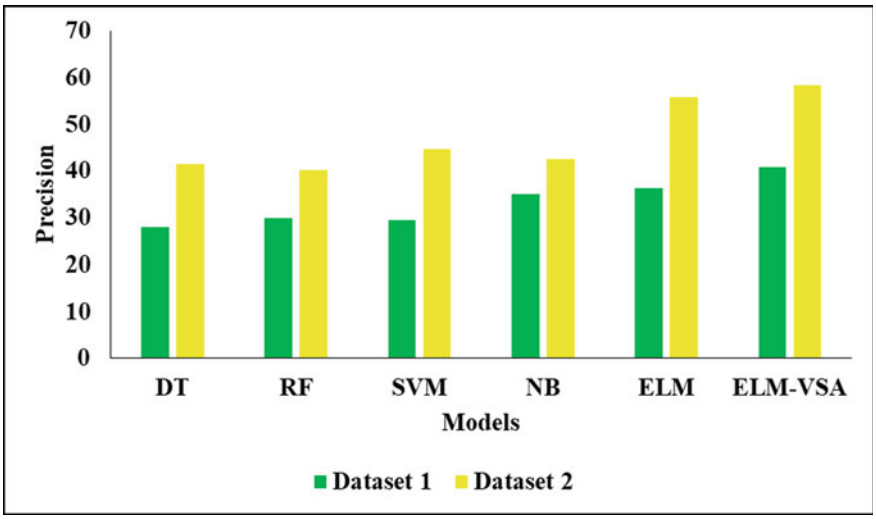

Fig. 4 Precision assessment on two datasets

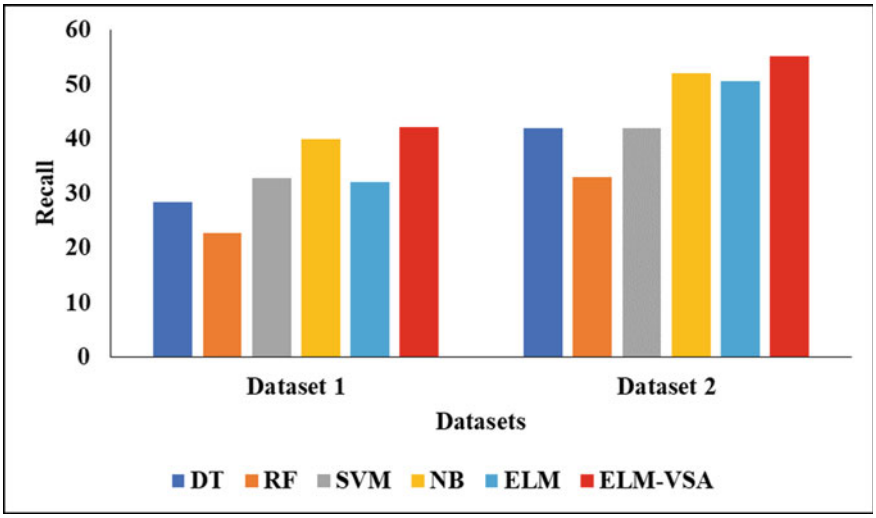

Fig. 5 Recall assessment on two datasets

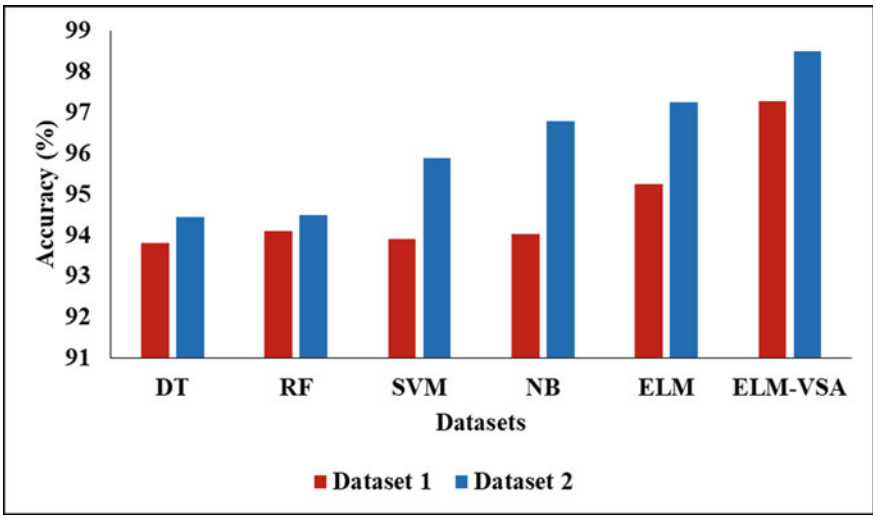

Fig. 6 Accuracy comparison on two datasets

5 Conclusion

Multimodal data, such as video, are quickly replacing text as the dominant form of online communication. Therefore, activities like financial prediction are increasingly dependent on the ability to extract feelings and polarity from films. In this research, we suggest utilising MKL to integrate speech, voice tone, and facial emotions into multimodal sentiment analysis. In specifically, the research fused diverse data acquired from three separate modalities (audio, video, and text) by using multiple kernels' learning. Afterwards, ELM is used to categorise the MSA, and VSA is used to choose the best values for the model's weight and bias. After analysing the data, we found that the ELM-VSA obtained 97.27% accuracy, whereas the ELM alone achieved 95% and the SVM just 93% on the first dataset. Future work should include using deep learning into the planned study to enhance the outcomes.

References

1. Huddar MG, Sannakki SS, Rajpurohit VS (2019) A survey of computational approaches and challenges in multimodal sentiment analysis. Int J Comput Sci Eng 7(1):876–883
2. Poria S, Majumder N, Hazarika D, Cambria E, Gelbukh A, Hussain A (2018) Multimodal sentiment analysis: addressing key issues and setting up the baselines. IEEE Intell Syst 33(6):17–25
3. Kasthuri S, Selvaraj R (2016) An effective text mining by using text pattern classification and relevant feature extraction. Int J Res Instinct (INJRI) 3(2):111–120. E-ISSN: 2348–2095

4. Hazarika D, Zimmermann R, Poria S (2020) Misa: modality-invariant and-specific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM international conference on multimedia, pp 1122–1131
5. Kasthuri S, Nisha Jebaseeli A (2019) Study on social network analysis in data mining. Int J Analytic Experiment Modal Anal (IJAEMA) (UGC CARE-A Journal), Impact Factor 6.3 XI(VIII):111–116. ISSN: 0886–9367
6. Sivaraman E, Manickachezian R (2019) Intelligent decision-making service framework based on analytic hierarchy process in cloud environment. Int J Networking Virtual Organ 21(2):221–236. https://doi.org/10.1504/ijnvo.2019.101787
7. Wang Z, Wan Z, Wan X (2020) Transmodality: an end2end fusion method with transformer for multimodal sentiment analysis. In: Proceedings of the web conference 2020, pp 2514–2520
8. Kim T, Lee B (2020) Multi-attention multimodal sentiment analysis. In: Proceedings of the 2020 international conference on multimedia retrieval, pp 436–441
9. Uma Maheswari V, Aluvalu R, Chennam KK (2021) Application of machine learning algorithms for facial expression analysis. Mach Learn Sustain Develop 9:77
10. Kamath R, Ghoshal A, Eswaran S, Honnavalli P (2022) An enhanced context-based emotion detection model using RoBERTa. In: 2022 IEEE international conference on electronics, computing and communication technologies (CONECCT), pp 1–6. https://doi.org/10.1109/CONECCT55679.2022.9865796
11. Kasthuri S, Jayasimman L, Nisha Jebaseeli A (2016) An opinion mining and sentiment analysis techniques: a survey. Int Res J Eng Technol (IRJET) 3(2):573–575. E-ISSN: 2395–0056
12. Agarwal A, Yadav A, Vishwakarma DK (2019) Multimodal sentiment analysis via RNN variants. In: 2019 IEEE international conference on big data, cloud computing, data science & engineering (BCD), IEEE. pp 19–23
13. Wang B, Ramazzotti D, De Sano L, Zhu J, Pierson E, Batzoglou S (2018) SIMLR: a tool for large-scale genomic analyses by multi-kernel learning. Proteomics 18(2):1700232
14. Ye J, Zhou J, Tian J, Wang R, Zhou J, Gui T, Zhang Q, Huang X (2022) Sentiment-aware multimodal pre-training for multimodal sentiment analysis. Knowl-Based Syst 258:110021
15. Mai S, Zeng Y, Zheng S, Hu H (2022) Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. IEEE Transact Affect Comput
16. Yang B, Wu L, Zhu J, Shao B, Lin X, Liu TY (2022) Multimodal sentiment analysis with two-phase multi-task learning. IEEE/ACM Transact Audio Speech Lang Process
17. Yan X, Xue H, Jiang S, Liu Z (2022) Multimodal sentiment analysis using multi-tensor fusion network with cross-modal modeling. Appl Artif Intell 36(1):2000688
18. Salur MU, Aydın İ (2022) A soft voting ensemble learning-based approach for multimodal sentiment analysis. Neural Comput Appl 34(21):18391–18406
19. Chen Q, Huang G, Wang Y (2022) The weighted cross-modal attention mechanism with sentiment prediction auxiliary task for multimodal sentiment analysis. IEEE/ACM Transact Audio Speech Lang Process 30:2689–2695
20. Han W, Chen H, Poria S (2021). Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. arXiv preprint arXiv:2109.00412.
21. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp 1532–1543
22. Han W, Chen H, Gelbukh A, Zadeh A, Morency LP, Poria S (2021) October. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In: Proceedings of the 2021 international conference on multimodal interaction. pp 6–15
23. Degottex G, Kane J, Drugman T, Raitio T, Scherer S (2014) COVAREP—a collaborative voice analysis repository for speech technologies. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE. pp 960–964
24. Yuan J, Liberman M (2008) Speaker identification on the SCOTUS corpus. J Acoustic Soc Am 125(5):3878–3878

25. Jain A, Vishwanathan S, Varma M (2012) Spf-gmkl: generalized multiple kernel learning with a million kernels. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 750–758